THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XXX OCTOBER 1951 NUMBER 4 PART 1

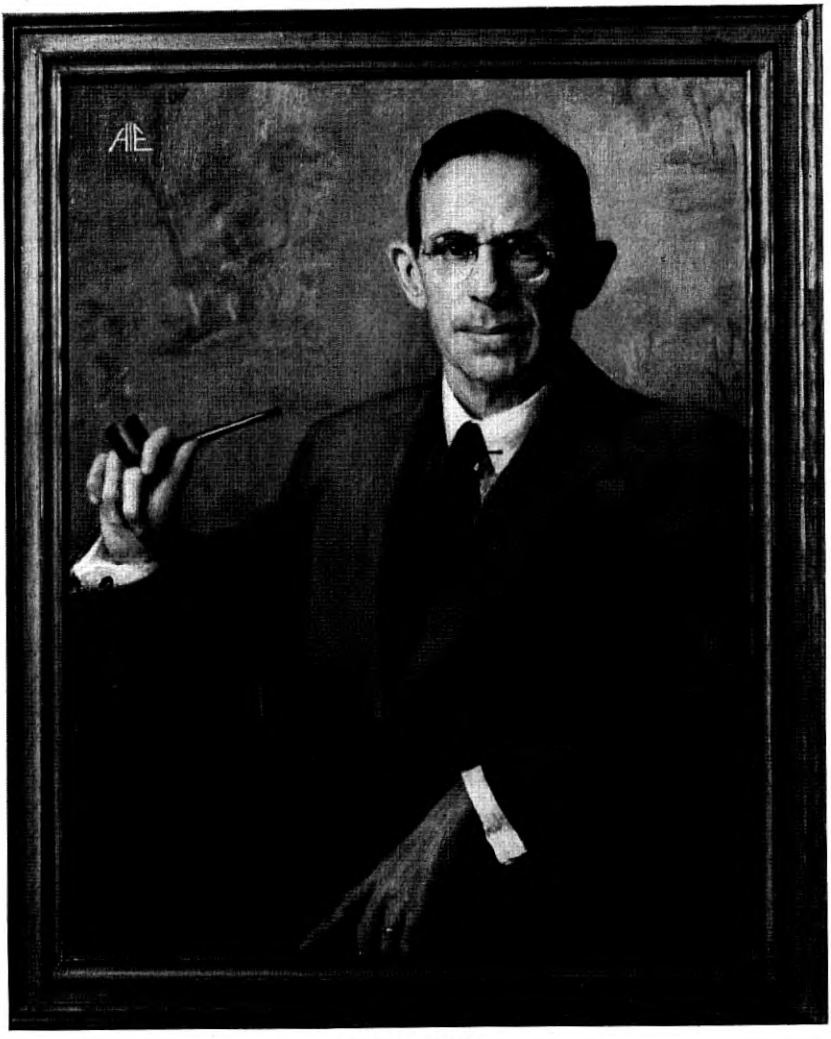
THIS ISSUE OF

THE JOURNAL CELEBRATES THE

SEVENTIETH BIRTHDAY OF

# CLINTON JOSEPH DAVISSON

IT IS CONTRIBUTED BY SOME OF HIS MANY FRIENDS AND FORMER ASSOCIATES IN THE BELL TELEPHONE LABORATORIES AS A TOKEN OF THEIR AFFECTION AND OF THEIR RECOGNITION OF THE VALUE OF HIS MANY CONTRIBUTIONS IN THE FIELD OF PHYSICAL RESEARCH



C. J. DAVISSON
From a portrait by H. E. Ives
of about 1938

The Bell System Technical Journal

*Vol. XXX* *October, 1951* *No. 4*



# Dr. C. J. Davisson

By M. J. KELLY

DR. DAVISSON, affectionately known to his large circle of friends as "Davy," joined the research section of the Engineering Department of the Western Electric Company in March, 1917 to participate in its World War I programs. He came on leave of absence from Carnegie Institute of Technology with the intention of returning to his academic post at the close of the war, but remained with its engineering organization, later to become Bell Telephone Laboratories, until 1946, when he retired at the age of sixty-five. He then accepted a research professorship in the Department of Physics at the University of Virginia.

When I joined Western's Engineering Department at the beginning of 1918, I had the good fortune to be assigned an office with Davisson. This was the beginning of a lifelong intimate friendship and an uninterrupted and close professional association terminated by his retirement.

Beginning in 1912 the Western Electric Company under the able leadership of Dr. H. D. Arnold pioneered in the development of the thermionic high-vacuum tube for communications applications. Although such devices already had important application as voice-frequency amplifiers in long-distance circuits at the time of our entrance into World War I, tubes were really not yet out of the laboratory, and the relatively few that were required for extending and maintaining service were made in the laboratories of the Engineering Department. Research and development programs directed to military applications of these new devices brought about a large expansion in the work of the laboratories. Davisson and I were assigned to the development of tubes for military use.

Important applications resulted from this work, and thermionic high-vacuum tubes had to be produced in what was for that time astronomical quantities. The science, technology, and art essential to such quantity production did not exist, and had to be created concurrently with a most rapid

build-up in production. All of the tubes employed an oxide-coated cathode, which was later to become the universal standard the world over for low-power thermionic vacuum tubes.

Davisson early took a position of leadership in problems of fundamental physics relating to the emitter and high-vacuum techniques. We were forced to move so rapidly that much of the work was necessarily empirical. Even in this atmosphere of empiricism Davisson's work was unusually fundamental and analytical. Increasingly all of us went to him to discuss fundamental problems that were in urgent need of an answer. He was always available and displayed a friendly interest; we rarely left him without benefit from the discussion. Frequently he would continue his study of the problem and come later to give the benefit of his more mature consideration.

During this period of intensive work performed in an atmosphere of urgency, Davisson displayed the characteristics that were important in determining the pattern of his work through the years and the nature of his contributions to our laboratories and to science. His inner driving force was always seeking complete and exact knowledge of the physical phenomena under study. Thoroughness was an outstanding characteristic. The rapid tempo of the work with the necessity of accepting partial answers and following one's nose in an empirical fashion were foreign to his way of doing things. As a war necessity he yielded to it, and performed as a good soldier. His interests were almost wholly scientific, but the needs of the situation forced upon him somewhat of an engineering role for which he had little appetite. As an adviser and consultant, he was unusually effective. In this he has few equals among scientists of my acquaintance. I believe that his success here is due to the high level of his interest in solving problems, to his broad area of curiosity about physical phenomena, and to his warm, friendly, and unselfish interest in the scientific aspects of the work of his associates.

Industry's scientific and technologic support of the war effort led to a rapid expansion of industrial laboratories in the postwar period. Our laboratories had expanded during the war period, and this was continued at a rapid rate throughout the following decade. The scientists who had come to the Laboratories during the war and the years immediately preceding it, with few exceptions moved out of the laboratory and assumed places of management and leadership in the research and development sections of the Laboratories' organization. At that early period in the life of industrial laboratories, the major emphasis was on applied research and development: there was very little research of a pure scientific nature.

Davisson was one of the few who did not gravitate to positions of man-

agement and leadership. His compelling interest in scientific research led Dr. Arnold to make a place in it for him, very rare in industrial laboratories of the time. A pattern of work of his own choosing gradually evolved, and he worked within it throughout his career. One or two young physicists and a few laboratory technicians made up the team that worked on his research problems. The young physicists and technicians did most of the work in the laboratory, although Davisson would frequently be found in the laboratory making observations in association with his co-workers. He took a leading part in planning the experiments and in designing the apparatus. His thoroughness and absorbing interest in detail were especially rewarding in this area for his experiments were always well conceived and their instrumentation was beautiful.

The maximum of reliability, long life (measured in years) and the highest electron-emitting efficiency from the cathode were early recognized as important to the full utilization of the thermionic high-vacuum tube in telecommunications. For several years after the close of the war, Davisson's researches were directed at a complete understanding of the emission phenomena of oxide-coated cathodes. This emitter is an unusually complex system. Chemical, metallurgical, and physical problems of great complexity are interleaved. Over the years, our laboratories have made great progress in reliability, long life, and high electron-emitting efficiency of thermionic vacuum tubes for telecommunication uses. The benefits of this work to the telephone user have been large, and annual savings to the Bell System of many millions of dollars have resulted. Davisson's researches during the five years following the close of the war. and his continuing advice to others through a longer period were significant in the advances that our laboratories have made.

As multigrid structures came into use and the tubes came to be used in circuits of ever increasing complexity, unwanted secondary electron emission from the grid structures became a major problem. The presence of this emission and its variation in amount from tube to tube brought about malfunctioning and unreliability. If it were to be controlled, its complete understanding was essential. A basic study of secondary emission was Davisson's next area of research. In these studies he came upon patterns of emission from the surface of single crystals of nickel that aroused his curiosity. His examination of these patterns led to his discovery of electron diffraction and the wave properties of electrons. In recognition of this masterful research with its important and highly significant results, he was awarded the Nobel Prize in 1937.

After the discovery of electron diffraction, Dr. L. H. Germer, who had worked with Davisson on the secondary emission researches, took the prob-

lem of applying electron diffraction to the study of the structure of thin surface films. He was the pioneer in utilizing electron diffraction in studies of surface structure, and made a large contribution to the science and technology of this new and important analysis technique. While Germer, graduating from his place as an aide to Davisson, worked independently on this problem, he benefited from frequent discussions with Davisson. After Germer had perfected an electron diffraction spectrometer, he operated it for a number of years as an analytical aid to many of the research and development projects of the Laboratories. The interpretation of the patterns and the determination of the crystalline structure of surface films were complex problems. During the period that Germer was developing techniques and getting order into the analysis of the patterns, Davisson often joined him in puzzling out the crystal structure revealed in photographs of the diffraction patterns of many different kinds of surfaces.

As a logical consequence of Davisson's interest in electron diffraction, he next concerned himself with a variety of problems in electron optics. He was one of the first to develop analytical procedures in the design of structures for sharply focusing electron beams. For many years, beginning in the early 1930's, Davisson gave much attention to the analytical side of electron optics and designed and constructed many structures for electron focusing. Prior to his work much of the vacuum-tube development work in our laboratories, as elsewhere where electron focusing was required, was largely empirical. Unfortunately he did not publish much of the fine work that he did, although he reported on portions of it to scientific and technical groups. However, the effect of his work and his ever increasing knowledge of electron optics on the programs and men of our laboratories concerned with electron dynamics was large. Dr. J. B. Fisk, Dr. J. R. Pierce, Dr. L. A. MacColl, Dr. Frank Gray and others of our laboratory obtained guidance and inspiration from Davisson, the consultant and adviser.

His work in electron optics came at a fortunate time in relation to our laboratories' studies of the transmission of television signals over coaxial conductor systems. Although it was possible to measure the amount and characteristics of the electrical distortion of signal currents, there were not available cathode ray tubes precise enough in their design for evaluating the degradation in the picture's quality resulting from the passage of the signal through the coaxial system. He undertook the development of a cathode-ray tube for this test purpose employing the principles of electron optics that he had worked out. In doing this he made one of his few excursions into technology. There resulted from his work a cathode-ray tube of great precision. By virtue of the fundamental design of the beam and deflecting system, the tube provided an extremely small rectangular spot on

the flourescent screen that remained in sharp focus over the entire screen area and had a much improved response characteristic. He took unusual pride in this project, and played a leading part in the design of every element of the complicated structure. The tube proved to be a useful tool in the evaluation of picture impairment resulting from different types of signal distortion.

Our laboratories steadily increased their participation in research and development activities for the military beginning in 1938. This effort expanded with terrific speed at the beginning of World War II, and soon became our major activity, continuing until the close of hostilities. Davisson was most anxious to contribute in any way that he could in our military work. While continuing his researches, he gave attention to the new and important multicavity magnetron that was receiving increasing attention. His background in electron optics made him invaluable as a consultant to Fisk, who led our magnetron work. As in World War I, speed was again the driving force in our programs, and substantially all of our research people turned to development. By keeping aloof from the rapidly moving development stream, he was able to give unhurried consideration to many of the basic electron dynamics problems of the magnetron.

When Dr. J. C. Slater joined us in 1943 to participate with Fisk in the basic magnetron problems, Davisson turned his attention to problems of crystal physics in relation to our programs on quartz crystal plates as circuit elements. Our laboratories were the focal point of a large national effort for the development, design, and production of quartz crystal plates for a multitude of electronic circuit applications. Drs. W. P. Mason, W. L. Bond, G. W. Willard, and Armstrong-Wood were the basic science team working on a multitude of problems that arose with the tremendous expansion of quartz plate production and use.

Davisson spent the major portion of his time from 1943 until his retirement in 1946 on a variety of crystal physics problems. He brought a fresh viewpoint into the crystal physics area. Through consultation, analyses, and experiments, he was of material assistance to our crystal physics group in the large contribution they made to the application of quartz plates to electronic systems for the military.

Davisson exerted a constructive influence on programs and men in the research and development areas of our laboratories throughout the thirty years of his active service. His door was open to all, and through his constructive interest in the problems presented, he developed large and continuing consulting contacts. This was not an assigned task but rather one that was personal to him, and its amount and continuance through the years were expressions of a facet of his personality. His contribution to the

adjustment of the young men to their change in environment from the university to industry as they came to our laboratories was considerable. It became a habit of the research directors to place with him for a year or so junior scientists on their entrance into our laboratories. Dr. J. A. Becker and Dr. William Shockley are typical of the men who were introduced into the Laboratories through a period of association with him.

We have always welcomed young scientists from the graduate schools of our universities for summer work. This gives them a view of the operation of an industrial laboratory, and is an aid to us in the selection of young research men from the schools. Several of them were assigned to work with Davisson. Some have since had distinguished careers in science. Drs. Lee A. DuBridge, Merle Tuve, and Philip Morse are among the graduate students who worked with Davisson during their summer employment with us.

It was fortunate that Davisson, who had come to stay for the duration of the war, elected to stay with the laboratories to work as a scientist in areas of physics important to our programs rather than return to university life. He established a pattern of fundamental research that has continued and enlarged in scope as our laboratories have evolved and reached maturity. Across the forefronts of the physics, mathematics, and chemistry, which are basic to telecommunication technology, we now have many scientists whose programs are directed, as was Davisson's, only at expanding fundamental knowledge, and who do not divert their energies even to the fundamental development phases of our technology. It is a tribute to Davisson's overpowering interest in science, and to his steadfastness in the pursuit of knowledge through the scientific method of experiment and analysis that during the pioneering and rapid expansion years of our laboratories, when development demanded the attention of most of our scientists, he gave almost undivided attention to the scientific aspects of our work. Throughout his career he has remained a scientist and has maintained a working knowledge at the forefront of a wide area of physics.

Throughout his thirty years at the Laboratories, Davisson's circle of friends among scientists steadily grew, not only within his own country but extending to Europe and the Orient. His capacity for friendships is large, and each of us at the Laboratories in daily contact with him has enjoyed a close friendship of exceptional warmth. The integrity and quality of his work are universally appreciated. He is held in high regard, not only for it but also because of his fine personal qualities. He is shy and modest. Because of this, it requires an association of some duration to know Davisson the man. He has a keen sense of humor, which flashes upon you in most unexpected ways. Unusually slight in stature with a fragile physical frame, his weight never exceeded 115 pounds, and for many years it hovered

around 100. While his health has been good, his store of energy has been limited, and it has been necessary for him to husband it carefully.

Davisson's modesty causes him to undervalue the importance and scope of his contributions. This characteristic, the low level of his energy, and the high standard he has always set for his work have combined to limit the amount of his publication. His influence on science and technology generally has, therefore, not been at so high a level as it has attained within the Laboratories, where his personal contact with individuals and their work has been more effective than publication.

In recognition of his constructive influence on the evolution of the basic science programs of our laboratories and his contribution of important new knowledge in the areas of thermionic emission, secondary electron emission, electron diffraction, and electron optics, the editors and the editorial board of The Bell System Technical Journal have invited members of our staff whose work and careers have benefited through association with Davisson to contribute papers to this issue of the Journal in celebration of his seventieth birthday.

# The Scientific Work of C. J. Davisson

By KARL K. DARROW

THE very first piece of work which is published by a physicist who is destined to be great is not often outstanding; but sometimes it has curious affinities, accidental rather than causal, with aspects of the work that was to come thereafter. In the first paper published by C. J. Davisson, we find him working with electrons, concentrating them into a beam by the agency of a magnetic field, directing them against a metal target, and looking to see whether rays proceed from the target. True, the electrons came from a radioactive substance, and therefore were much faster than those of his later experiments. True also, he did not actually focus the electron-beam. True also, the rays for which he was looking were X-rays, and in these he took no further interest. Yet in nearly all of his subsequent researches he was to use some of the principles of electron-focussing or electron-microscopy; in many, he was to look for things that were emitted by the target on which his electrons fell. This maiden paper was presented before the American Physical Society at its meeting in Washington in April 1909; the printed version may be found in the Physical Review, page 469 of volume 28 of the year 1909. It was signed from Princeton University, whither Davisson had gone as a graduate student.

Another characteristic of Davisson's work in his later years was his frequent study and use of thermionics. Already in 1911 we find him working in this field—but it was thermionics with a difference. The word "thermionics" now signifies, nearly always, the emission of electrons from hot metals; but at first it included also the emission of positive ions from hot metals and hot salts. Though neither useless nor uninteresting, the emission of positive ions is now rated far below the effect to which we now confine the name of thermionics: emission of electrons from hot metals is one of the fundamental phenomena of Nature, and its uses are illimitable. It may be plausibly conjectured that in 1911 the difference in the importance of the two phenomena—emission of positive ions and emission of electrons—was far less evident than it is now. Davisson, working under the British physicist O. W. Richardson who was then professor at Princeton, established that the positive ions emitted from heated salts of the alkali metals are once-ionized atoms of these metals—that is to say, atoms lacking a single electron. He also showed that if gas is present in the tube, it may enhance the number of the ions but does not change their character. This

work was presented before the April meeting of the American Physical Society in 1911. Abstracts of the papers which he there gave orally may be found in *Physical Review*, but the publications in full appeared (in 1912) in *Philosophical Magazine*. Davisson's choice of a British journal was advised by his transplanted teacher, but it must be realized that in 1912 the *Physical Review* had by no means ascended to the rank that it holds today. With this work came to their end the contributions of his student years, and next we find him publishing as an independent investigator.

From Davisson's years (1912–17) at the Carnegie Institute of Technology there is a paper embodying an attempt to calculate the optical dispersion of molecular hydrogen and of helium from Bohr's earliest atom-model. It shows him possessed of no mean mathematical technique, but is based—as the date by itself would make evident—on too primitive a form of quantum-theory.

In June 1917, in the midst of World War I, Davisson came for what he thought would be a temporary job at the institution then known as the Research Laboratories of the American Telephone and Telegraph Company and the Western Electric Company, thereafter—from 1925—as Bell Telephone Laboratories. Not for a year and a half was he able to devote himself to work untrammeled by the exigencies of war. So far as publication is concerned, his second period began in 1920, when he presented two papers before the American Physical Society: one at the New York meeting in February, one at the Washington meeting in April. In the former of these his name is linked with that of L. H. Germer, a name associated with his in the great discovery of electron waves; in the latter it is linked with that of the late H. A. Pidgeon.

These two papers are represented only by brief abstracts; and this is the more regrettable, as they form the only contributions published under Davisson's name to the dawning science of the oxide-coated cathode. In the former, he established that the remarkably high electron-emission of oxide-coated metals—as contrasted with bare metals—is *not* due, as had been elsewhere suggested, to the impacts of positive ions from the gas of the tube against the coatings: it is true thermionic emission. In the latter, he studied the rise and eventual fall of the thermionic emission as more and more oxide is laid down upon the metal surface, and concluded that the emission occurs when a definite number of oxide molecules is assembled into a patch of definite size on the surface: the number of patches of just the right size first rises, then declines as the deposition continues. According to colleagues of his, these two papers fall short by far of indicating the extent of his contributions to this field; and one of them has said that Davisson was excessively scrupulous about putting his work into print, being

unwilling to publish his observations until he felt sure that he understood all that was taking place. It is in an article by another—the late H. D. Arnold, first to hold the post of Director of Research in Bell Telephone Laboratories and its antecedent organization—that we find a description of Davisson's "power-emission chart," now standard in the art. In Arnold's words: "Dr. Davisson has devised a form of coordinate-paper in which the coordinates are power supplied to the filament (abscissae) and thermionic emission (ordinates). The coordinate lines are so disposed and numbered that if the emission from a filament satisfies Richardson's relation, and the thermal radiation satisfies the Stefan-Boltzmann relation, then points on the chart coordinating power and emission for such a filament will fall on a straight line."

In a paper presented at a meeting toward the end of 1920 (it was a joint paper of himself and J. R. Weeks) Davisson gives the theory of the emission of light from metals, deduces a deviation from Lambert's law and verifies this by experiment. A connection between this and the study of thermionics may be inferred from the words which I quoted earlier from Arnold's description of Davisson's power-emission chart. This work was published in full, some three years later, in the *Journal of the Optical Society of America*.

We turn now to Davisson's investigations of thermionic emission from metals.

Those whose memories go back far enough will recall that two laws have been proposed for the dependence of thermionic emission on temperature. Both were propounded by O. W. Richardson, and each, somewhat confusingly, has at times been called "Richardson's law." The earlier prescribed that the thermionic current $i$ should vary as $T^{\frac{1}{2}}\exp(-b/T)$, $T$ standing for the absolute temperature; the later prescribes that $i$ should vary as $T^2\exp(-b/T)$. The former is derived from the assumption that the velocities and energies of the electrons inside the metal are distributed according to the classical Maxwell-Boltzmann law. The latter follows from the assumption that these velocities and energies are distributed according to the quantum-theory or Fermi-Dirac law: it was, however, derived from thermodynamic arguments some thirteen years before the Fermi-Dirac theory was developed, and the experiments about to be related were performed during this thirteen-year period.

In the interpretation of either law, $b$ is correlated with the work of egress which an electron must do (at the expense of its kinetic energy) in order to go from the inside to the outside of the metal. I will leave to a later page the phrasing of this correlation, and say for the moment that $b$ multiplied by Boltzmann's constant $k$ represents what used to be called and is still some-

times called the "thermionic work-function" of the metal. If a given set of data is fitted first by the $T^{\frac{1}{2}}$ law and then by the $T^2$ law, different values of $b$ and therefore different values of the thermionic work-function are obtained. Which is right?

This question can be answered if the thermionic work-function can be measured with adequate accuracy by some other method. Such a method exists: it is called the "calorimetric" method. Suppose an incandescent wire surrounded by a cylindrical electrode. If the latter is negative with respect to the former, the emitted electrons will return to the wire, and there will be no net thermal effect due to the emission. If, however, the cylinder is positive with respect to the wire, the electrons will be drawn to it, and the wire will fall in temperature: this is the "cooling-effect" due to the emission. The resistance of the wire will decrease, and if the current into the wire is held constant, the voltage between its terminals will be lessened.

The experiment may sound easy, and so it might be if all of the current flowed within the wire from end to end; but the bleeding of electrons through the entire surface makes the current vary from point to point along the wire, and complicates the test enormously. Others elsewhere had tackled this difficult problem of experimentation; but Davisson and Germer found a better way to handle it, and their results for tungsten were presented at a meeting at the end of 1921 and published fully the following year. From their data they calculated the thermionic work-function of the metal, which when thus determined we may denote by $e\phi$. It agreed with the value of $kb$ obtained from the newer form of "Richardson's law," disagreed with the other. Thus Davisson was in the position of having confirmed the Fermi-Dirac distribution-law before it had been stated!

It remains to be said that, years later, Davisson and Germer repeated this experiment upon an oxide-coated platinum wire. Here they came upon a complication from which clean metal surfaces are fortunately exempt. The character of the oxide-coated wire changed with the temperature; and, since the measurement of the "constant" $b$ requires a variation of the temperature, its value did not provide a reliable measure of the work at any single temperature, whereas the "calorimetric" measurement did.

Now at last we are ready to attend to the early stages of the studies which were destined to lead to the discovery of electron-waves. These were studies of what I shall call the "polycrystalline scattering patterns" of metals: the name is descriptive rather than short. A beam of electrons is projected against a metal target which is in the condition, normal for a metal, of being a complex of tiny crystals oriented in all directions. Some of these electrons swing around and come back out of the metal with undiminished energy: these are the electrons that are "elastically scattered,"

(Davisson records that elastic scattering had previously been observed only with electrons having initially an energy of 12 electron-volts or less). A collector is posted at a place where it collects such electrons as are scattered in a direction making some chosen angle $\theta$ with the direction exactly opposite to that of the original or "primary" beam. There may be inelastically-scattered or secondary electrons which travel toward the collector: its potential is so adjusted as to prevent the access of these.

The collector is moved from place to place so as to occupy successively positions corresponding to many values of the angle $\theta$. It is always in the same plane passing through the primary beam, and so the curve of number-of-scattered-electrons (per unit solid angle) plotted against $\theta$ is a cross-section of a three-dimensional scattering pattern; but, for obvious reasons of symmetry, the three-dimensional pattern is just the two-dimensional pattern rotated around the axis which is provided by the primary beam. This two-dimensional pattern is what I have called the polycrystalline scattering-pattern. It is a curve plotted, in polar coordinates or in Cartesian, against $\theta$ over a range of this angle which extends from $-90°$ to $+90°$; but the part of the curve which runs from $\theta = -90°$ to $\theta = 0°$ is the mirror-image of the other part, and either by itself suffices. The curve cannot be plotted in the immediate vicinity of $\theta = 0°$, because the source of the electrons gets in the way.

The first published report of such an experiment is to be found, under the names of Davisson and C. H. Kunsman, in *Science* of November 1921; in that same November Davisson presented the work before the American Physical Society. The metal was nickel, and the pattern had two most remarkable features. These were sharp and prominent peaks; one inferred from the trend of the curve in the neighborhood of $\theta = 0°$ and presumably pointing in exactly that direction, consisting therefore of electrons which had been turned clear around through 180 degrees; the other pointing in a direction which depended on the speed of the electrons, and for 200-volt electrons was at 70°.

Any physicist who hears of experiments on scattering is likely to think of the scattering-experiments performed by Rutherford now more than forty years ago, which established the nuclear atom-model. These were measurements of the scattering-pattern of alpha-particles, and this does not look in the least like the curve observed by Davisson and Kunsman: it shows no peaks at all. Alpha-particles, however, are seven thousand times as massive as electrons: they are deflected in the nuclear fields, and so great is the momentum of an alpha-particle that it does not suffer any perceptible deflection unless and until it gets so close to a nucleus that there are no electrons at all between the nucleus and itself. But with so light a particle

as an electron, and especially with an electron moving as slowly as Davisson's, the deflection commences when the flying electron is still in the outer regions of the atom which it is penetrating. The deflection of the individual electron and the scattering-pattern of the totality of the atoms are, therefore, conditioned not only by the nuclear field but by the fields of all the electrons surrounding the nucleus. How shall one calculate the effect of all these?

This is a very considerable mathematical problem, and Davisson simplified it to the utmost by converting the atomic electrons into spherical shells of continuous negative charge centered at the nucleus. The simplest conceivable case—not to be identified with that of nickel—is that of a nucleus surrounded by a single spherical shell having a total negative charge equal in magnitude to the positive charge of the nucleus itself. Within the shell the field is the pure nuclear field, the same as though the shell were not there at all; outside of the shell there is no field at all. This is what Davisson called a "limited field." Calculation showed that the scattering-pattern of such a system would have a peak in the direction $\theta = 0°$, so long as the speed of the electrons did not exceed a certain ceiling-value! And there was more: "the main features of the scattering-patterns (Davisson said "distribution-curves") for nickel, including the lateral maximum of variable position, are to be expected if the nickel atom has its electrons arranged in two shells."

Nickel in fact is too complicated an atom to be represented, even in the most daring allowable approximation, as a nucleus surrounded by a single shell; two shells indeed seem insufficient, but the fact that a two-shell theory leads in the right direction is a significant one. Magnesium might reasonbly be approximated by a single-shell model; Davisson experimented on this metal, and published (in 1923) scattering-patterns which lent themselves well to his interpretation. He measured scattering-patterns of platinum also, and these as to be expected are much more wrinkled with peaks and valleys; the task of making calculations for the platinum atom with its 78 electrons was too great.

Nickel continued to be Davisson's favorite metal, and four years later (1925) his study of its polycrystalline scattering-pattern was still in progress. In April of that year occurred an accident, of which I quote his own description from *Physical Review* of December 1927. "During the course of his work a liquid-air bottle exploded at a time when the target was at a high temperature; the experimental tube was broken, and the target heavily oxidized by the in-rushing air. The oxide was eventually reduced and a layer of the target removed by vaporization, but only after prolonged heating at various high temperatures in hydrogen and in vacuum. When the

experiments were continued it was found that the distribution-in-angle of the scattered electrons had been completely changed. . . . This marked alteration in the scattering-pattern was traced to a re-crystallization of the target that occurred during the prolonged heating. Before the accident and in previous experiments we had been bombarding many small crystals, but in the tests subsequent to the accident we were bombarding only a few large ones. The actual number was of the order of ten."

I do not know whether Davisson ever cried out *O felix culpa!* in the language of the liturgy; but well he might have. The exploding liquid-air bottle blew open the gate to the discovery of electron-waves. Fatal consequences were not wanting: the accident killed the flourishing study of polycrystalline scattering-patterns, and countless interesting curves for many metals are still awaiting their discoverers. This may illustrate a difference between the industrial and the academic career. Had Davisson been a professor with a horde of graduate students besieging him for thesis subjects, the files of *Physical Review* might exhibit dozens of papers on the scattering-patterns of as many different metals, obtained by the students while the master was forging ahead in new fields.

Now that we are on the verge of the achievement which invested Davisson with universal fame and its correlate the Nobel Prize, I can tell its history in words which I wrote down while at my request he related the story. This happened on the twenty-fifth of January, 1937: I have the sheet of paper which he signed after reading it over, as also did our colleague L. A. MacColl who was present to hear the tale. This is authentic history such as all too often we lack for other discoveries of comparable moment. Listen now to Davisson himself relating, even though in the third person, the story of the achievement.

"The attention of C. J. Davisson was drawn to W. Elsasser's note of 1925, which he did not think much of because he did not believe that Elsasser's theory of his (Davisson's) prior results was valid. This note had no influence on the course of the experiments. What really started the discovery was the well-known accident with the polycrystalline mass, which suggested that single crystals would exhibit interesting effects. When the decision was made to experiment with the single crystal, it was anticipated that 'transparent directions' of the lattice would be discovered. In 1926 Davisson had the good fortune to visit England and attend the meeting of the British Association for the Advancement of Science at Oxford. He took with him some curves relating to the single crystal, and they were surprisingly feeble (surprising how rarely beams had been detected!). He showed them to Born, to Hartree and probably to Blackett; Born called in another Continental physicist (possibly Franck) to view them, and there was much

discussion of them. On the whole of the westward transatlantic voyage Davisson spent his time trying to understand Schroedinger's papers, as he then had an inkling (probably derived from the Oxford discussions) that the explanation might reside in them. In the autumn of 1926, Davisson calculated where some of the beams ought to be, looked for them and did not find them. He then laid out a program of thorough search, and on 6 January 1927 got strong beams due to the line-gratings of the surface atoms, as he showed by calculation in the same month."

Now I will supplement this succinct history by explanations. The first name to be mentioned in the explanations must be one which does not appear in the quotation: that of Louis de Broglie.

Louis de Broglie of Paris had suggested that electrons of definite momentum—let me denote it by $p$—are associated with waves of wavelength $\lambda$ equal to $h/p$, $h$ standing for Planck's constant. This suggestion he made in an attempt to interpret the atom-model of Bohr, a topic which is irrelevant to this article. Irrelevant also is the fact that Louis de Broglie's suggestion led Schroedinger to the discovery of "wave-mechanics," but I mention it here because Schroedinger's name appears in the quotation. Highly relevant is the inference that the "de Broglie waves," as they soon came to be called, might be diffracted by the lattices of crystals, and that the electrons of an electron-beam directed against a crystal might follow the waves into characteristic diffraction-beams such as X-rays exhibit.

This inference was drawn by a young German physicist Walther Elsasser by name, then a student at Goettingen. It was one of the great ideas of modern physics; and, in recording that its expression in Elsasser's letter was not what guided Davisson to its verification, I have no wish to weaken or decry the credit that justly belongs to Elsasser for having been the first to conceive it. Dr. Elsasser has authorized me to publish that he submitted his idea to Einstein, and that Einstein said "Young man, you are sitting on a gold-mine." The letter which I have mentioned appeared in 1925 in the German periodical *Die Naturwissenschaften*. As evidence for his idea Elsasser there adduced the polycrystalline scattering-patterns, in particular those for platinum, that had been published by Davisson and Kunsman. But Davisson as we have seen did not accept this explanation of the patterns; and never since, so far as Elsasser or I are aware, has anyone derived or even tried to derive the polycrystalline scattering-patterns from the wave-theory of electrons. This must be listed as a forgotten, I hope only a temporarily forgotten, problem of theoretical physics.

Essential to the application of Elsasser's idea is the fact that the wavelengths of the waves associated with electrons of convenient speeds are of the right order of magnitude to experience observable diffraction from a

crystal lattice. It is easy to remember that 150-volt electrons have a wavelength of one Angstrom unit, while the spacings between atoms in a solid are of the order of several Angstroms. This fact of course did not escape Elsasser, and it figures in his letter.

From the quotation it is clear that the earliest patterns obtained from the complex of large crystals were obscure, and the definitive proof of Elsasser's theory was obtained only when Davisson instituted his "program of thorough search" and simultaneously in England G. P. Thomson instituted his own. Two other items in the quotation require to be explained. The hypothesis of "transparent directions" I will consider to be explained by its name. Were it correct, the directions of the beams would be independent of the speed of the electrons; since they are not, the hypothesis falls. The reference to the "line-gratings of the surface atoms" induces me to proceed at once to one of the principal contrasts between diffraction of electrons and diffraction of X-rays.

An optical grating is a sequence of parallel equidistant grooves or rulings on a surface of metal or glass. The atoms on a crystalline surface are arranged in parallel equidistant lines, and one might expect X-rays or electrons to be diffracted from them as visible light is diffracted from an optical grating. This expectation is frustrated in the case of X-rays, because their power of penetration is so great that a single layer of atoms, be it the surface-layer or any other, diffracts but an inappreciable part of the incident X-ray beam; only the cumulative effect of many layers is detectable. Electrons as slow as those that Davisson used are not nearly so penetrating. With these indeed it is possible, as he was the first to show, to get diffraction-beams produced by the surface-layer only. Such beams, however, are detectable only when the incident (or the emerging) beam of electrons almost grazes the surface; and nearly always, when a beam is observed, it is due to the cumulative effect of many atom-layers as is the rule with X-rays. But the cumulative effect requires more specific conditions than does diffraction by the surface-layer: if the incident beam falls at a given angle upon the surface, the momentum of the electrons and the wavelength of their waves must be adjusted until it is just right, and, reversely, if the momentum of the electrons has a given value the angle of incidence must be adjusted until it is just right. This also Davisson verified.

As soon as Davisson made known his demonstration of electron-waves, he was bombarded by entreaties for speeches on his work and for descriptions to be published in periodicals less advanced and specialized than *Physical Review*. To a number of these he yielded, and I recommend especially the talk which in the autumn of 1929 he gave before the Michelson Meeting of the Optical Society of America; one finds it in print in volume

18 of the Journal of that Society. It is written with such clarity, grace and humor as to make one regret that Davisson was not oftener tempted to employ his talents for the benefit not of laymen precisely, but of scientists who were laymen in respect to the field of his researches. I quote the first two sentences: "When I discovered on looking over the announcement of this meeting that Arthur Compton is to speak on 'X-rays as a Branch of Optics' I realized that I had not made the most of my opportunities. I should have made a similar appeal to the attention of the Society by choosing as my subject 'Electrons as a Branch of Optics.' "

Though in this period his duties as expositor took a good deal of his time, Davisson found opportunity to prosecute his work and to begin on certain applications. One obvious development may be dismissed rather curtly, as being less important than it might reasonably seem. One might have expected Davisson to strive to verify de Broglie's law $\lambda = h/p$ to five or six significant figures. This would have been difficult if not impossible, since the diffraction-beams of electrons are much less sharp than those of X-rays; this is a consequence of the fact that the diffraction is performed by only a few layers of atoms, the primary beam being absorbed before it can penetrate deeply into the crystal structure. But even if it had been easy the enterprise would probably have been considered futile, for de Broglie's law quickly achieved the status of being regarded as self-evidently true. Such a belief is sometimes dangerous, but in this case it is almost certainly sound: the law is involved in the theories of so many phenomena, that, if it were in error by only a small fraction of a per cent, the discrepancy would have been noted by now in more ways than one. Davisson established the law within one per cent, and there are few who would not regard this as amply satisfactory.

The greatest of the uses of electron-diffraction lies in the study of the arrangement of atoms in crystals and in non-crystalline bodies. Here it supplements the similar use of X-ray diffraction, for it serves where X-ray diffraction does not, and *vice versa*. Once more I quote from a lecture of Davisson's: "Electrons are no more suitable for examining sheets of metal by transmission than metal sheets are suitable for replacing glass in windows. To be suitable for examination by electrons by transmission, a specimen must be no more than a few hundred angstroms in thickness. It must be just the sort of specimen which cannot be examined by X-rays. Massive specimens can be examined by electrons by reflection. The beam is directed onto the surface at near-grazing incidence, and the half-pattern which is produced reveals the crystalline state of a surface-layer of excessive thinness. . . . Invisible films of material, different chemically from the bulk of the specimen, are frequently discovered by this method." Many experi-

ments of this type were done at Bell Telephone Laboratories by L. H. Germer; they do not fall within the scope of this article, but we may be sure that Davisson was interested in them.

Davisson also studied the refraction of electrons at the surface of nickel, and this is work which in my opinion has never received the attention that it merits. Let us consider its importance.

I have already spoken of the work which an electron must do in order to quit a metal, and have mentioned two ways employed by Davisson (and by others) to ascertain its value—the measurement of the constant $b$ which figures in Richardson's equation, and the measurement of the quantity $\phi$ by the calorimetric method. It is customary to ascribe this work of egress to the presence of a "surface potential-barrier," usually imagined as an infinitely sudden potential-drop occurring at the surface of the metal: the potential immediately outside the metal is supposed to be less than the potential immediately inside by a non-zero amount, which I will denote by $X$. One is tempted to identify $X$ with $\phi$ and with $kb/e$; but this is an over-simplification. By the classical theory there is a difference which is small but not quite negligible. By the new theory there is a difference which is neither small nor by any means negligible. By the new theory, in fact, $X$ is greater than $\phi$ by an amount which is equal to the so-called "Fermi energy"—the kinetic energy of the electrons which, if the metal were at the absolute zero of temperature, would be the fastest-moving electrons in the metal. Now, this last amount is of the order of half-a-dozen electron-volts for the metals of major interest in thermionic experiments, and so also is the value of $\phi$. Thus, if there were a method for determining the height of the surface potential-barrier, this would be expected to yield a value of the order of six volts if the old theory were right and a value of the order of twelve volts if the new theory were correct.

Well, there *is* such a method, and it consists precisely in observing and measuring the refraction of the electron-waves as they pass through the surface of the metal. This refraction has a deceptive effect; it alters the orientations of the diffraction-beams as though the crystal were contracted in the direction normal to its surface. Once this is comprehended, the refractive index may be calculated from the observations, and from the refractive index the value of $X$ the surface potential drop. This was done by Davisson and Germer for nickel, and published in the *Proceedings of the National Academy of Sciences* for 1928. The value which they found for the surface potential-drop was 18 volts—three times as great as the value prescribed by the old theory, half again as great as the value afforded by the new. Thus the experiments speak for the new theory over the old, yet not with unambiguous support of the new. This has been described to me, by a dis-

tinguished physicist, as one of the situations in which the concept of a single sharp potential-drop becomes most palpably inadequate. Work of this kind continued to be done, especially in Germany, until the later thirties, and then regrettably flickered out.

In 1937 the Nobel prize was conferred on Davisson, and he had the opportunity of enjoying the ceremonies and festivities which are lavished upon those who go to Stockholm and receive it. He shared the prize with G. P. Thomson, who must not be entirely neglected even in an article dedicated explicitly to Davisson. There was little in common between their techniques, for Thomson consistently used much faster electrons which transpierced very thin polycrystalline films of metal and produced glorious diffraction-rings. He too founded a school of crystal analysts.

Finally I mention three notes—two abstracts of papers given before the American Physical Society in 1931 and 1934, and one Letter to the Editor of Physical Review—bearing on what has been described to me, by an expert in the field, as the first publication of the principle of the "electrostatic lens" useful in electron-microscopy. These are joint papers of Davisson and C. J. Calbick. They report, in very condensed form, the outcome of an analysis which showed that a slit in a metal cylinder treats electrons as a cylindrical lens treats light, and a circular hole in a metal plate treats electrons as a spherical lens treats light: in both cases the field-strengths on the two sides of the metal surface (cylinder or plate) must be different. Experiments were performed to test the theory, and succeeded; and in the latest of the notes we read that Calbick and Davisson used a two-lens system to form a magnified image of a ribbon-filament upon a fluorescent screen. Calbick recalls that the magnification was of the order of twentyfold.

During the time of his researches on electron-waves, Davisson's office was on the seventh floor of the West Street building, on the north side about seventy-five paces back from the west facade: his laboratories were at times beside it, at times across the corridor. This illustrates a disadvantage of our modern architecture. If Davisson had done his work in a mediaeval cathedral, we could mount a plaque upon a wall which had overlooked his apparatus, and plaque and wall would stand for centuries. But the inner walls of Davisson's rooms are all gone, and the outer wall consisted entirely of windows; and nothing remains the same except the north light steaming through the windows, which we may take as a symbol of the light which Davisson cast upon the transactions between electrons and crystals.

# Inorganic Replication in Electron Microscopy

By C. J. CALBICK

Contrast and resolution in electron micrographs from thin replica films are determined by the geometrical relationships between the directions of incidence of the condensing atom beam and the local surface normal, during film formation by evaporation *in vacuo*, and the direction of incidence of the electron beam, during subsequent exposure in the microscope. Replica films may be formed of any material suitable for vacuum evaporation. Metal atoms in general tend to stick where they strike, moving only short distances, 100 Å or less, to nucleating centers where they form small crystallites. Oxides such as silica and silicon monoxide, and also the semi-metal germanium, form amorphous films. A portion of the incident material, about 50% in the case of silica, migrates large distances, 5000 Å or more, before finally condensing; the remainder sticks where it first strikes the surface.

The existence of a minimum perceptible mass thickness difference, about 0.7 $\mu g/cm^2$ for 50 kv electrons, results in an optimum replica mass thickness of about 10 $\mu g/cm^2$. The resolution of the replica film is proportional to its linear thickness and hence is inversely proportional to its density. Micrographs of silica, chromium, gold-manganin, aluminum, aluminum-platinum-chromium and germanium replicas are shown. The importance of stereoscopic methods in interpretation of micrographs is discussed.

THE basic purpose of micrography of surfaces is to exhibit structural topography. Present day electron microscopes are transmission-type instruments. Practical limitations of experimental technique establish a voltage of the order of 50 kv as the most useful accelerating potential for the electrons used for illumination. In bright field transmission microscopy, the image consists of a field with local variations of intensity produced because the object has partially absorbed, or scattered, the incident radiation. In electron imaging scattering is the predominant factor, limiting direct examination to objects whose mass thickness does not exceed about 50 $\mu g/cm^2$.* Thicker specimens can be examined only in profile.

Optical microscopy of surfaces is concerned with their appearance as seen by reflected light, the counterpart of which is not practicable[1] with electrons. The electron microscopist has therefore devised means of transferring surface structural details to thin films called replicas.[2] These films must present to the electron beam locally varying thickness corresponding to the surface details. A simple type is the plastic replica[3] consisting of an appropriately thin plastic film stripped from the surface. A second type

* Some microscopes provide a range of accelerating potentials, up to 100 kv or more, permitting direct examination of thicker objects.

[1] Zworykin et al. "Electron Optics and the Electron Microscope," pp. 98–106.

[2] *J. Roy. Micro. Soc., 70*, 1950, "The Practise of Electron Microscopy," ed. by D. G. Drummond, see Chapters II and V.

[3] V. J. Schaefer and D. Harker, *Jl. App. Phys., 13*, 427 (1942).

is the oxide film,[2] produced by controlled oxidation of the surface when it is aluminum or another suitable metal. For other materials, a pressure mold of the surface in pure aluminum may be utilized as an intermediate replica in a two-step process.[4] A third type is the silica replica,[5] [6] produced by the condensation of silica vaporized by a hot source *in vacuo*, either on the surface in a one-step, or on a plastic mold of the surface in a two-step, process. A fourth type is the shadow-cast plastic replica,[7, 8] produced by similar deposition of a suitable metal at near-glancing incidence upon a plastic replica.

The purpose of this paper is to discuss the process of evaporated film formation as it is related to properties important in microscopy. The resolution and range of contrast available in the finished micrograph determine the faithfulness with which the original surface is depicted, and depend on the relative orientation of the surface, vapor source, and electron beam, on the density and average thickness of the replica, and on the mechanism of condensation. In principle, it is pointed out that any material of suitable physical and chemical properties may be used for evaporated replica films, and a number of examples are shown in micrographs. Inorganic replica films retain the third or vertical dimension, a fundamental advantage which permits stereoscopic study. The material presented perhaps provides a unified view of replication techniques and a method for the evaluation of micrographs relative to the faithfulness of portrayal of the original surface.

## 1. Local Thickness of Condensed Material

Figure 1 illustrates how the thickness $t_e$ in the direction of the electron beam is dependent on the local surface normal $n$. The thickness $t_a$ in the direction of the atom source is constant, depending only on the amount of material reaching the surface. The thicknesses due to two or more sources obviously may be vectorially added, and hence *an arbitrary assembly of sources may be replaced by a single source properly located*, since each yields the same thickness distribution on the surface $S$. A simple analogy is the shading produced when ordinary objects are illuminated by direct light. It is clear that atom source and electron beam must differ in direction for shading to occur. The atoms or molecules of some materials, notably silica and silicon monoxide, do not all stick where they strike, but some wander over the surface[9] as a "two-dimensional gas" before condensing. This

[4] J. Hunger and R. Seeliger, *Metallforschung*, *2*, 65 (1947).
[5] R. D. Heidenreich and V. G. Peck, *Jl. App. Phys.*, *13*, 427 (1943).
[6] C. H. Gerould, *Jl. App. Phys.*, *17*, 23 (1947).
[7] H. Mahl, *Korrosion u. Metallschutz*, *20*, 225 (1945).
[8] R. C. Williams and R. W. G. Wyckoff, *Jl. App. Phys.*, *17*, 23 (1946).
[9] R. D. Heidenreich, *Jl. App. Phys.*, *14*, 312 (1943).

results in a somewhat different shading distribution. The distributions are mathematically formulated in Appendix I.

### 1.1 *Shadows*

In certain regions the local surface is not exposed to the source, resulting in "shadows" within which $t_e$ is zero. These shadows are bounded by two lines, the shadow-casting profile and the shadow edge. The shape of the shadow edge depends both on the shadow profile and on the topography in the vicinity of the edge and in consequence interpretation of shadowing is

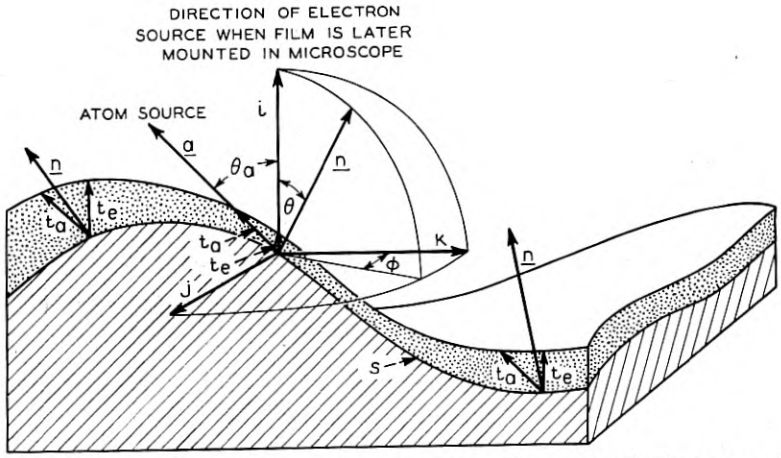



Fig. 1—Diagrammatic representation of thin film replicas produced when every atom sticks where it strikes.

difficult unless one of the surfaces is smooth.[10] With multiple or extended sources, partial shadows and shading due to partial shadowing occur.

### 1.12 *Negative shadows*

There may be other regions of the surface which are exposed to the atom source, but not to the subsequently incident electron beam. When the replica film is mounted in the microscope, these regions appear reentrant to the electron beam, which must pass through three rather than one layer of replica. These regions appear as very lightly exposed areas in micrographs, and since they are essentially "negative shadows" are subject to the same considerations of shape and interpretation as ordinary shadows.

[10] "Physical Methods in Chemical Analysis," Vol. 1, 1950. On pp. 571–3, R. D. Heidenreich reports some unpublished work of S. G. Ellis and W. G. Gross, showing great differences in appearance of shadow-cast replicas as the azimuth of the atom source is varied.

## 2. Intrinsic Resolution of Replica Films

In the vicinity of a sharp change in surface gradient, the local thickness $t_e$ does not change abruptly, but the change occurs over a short distance $d$ as illustrated in Fig. 2. A mathematical formulation for this intrinsic resolution $d$ is given in Appendix II, with some detailed discussion. Intrinsic resolution varies locally over the surface, dependent on the geometrical relationship between atom source, electron beam, and local surface topography. *Observable* resolution includes also instrumental resolution and contrast factors. Except at shadow boundaries it is probably never less than $\frac{1}{2}\bar{t}_e$, where $\bar{t}_e$ is the average value of $t_e$, and perhaps may be as poor as

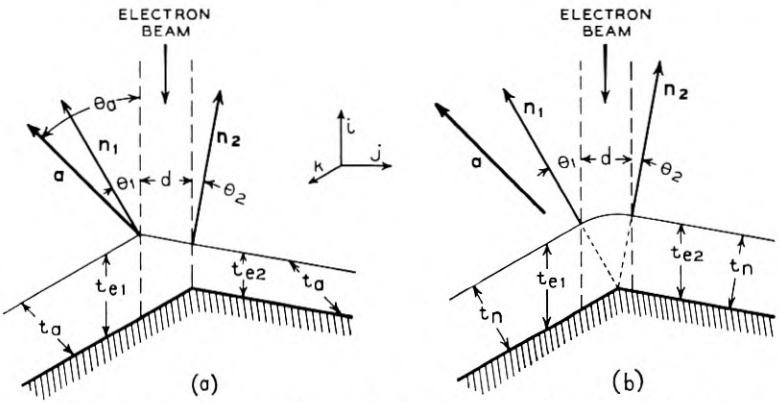



Fig. 2—Diagram showing resolution of replica film at a sharp corner. (a) Every atom sticks where it strikes. (b) All atoms diffuse, finally condensing into film of uniform local thickness.

several times $\frac{1}{2}\bar{t}_e$. Shadow profiles and edges often show short lengths of extreme sharpness. The resolution across these portions of shadows may often be assumed to be the instrumental resolution. The reasons for this are developed in the Appendix.

### 2.1 *Effect of Film Thickness*

The average linear thickness $\bar{t}_e$ is a factor in the expressions for resolution. To reduce $d$ and thereby improve the resolution, the most effective method is to reduce $\bar{t}_e$. With a given material, as the replica film is made thinner, the contrast is reduced also, so that, just as in photography a feature that is visible in a properly exposed negative may be lost in a thin negative, some features may not be replicated with sufficient contrast to be detectable. For 50 kv electrons, the *optimum* average thickness of a replica lies between 5 and 10 $\mu g/cm^2$. This is 7–15 times the minimum perceptible thickness

difference, a rule applicable for electrons of any energy. The local thickness of the replica varies from a fraction of to several times its average thickness, and a replica of near optimum thickness provides a range of 20–40 detectably different shades of contrast. Although for special purposes a replica film may be made thinner or thicker, usually the average thickness should be selected in the optimum range. The resolution is then determined by its density, or more precisely, by its electron scattering power. Because the denser materials are composed of elements of higher atomic number which are more efficient scatters, scattering power† tends to increase rather faster than density, but the difference is not sufficiently great to invalidate for

TABLE I
RESOLUTIONS OF REPLICA FILMS
10 $\mu g/cm^2$

| Film Material | Density | $\bar{l}_e$ | Resolution $\frac{1}{2}\bar{l}_e$ |
|---|---|---|---|
| Plastic | 1 | 1000 Å | 500 Å* |
| Silicon | 2.4 | 416 | 208 |
| Silica (Silicon Monoxide) | 2.5 | 400 | 200 |
| Aluminum | 2.7 | 370 | 165 |
| Aluminum oxide | 3.7 | 270 | 135 |
| Germanium | 5.4 | 185 | 92 |
| Chromium | 6.9 | 144 | 72 |
| Gold 50%**, Manganin 50% | 14 | 72 | 36 |
| Gold 67%**, Manganin 33% | 16 | 62 | 31 |
| Uranium | 18.7 | 50 | 25*** |
| Gold | 19.3 | 50 | 25 |
| Platinum | 21.5 | 44 | 22 |

* Included for comparison only.
** By volume.
*** On exposure to air, *U* oxidizes.

the present purposes the assumption that the two are proportional. The existence of an optimum average mass thickness then implies that *intrinsic resolution of replica films is inversely proportional to the density of the material of which they are composed.*

Table I presents a comparison of the resolutions associated with various materials, based on the assumption that the resolution is $\frac{1}{2}\bar{l}_e$. As discussed above, this is about the best observable resolution, for favorable topographic features. The resolution of plastic films is not susceptible to calculation, and is probably greater than indicated. The resolutions of evaporated films decrease from about 200 Å for silicon and silica to about 25 Å for the very heavy metals. However, it is difficult to process the exceedingly thin films of these metals particularly if they recrystallize as does gold. Although

† Ref. 1, Chap. 19 and p. 158. See also C. E. Hall, *Jl. App. Phys. 22*, 658, 1951.

gold-manganin films are only slightly thicker, they are not particularly difficult to process, especially if 500-mesh supporting screen is used.

### 2.2 *Granularity*

Resolution is also affected by the short-range migration which culminates in recrystallization of many metallic films.[11, 12, 13, 14] If the crystallite size is smaller than the resolution, i.e. less than $\frac{1}{2}l_e$, this effect is not too important, even though the granularity may be objectionable at high magnification from an esthetic viewpoint. A fairly extreme example of recrystallization is shown by the aluminum replica in Fig. 9.

A second source of granularity is due to properties of plastics when plastic molds are used as intermediate replicas. Because plastic molecules are large, and because they associate into domains,[15] the plastic surface is actually granular on a scale of the order of 100 Å. Plastic granularity is not observed in silica replicas because of insufficient resolution, but it becomes very evident in shadow-cast replicas on account of the near-glancing incidence of the shadowing material.[16, 17] The occurrence of granularity due to this cause in replica films of denser materials is an indication of their good resolution. Since this granularity is real on the plastic surface, it shows clearly the azimuth of the incident atom beam, whereas granularity due to recrystallization shows no directional effect.

## 3. Experimental Observations

The foregoing material presents a rather idealized picture of the process of replica film formation by condensation of inorganic substances evaporated under good vacuum, i.e. at pressures preferably less than $10^{-5}$ mm, and certainly not greater than $10^{-4}$ mm of mercury. Subsequent to film formation in the vacuum, it must be subjected to gross physical and chemical processing to prepare it for electron microscopic examination. It must be exposed to air, which may cause oxidation. Uranium films, for example, appear to oxidize completely, and it is believed that SiO films oxidize to $SiO_2$. Most metal films yield good electron diffraction patterns characteristic of the metal, although this does not preclude the possibility of surface oxidation, since the oxides are usually amorphous and diffuse rings due to thin oxide layers would be difficult to detect. Then the films must be sepa-

[11] R. G. Picard and O. S. Duffendack, *Jl. App. Phys., 14,* 291 (1943).
[12] H. A. Stahl. *Jl. App. Phys., 20,* 1, (1949).
[13] H. Levinstein, *Jl. App. Phys., 20,* 306 (1949).
[14] R. S. Sennett and G. D. Scott, *Jl. Opt. Soc. Am., 40,* 203 (1950).
[15] C. C. Hsiao and J. A. Sauer, *Jl. App. Phys., 21,* 1070 (1950).
[16] R. C. Williams and R. C. Backus, *Jl. App. Phys., 20,* 98 (1949).
[17] Metallurgical Applications of the Electron Microscope, p. 11, *Symp. of Inst. of Met.,* November 1949.

rated from the surface replicated, usually by dissolution of the latter. This process subjects the film to considerable strain. Finally it must be removed from the solvent, usually on a piece of 200-mesh screen, rinsed at least once, and finally allowed to dry. It is not surprising that films sometimes

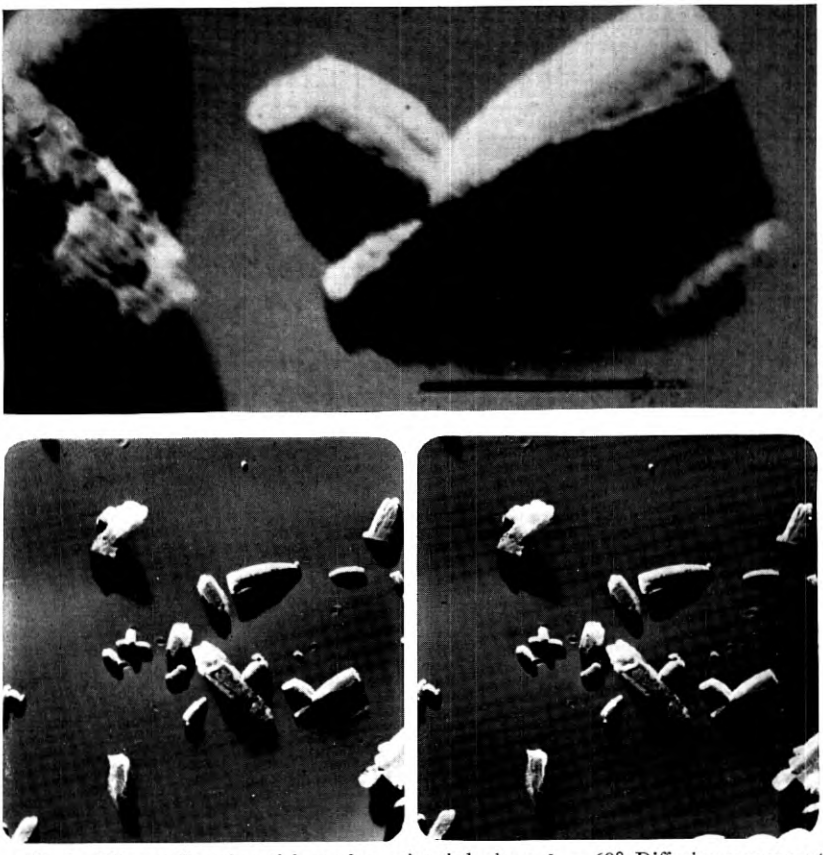


Fig. 3—Silica replica of particles and associated shadows. $\theta_a = 60°$. Diffusing component replicates particles within shadows. Oval hole shows presence of film in shadow. Black lines = 1 $\mu$.

exhibit cracks and holes. Stereoscopic examination of good replicas show that despite all the processing violence, a faithful picture of surface topography is obtained, at least for features up to a few microns in size.

The nature of silica condensation is indicated by the micrographs of Fig. 3, of a silica replica of particles of an alkaline earth carbonate. These were dispersed on a plastic-coated microscope slide, and silica evaporated at

$\theta_a = 60°$. The plastic film was then "floated off" on water and picked up on 200-mesh screen; the plastic and particles were successively dissolved, leaving the silica replica.† The micrographs clearly show (a) shadow-edges and (b) a film in the shadows. The enlargement shows a shadow within which there exists a hole in the film and replication of completely shadowed particles by the diffusing component. It follows from (a) that a part of the incident material must stick where it strikes, and from (b) that a part must diffuse into the shadows, somewhat in the manner diagrammatically illustrated in Fig. 4. Densitometer traces through shadow edges show that the film thins down a little as the edge is approached from the unshadowed region, drops more or less abruptly at the edge, and continues to thin down within the shadow as illustrated. Now the diffusing part must finally con-

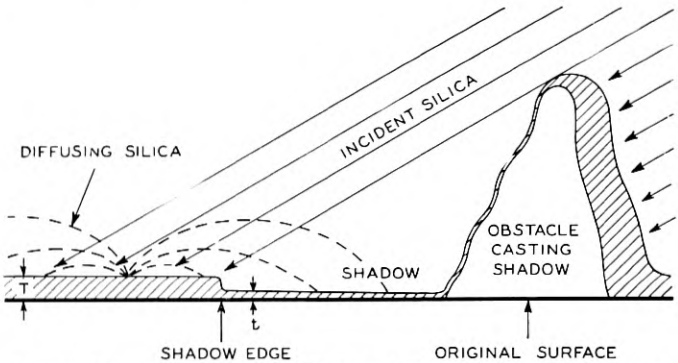



Fig. 4—Diagram illustrating diffusion of silica into shadowed regions.

dense, and it is natural to assume that at each collision with the surface the probability of sticking is $\alpha$, and of diffusing is $(1 - \alpha)$, and that the average molecule travels between collisions a small distance. Analysis of densitometer curves on these assumptions leads to a value of $\alpha$ of about $\frac{1}{2}$, and a range of about $\frac{1}{2}\ \mu$, with a rather wide spread of values.[18] However, these assumptions would require the film to be vanishingly thin at distances more than $3\ \mu$ from the shadow-edge. In fact, an *extremely* thin film is found at even greater distances. The probability of sticking upon collision, and the distance a molecule moves as a two-dimensional gas molecule between collisions with the surface, thus must be assumed to depend on such factors as angle of impingement, energy of molecule, and perhaps the nature of the surface. This latter initially is a plastic, probably covered

† In particle study, a second evaporation of silica 180° in azimuth from the first produces a "thin-shell" replica more suitable for stereoscopic study of the particles.

[18] C. J. Calbick, *Jl. Appl. Phys.*, *19*, 119 (1948).

with adsorbed gas, and later during the deposition is freshly condensed silica. Films are also found in shadows of replicas made of other materials such as silicon monoxide, germanium, gold-manganin, and even chromium. Whether the interpretation should always be the same as for silica, or whether in some cases the diffusing component is different from the sticking component, is uncertain.

A high degree of contrast is commonly attributed to silica replicas, which are supposedly deposited at near-normal incidence. In the writer's experience, distortion of the conical-tungsten-basket silica evaporator often results in values of $\theta_a$ greater than 10°. If normal incidence, characterized by absence of shadows, is actually attained, the resulting replica exhibits

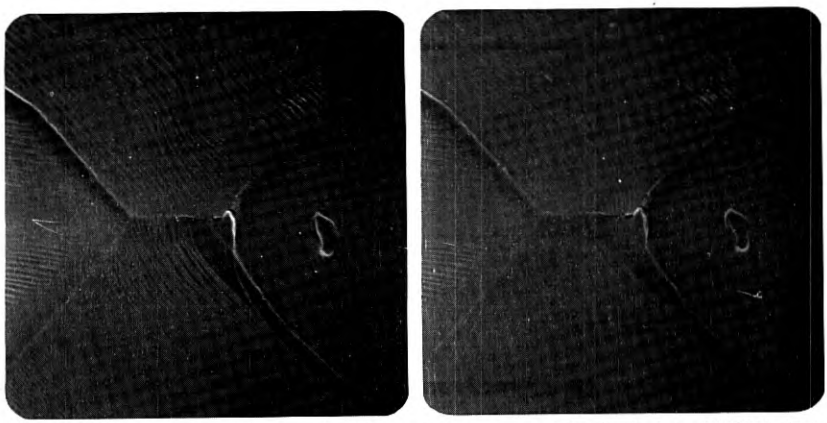


Fig. 5—Silica replica of natural surface of thermistor flake heated to 1425°C. Note difference in contrast between the two pictures of the stereogram.

poor contrast unless steep slopes are present. An example is shown in Fig. 5, in which difference in contrast due to the fact that $\theta_a$ differs for the two pictures of the pair by approximately 8°, the stereoscopic angle, is evident. Note that the shadow in the surface feature in the center of the grain at the right is more pronounced in the left-hand picture which shows the greater contrast. The replica is from the surface of a thermistor flake, sintered briefly at 1425°C.

Figure 6 is from a silica replica about 400 Å thick, of a portion of the surface of a thermistor disk sintered 6 hrs. at 1175°C. The white line is 0.1 $\mu$ long. The striations show a minimum separation of about 250 Å, and the character of the shading indicates that this is near the limit of resolution of the micrograph, about 200 Å (Table I). Higher resolutions claimed[5, 9] are probably due either to shadow effects, to special situations on the sides of steep slopes, or perhaps to the use of extremely thin replicas.

The natural surfaces of sintered thermistor flakes, prepared by heating in air thin (10 $\mu$) sheets of a mixture of NiO and $Mn_2O_3$ powders, exhibit well defined planes of sizes suitable for electron micrographic study.[19] Flakes sintered briefly at 1175°C were selected as suitable objects for experimental study of replication. They were molded into the surface of lucite blocks at 150–160°C and 2500 lb/in$^2$. They were then dissolved in HCl,§ and replica

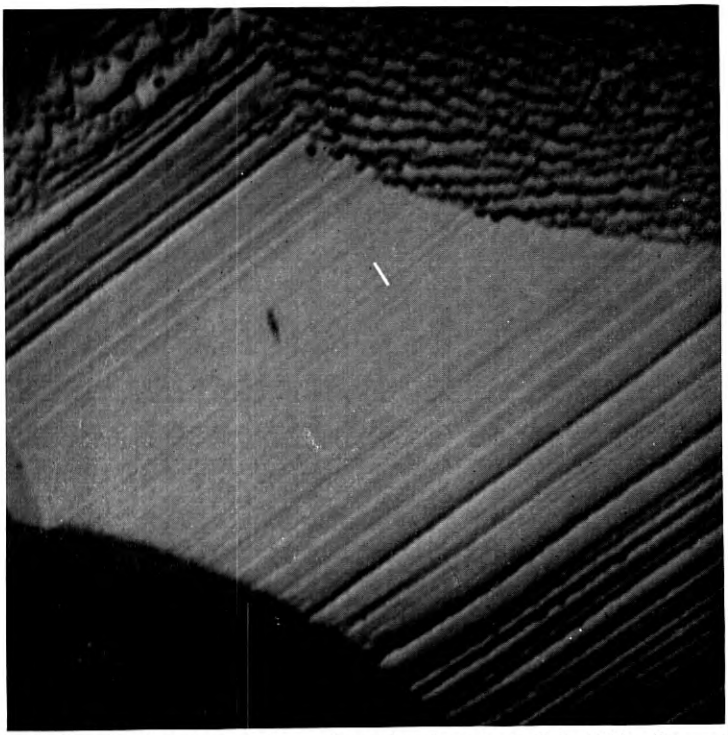


FIG. 6—A striated region on the surface of a thermistor disk. White line = 0.1 $\mu$. Striations are near the limit of resolution of the silica replica, which is about 400Å thick.

films deposited on the plastic molds at pressures not greater than $10^{-4}$ mm (usually about $2 \times 10^{-5}$ mm). The replica surface was then scored into small (about 1.5 mm) squares and immersed in ethyl bromide.¶ In a few minutes the replica films drift free and are then "fished" from the solvent

[19] H. Christensen and C. J. Calbick, *Phys.Rev.*, *74*, 1219 (1948).

§ A one-step process was precluded because some of the replicating materials are soluble in HCl.

¶ Ethyl bromide is not a good solvent for lucite, while chloroform is. Extended tests have produced better results when a poor solvent, which perhaps frees the replica by creeping between it and the plastic mold without appreciable dissolution, is used.

on pieces of 200- or 500-mesh screen. After drying, the screen is immersed in chloroform to remove the last traces of plastic, and is then ready for electron microscope use.

The electron micrographs of Figs. 7–11 are presented as stereoscopic pairs and enlargements of selected areas, the scale being given by dark lines of 1 $\mu$ length. Figure 7 is a micrograph from a chromium replica[20] for which

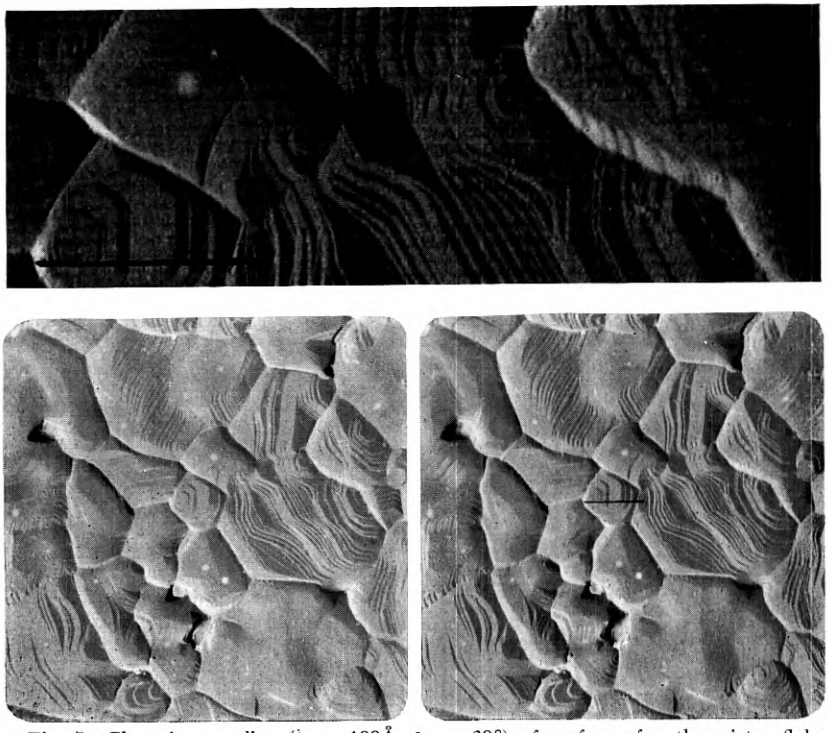


Fig. 7—Chromium replica ($t_e$ = 100Å, $\theta_a$ = 30°) of surface of a thermistor flake. Resolution about 50Å. Granularity is due to plastic mold.

$t_e$ = 100 Å, $\theta_a$ = 30°. The shadows show that the azimuth of incidence was to the right at about 30° above the horizontal. The granularity evident in the enlargement shows evidence of this direction, and may be ascribed to plastic granularity, except for a few of the larger hills or pits which are

[20] J. Ames, T. L. Cottrell and A. M. D. Sampson, *Trans. Far. Soc. 46*, 938 (1950). This paper, which appeared while the present paper was in preparation, exhibits micrographs of chromium and other metallic replicas of surfaces of crystals grown from solution. The characteristic surface structures reported are in some ways similar to those of the sintered thermistor flakes here shown.

probably due to features of the original surface. Resolution ranges upward from about 50 Å.

The detection of films in shadows is difficult when these films are so thin as to approach the minimum perceptible thickness difference. The film may be flawless and present, or ruptured during processing and completely eliminated, and the difference between these two conditions is difficult to

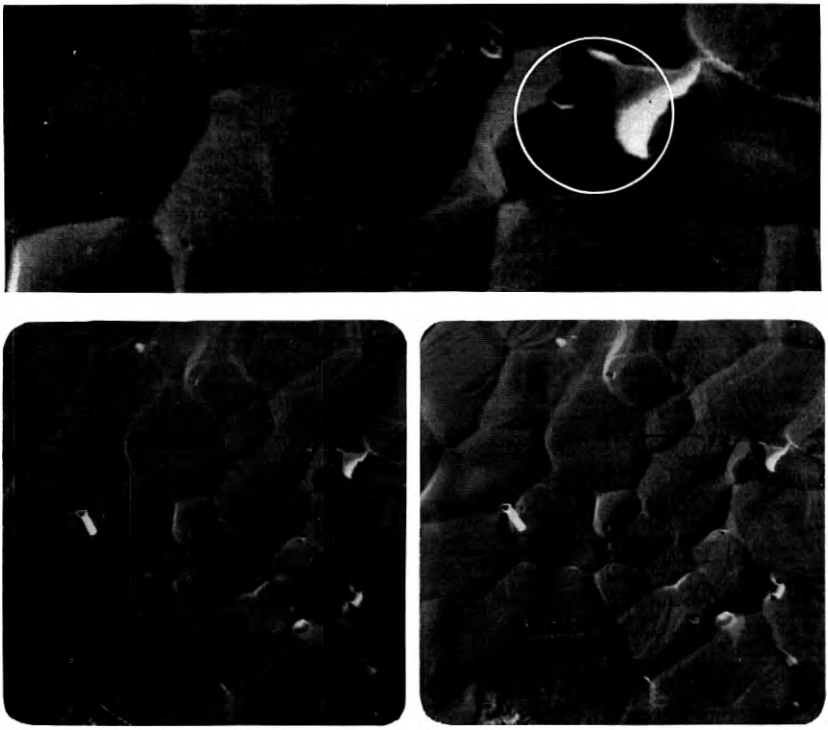


Fig. 8—Chromium replica ($t_e$ = 150Å, $\theta_a$ = 30°) produced by two evaporations differing in azimuth by 90°. Region within the circle shows partial shadows.

discern. Only when structure or edges due to partial rupture appear is it easily detectable. Some shadows in the replica from which Fig. 7 was made showed such a film, which was estimated to be less than 10 Å in thickness, but in most of the shadows no evidence of a film could be seen.

A chromium replica produced by two evaporations, from azimuths 90° apart, was the subject of the micrograph of Fig. 8. This replica was not washed in chloroform, which accounts for the presence of the residual plastic rings. This particular micrograph was selected to show the partial

shadowing effect shown clearly in the enlargement, but is not suitable for determination of resolution or exhibition of plastic granularity on account of the slightly imperfect focus, noticeable in the enlargement. Because the replica was about 150 Å thick, this granularity was scarcely evident even in a perfectly focused micrograph, from which the resolution was estimated as ranging upward from about 100 Å. It should be observed that the direction

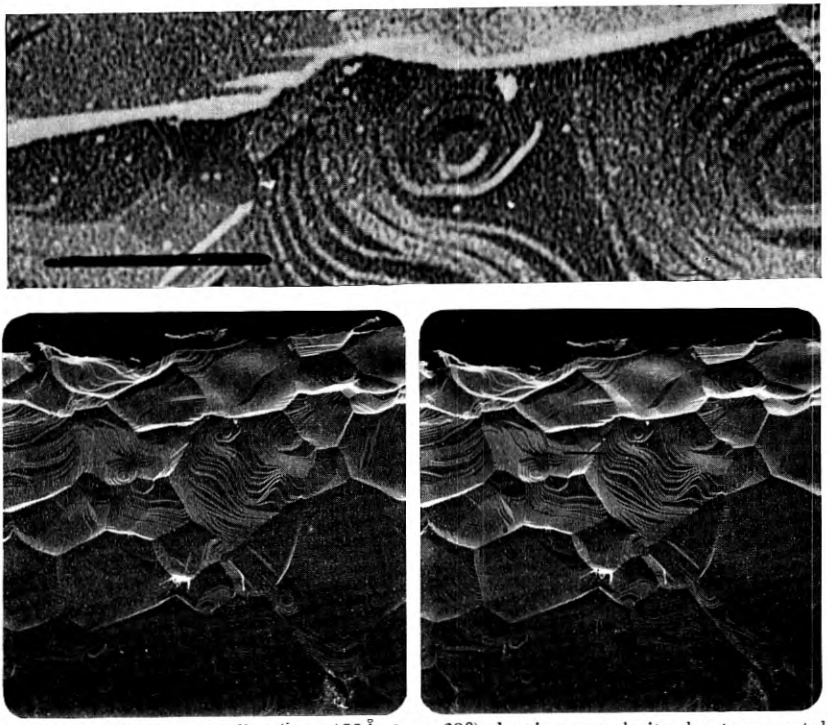


Fig. 9—Aluminum replica ($l_e$ = 180Å, $\theta_a$ = 30°) showing granularity due to recrystallization. Local curling of replica has resulted in "negative shadows" in upper left region of stereogram. Resolution about 100Å.

of maximum contrast in shading is as if a single source were toward the lower right, about midway between the two actual sources whose azimuths can be determined in the stereogram.

The appearance of a micrograph of a replica which has undergone recrystallization on a scale comparable to $l_e$ is shown in Fig. 9. The replica was of aluminum; $l_e$ was about 180 Å, the lower limit of the optimum range; $\theta_a$ was 30°. The non-directional character of the granularity is evident. The micrograph selected shows an area near a torn edge which has curled up-

ward, thus tilting the replica to various angles. In the extreme upper left, this tilt produces *negative* shadows, regions where the electrons pass through three thicknesses of replica. In none of the more usual positive shadows observed in other parts of the replica was any film observed, or in a replica twice as thick also studied. The conclusion is that aluminum does not diffuse. It is tempting to speculate that the short-range forces responsible

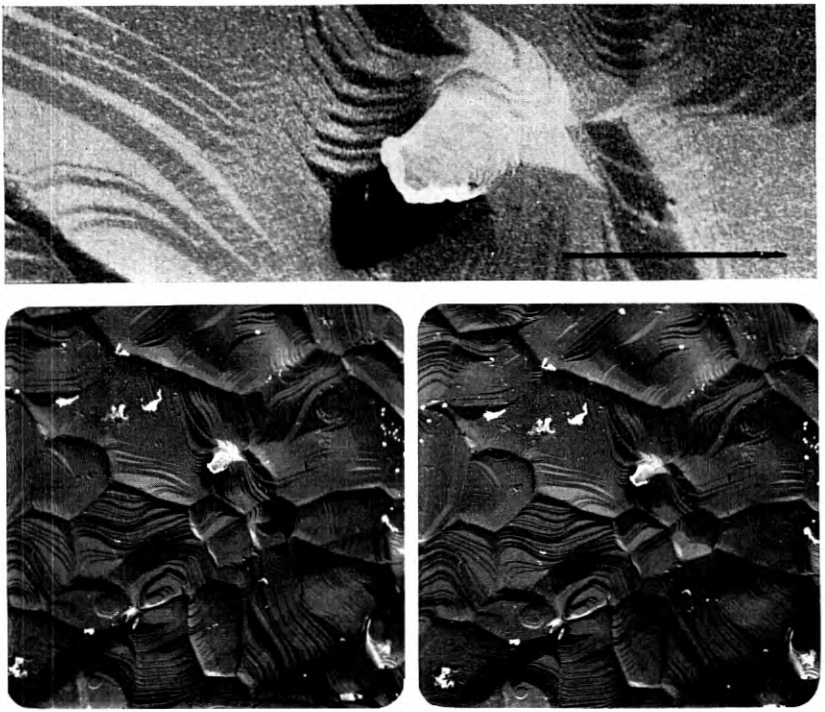


Fig. 10—Gold-manganin replica ($t_e$ = 150Å, $\theta_a$ = 20°). Resolution about 100Å. Replica is much thicker than optimum. A film in the shadow, clearly evident in the original micrograph, does not show except for two whitish areas where it has torn and curled.

for recrystallization do not permit diffusion and, that when diffusion does occur with materials such as chromium, it is due to some other component such as an oxide. Resolution, although complicated by the granular structure, appears to be about 100 Å.

The micrograph of Fig. 10 is from a gold-manganin replica, produced by simultaneous evaporation of two volumes of gold and one of manganin (alloy, 84% Cu, 12% Mn, 4% Ni) at $\theta_a$ = 20°. The total thickness was about 150 Å, so mass thickness, about 24 $\mu g/cm^2$, was much greater than the

optimum. Crystallite size is probably less than 50 Å. As in Fig. 10, the granularity is probably due to plastic. A thin, torn film appears in the shadow. From the extreme sharpness of a portion of one edge of the shadow, although slightly complicated by granularity, one concludes that instrumental resolution is better than 50 Å. Replica resolution is about 100 Å. Despite the smaller value of $\theta_a$, the thick replica provides a greater range of tone

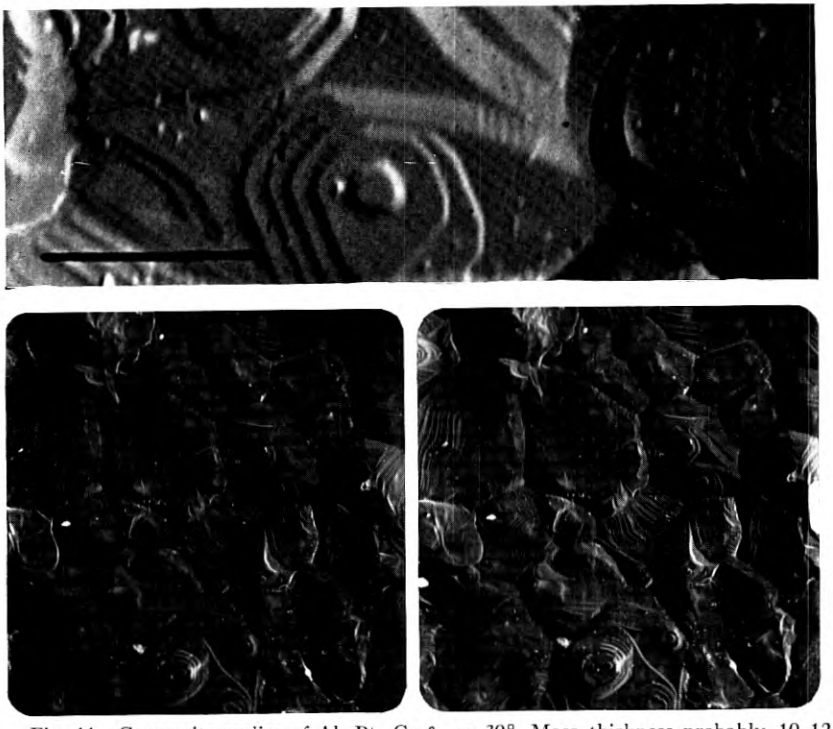


Fig. 11—Composite replica of Al, Pt, Cr. $\theta_a = 30°$. Mass thickness probably 10–12 $\mu g/cm^2$. Resolution about 100Å.

values, as compared with Fig. 7, although this may not be evident in the reproductions.

Figure 11 shows a micrograph from a composite replica of aluminum, platinum, and chromium. Aluminum and platinum were simultaneously evaporated from a tungsten wire which burned out before the evaporation was complete. If all the aluminum evaporated before the platinum, its thickness was about 50 Å. The amount of Pt is problematical, but the chromium, evaporated after the tungsten wire was replaced, had a thickness of about 100 Å. The same angle, $\theta_a = 30°$, was used in the two evaporations.

The granularity barely discernible in the enlargement, which again shows a slight imperfection in focusing, is attributable to plastic. Resolution is perhaps 100 Å.

Metallic film replicas could easily be used to replicate surfaces in sealed-off tubes. For example, chromium can be plated on a tungsten wire which is suitably mounted in the tube, and thoroughly outgassed during pumping, the surface to be replicated being shielded from the Cr source during the process. Later it is evaporated to form the replica. As an illustration, Fig. 12 is from a Cr replica of the activated surface of an oxide-coated cathode. It was actually prepared by evaporation at a pressure of $2 \times 10^{-6}$ mm and not in a sealed-off tube, but the suggested technique is certainly practicable. No films were observed in shadows in this replica, which was less than

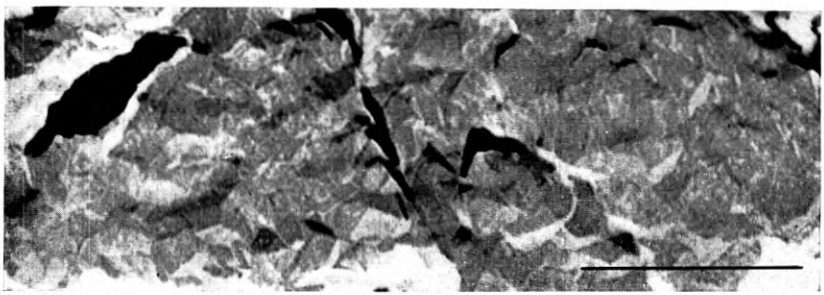


Fig. 12—Chromium replica ($t_e$ = 100Å, $\theta_a$ = 30°) of surface of an oxide-coated cathode. One-step replica. Whitish areas are due to impurity in the oxide (probably silica) redeposited on the replica.

100 Å thick with $\theta_a = 30°$. However, the film was very flimsy and exhibited a large number of cracks, usually originating in shadows (e.g., the elongated black area). The film was freed from the surface by dissolving the oxide in dilute acid; since no plastic mold was involved, the fine scale features are characteristic of the oxide surface. The whitish areas (near bottom) are due to some impurity in the oxide redeposited on the replica, probably silica from the ball-milling process to which the original barium carbonate powder had been subjected. Replica resolution is perhaps 100 Å, although suitable features to test higher resolving power are not present. Shadow edges indicate instrumental resolution less than 50 Å.

In Figs. 13 and 14 the use of germanium as a replicating material is illustrated. Germanium is easily evaporated from a conical carbon crucible supported and heated by a conical helix of tungsten. As in the case of silica and silicon monoxide, the thickness of the resulting film is only roughly known. For the replica of Fig. 13, 2 mg of Ge at 8 cm distance, $\theta_a = 30°$,

was used, and $\bar{l}_e$ is estimated to lie between 200 and 300 Å. For that of Fig. 14, 1 mg at 8 cm with $\theta_a = 35°$ was used and $\bar{l}_e$, between 100 and 150 Å, is in the optimum thickness range. The micrograph has the appearance of a correctly exposed photographic negative, whereas Fig. 13 resembles an over-exposed negative. Since germanium films are amorphous unless heated to temperatures higher than 300°C, the fine structure is perhaps due to

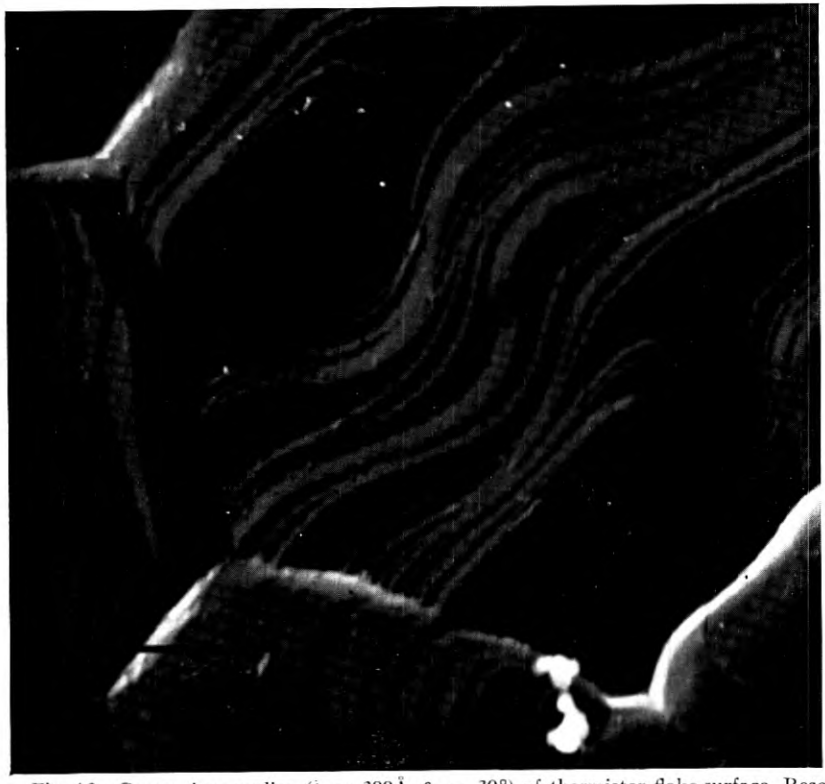


Fig. 13—Germanium replica ($\bar{l}_e \simeq 300$Å, $\theta_a = 30°$) of thermistor flake surface. Resolution about 150Å. Over-exposed appearance shows replica is thicker than optimum.

plastic granularity, although some features are probably real in the thermistor flake surface. Resolutions are perhaps 150 Å (Fig. 13) and 75 Å (Fig. 14). A film clearly appears in the shadow in Fig. 14. Germanium shadow films are relatively thinner than silica, indicating that $\alpha$ is greater than 0.5 (perhaps 0.7 to 0.8), but thicker than chromium or gold-manganin shadow films.

Germanium has many advantages as a replicating material. It is easily

evaporated; the films are substantially amorphous; they are fully as rugged as silica films in processing and, in contrast to silica films, are easily seen during the "fishing" part of this procedure. Because they are conducting and do not tend to charge up in the microscope, germanium films are more stable than silica. Finally, and most important, because germanium is

Fig. 14—Germanium replica ($t_e \simeq 150$Å, $\theta_a = 35°$) of near optimum thickness. Resolution about 75Å. Ruptured film evident in shadow.

twice as dense as silica, the intrinsic resolution (Table I) is better by a factor of two. Many of these remarks apply also to chromium and gold-manganin replicas; however, they are not amorphous and are less rugged in processing, even though they are perhaps even more stable in the microscope. Stereoscopic pairs from germanium replicas are not shown only because it is desired to present the more extensive enlargements of Figs. 13 and 14.

Figure 15 shows electron diffraction patterns obtained from the several replica films, indicating that crystallite size ranges from greater than 100 Å for aluminum down to almost amorphous for germanium and amorphous for silica.

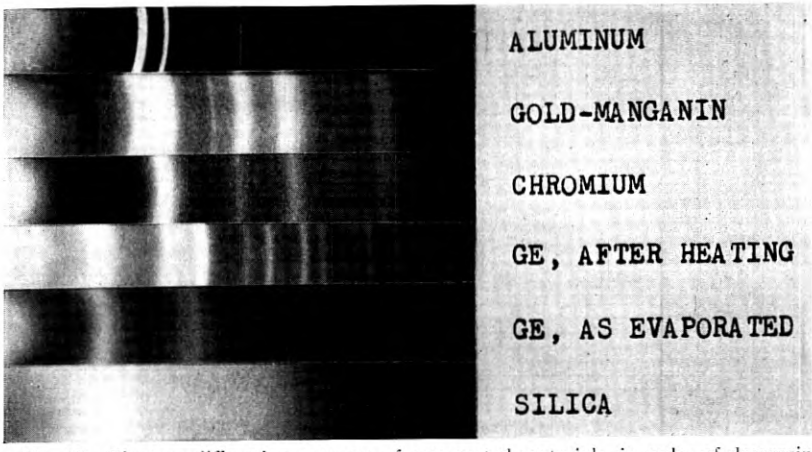



Fig. 15—Electron diffraction patterns of evaporated materials, in order of decreasing crystallite size. The aluminum pattern is due to crystallites of about 100Å. The gold-manganin pattern shows at least two sharp rings not due to gold and diffuse bands due to very small crystallites with the structure of gold. The chromium pattern is due to very small crystallites of Cr. The pattern of a germanium replica film after heating in air to 390° shows partial recrystallization, with rings due to quite small crystallites superposed on the amorphous pattern of the Ge as evaporated.

### 4. Interpretation of Micrographs of Replicas

The electron image pattern due to the replica is reproduced by the exposure of a photographic plate. The complex problems associated with photographic reproduction[21] cannot be discussed here. For a number of reasons, blackening of the plate is not linearly related to replica film thickness. The number of electrons scattered out of the beam* is proportional to thickness only as a first approximation, the blackening vs. exposure curve of the plate is not linear, and there is also a roughly uniform background due to inelastically scattered electrons. At the lower electronic magnifications, field distortion is also a factor affecting local intensity in the electron image. Furthermore, the geometrical relation between the electron beam incident upon the thin mesh-supported film and the atom-beam is usually not known accurately, and indeed may vary locally over the replica in the manner of which Fig. 9 is an extreme example. In con-

[21] W. T. Wintringham, *Proc. I. R. E.*, *38*, 1284 (1950).
* Ref. 1, ch. 19 or ref. 7, p. 541.

sequence of these factors, *the precise interpretation of density variations in micrographs is not practicable.* Also many replicas contain artifacts, i.e., features in the micrograph not due to the original surface, but introduced somewhere in the processing. Recrystallization, plastic granularity, specks of dirt or other foreign material, tears in the replica films, and defects in the photographic emulsion are examples of artifacts. Few micrographs are completely free of these effects.

The interpretation of micrographs therefore has as its object not so much a detailed topographical map of the surface, but rather its characteristic features, repeated in many micrographs. For example, the micrographs presented show that sintered NiO-$Mn_2O_3$ flakes develop grains with extensive crystallographic planar surfaces but that thick disks develop a striated or hill-and-valley surface structure on individual crystallites. In both cases the structure is very compact, pores between grains being almost nonexistent. (Incidentally, these materials can be subjected to a heating cycle in which pores are a predominant feature.) Naturally, the greater the range of contrast and the better the resolution, the more surely can characteristic features on an exceedingly fine scale be detected. In general, the method of replication which portrays best the characteristic features under study should be selected. Even in the study of a single material, more than one method may be desirable. For example, a porous structure is probably most easily reproduced by a silica replica using the two-step process, on account of the fact that the diffusing component forms films over reentrant regions not exposed directly to the source; but fine surface detail might best be revealed by a germanium or chromium replica using the one-step process.

## 5. Stereoscopy

Electron microscopy has a fundamental advantage in that, because of great depth of focus, stereoscopic study of surfaces at high magnification is possible. This advantage is sometimes indispensable; for example, porous structures result in complex micrograms which can be understood only by stereograms. More generally, *stereographic portrayal, by fully delineating surface topography, achieves the chief purpose of microscopy and makes unnecessary the precise interpretation of density variations.* However, resolution and contrast are important factors in stereograms.[22] It is obvious that *the replica must retain the third dimension; inorganic replicas in general do,* but thin film plastic replicas, unless heavily shadow-cast to make them effectively inorganic, change under the electron bombardment and draw down to a planar film of variable thickness.[23]

[22] A. W. Judge, "Stereoscopic Photography," p. 28.
[23] C. J. Calbick, *Jl. App. Phys., 19,* 1186 (1948).

Many artifacts are also detectable by stereoscopy. Since two pictures of the same field are available, photographic emulsion flaws are readily detected. The three-dimensional view also helps to identify many other artifacts, such as foreign material present on the replica or local wrinkling of the replica film.

Unfortunately, half-tone reproductions are not suitable for stereograms, because half-tone detail is objectionably enlarged by the viewing stereoscope. Despite this, some idea of the value of stereograms may be obtained from the figures.

## Conclusions

A unified picture of replication by evaporated films has been presented. Thin-film replicas may be made by any material which can be evaporated *in vacuo* and whose physical and chemical properties are suitable. For good contrast, a considerable angle should separate the directions of incidence of atom- and electron-beams. The intrinsic resolution of the replica is about half the film thickness and is therefore inversely proportional to the density of the replicating material. Multiple point sources and sources of extended area are equivalent from a shading standpoint to a single point source properly placed. The oxides SiO and $SiO_2$, and presumably many others, form amorphous films, whereas the metals tend to recrystallize although the crystallite size may be less than 50 Å for some metals. Germanium, a semi-metal, forms an amorphous film. Although not dense compared to the heavy metals, it is more than twice as dense as $SiO_2$, and should be valuable as a replicating material because it combines electrical conductivity and high resolving power in an amorphous film and is chemically rather inert. Among the metals, chromium appears to be the most generally useful replicating material. Gold-manganin has sufficiently small crystallite size for many purposes, and is very easy to evaporate. The platinum group suffers from the disadvantage of being very difficult to evaporate.

Finally, all inorganic replica films studied retain the third dimension. Electron stereo-micrograms may be used to reveal three-dimensional topography, largely eliminating the need for correlation of photographic density variations with the surface structure.

## APPENDIX I
## CONTRAST IN REPLICA FILMS

(a) Local Thickness When Every Atom Sticks Where It Strikes.

Referring to Fig. 1, it is evident that:

$$t_c = t_a \frac{\underline{a} \cdot \underline{n}}{\underline{i} \cdot \underline{n}} = t_a \cos \theta_a \, (1 + \tan \theta_a \cos \varphi_a \tan \theta \cos \varphi). \qquad (1)$$

When the principal azimuth is defined by the plane containing $\underline{a}$ and $i$, $\cos\varphi_a = \pm 1$, and, as illustrated, $\cos\varphi_a = -1$. With more than one source, in directions $\underline{a}_1, \underline{a}_2 \cdots \underline{a}_m$ :

$$t_e = \frac{(t_{a_1}\underline{a}_1 + t_{a_2}\underline{a}_2 + \cdots t_{am}\underline{a}_m)\cdot\underline{n}}{i\cdot\underline{n}} \tag{2}$$

and the mass thickness is given by

$$\rho_a t_e = \left(\sum_{j=1}^{m} \frac{\rho_j t_{aj}\underline{a}_j\cdot\underline{n}}{i\cdot\underline{n}}\right) \tag{3}$$

where $\rho_j$ is the density of the material deposited by the source $t_{a_j}\underline{a}_j$, and $\rho_a$ is determined by the equation $\rho_a t_a\underline{a} = \sum_{j=1}^{m} \rho_j t_{aj}\underline{a}_j$, which defines the equivalent single source. The distribution of mass thickness of any evaporated film, when atoms stick where they strike, is therefore given by eq. 1, with values $\theta_a$, $\varphi_a$ determined by the distribution of sources. The replica may be called an *incidence-shaded* replica.

(b) Calculated Thickness When a Fraction of the Incident Atoms Diffuses Over the Surface

Figure 3 has shown that, in the case of silica, a fraction $\alpha$ of the impinging molecules stick where they first strike, the remainder diffusing. If the latter condense into a film of uniform thickness normal to the local surface, the shading equation becomes

$$t_e = t_a \cos\theta_a \left[\alpha(1 + \tan\theta_a \cos\varphi_a \tan\theta\cos\varphi) + (1-\alpha)\frac{A_0}{A}\cdot\frac{1}{\cos\theta}\right] \tag{4}$$

where $A$ is the total surface area and $A_0$ its projection on the $jk$ plane (Fig. 1). Many other materials, and in particular silicon monoxide and germanium, condense likewise with values of $\alpha$ less than unity.

The assumption of uniform local thickness for the diffusing component is not quite correct. More strictly, $A_0/A$ in eq. 4 should be evaluated over areas of the order of the square of the range ($\frac{1}{2}\mu$). Since the range is large compared to the resolution, and since $A_0/A$ can only be estimated in any event, this refinement is of little value. It has been suggested[9] that diffusing silica has a higher probability of condensing in regions of change of gradient. This suggestion is difficult to sustain theoretically; further, observations upon thin shell replicas such as those of Fig. 3 have tended to indicate almost-uniform condensation.

(c) Replica Contrast as Determined by Relative Thickness vs. Colatitude Angle Curves

In any photographic presentation, two densities are detectably different

only if they differ by some small fraction on the density scale. Correspondingly, a finite difference in $t_e$ is required. For 50 kv electrons, the minimum *perceptible* contrast difference is produced by a mass thickness difference of about 0.7 $\mu g/cm^2$. By plotting eq. (4) for particular choices of $\theta_a$, $\varphi_a$ and assumed values of $\alpha$ and $A_0/A$, the geometrical conditions required to pro-

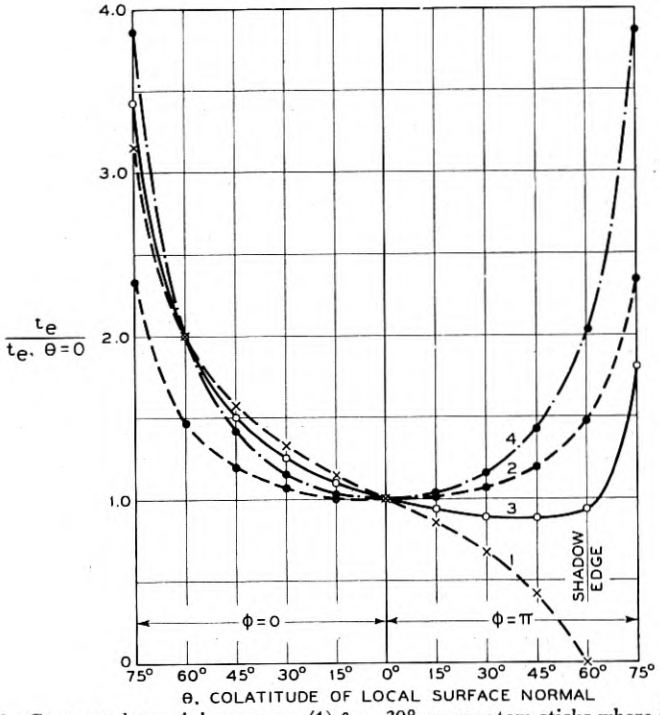



Fig. 16—Contrast-determining curves. (1) $\theta_a = 30°$, every atom sticks where it strikes. (2) $\theta_a = 0°$, half the atoms diffuse. (3) $\theta_a = 30°$, half the atoms diffuse. (4) $\theta_a = 0°$, all atoms diffuse over surface.

duce detectable differences in shading can be displayed. To make the curves more general, relative thickness rather than $t_e$ is plotted in Fig. 16 for azimuth $\varphi_a = 0$. The four curves are, respectively, for

$$(1)\ \theta_a = 30°, \quad \alpha = 1, \qquad (2)\ \theta_a = 0°, \quad \alpha = \tfrac{1}{2},$$

$$(3)\ \theta_a = 30°, \quad \alpha = \tfrac{1}{2}, \qquad (4)\ \theta_a = 0°, \quad \alpha = 0.$$

$A_0/A$ has been chosen as 0.833.

Let us assume that $\rho t_e$ when $\theta = 0$ is 7 $\mu g/cm^2$. Then a thickness difference of 0.1 on the ordinate scale is barely perceptible. For incidence shading,

curve (1) shows that two adjacent planes differing in colatitude angle by not less than 7° in the azimuth of incidence are detectable. For pure diffusion shading,* curve (4) shows that considerably larger differences in angle are required near $\theta = 0$; moreover, planes differing by large angles but symmetrical with respect to the electron beam yield the same $t_e$ and hence, if adjacent, cannot be detected. This difference between diffusion and incidence shading is even more pronounced, because curve (4) is independent of azimuth, whereas for curve (1), $(1 - t_e/t_{e,\theta=0})$ is proportional to $\cos\varphi$. Although this introduces very low contrast in azimuths near $\varphi = 90°$, and also the possibility that two adjacent planes differing considerably in azimuth may yield the same density, incidence shading remains in general more favorable for portrayal of surface topography.

The intermediate curves (2) and (3) for $\alpha = \frac{1}{2}$ correspond to the case of condensation of silica at $\theta_a = 0°$ and 30° respectively. It is evident that, at normal incidence, appreciable contrast occurs only on the steeper slopes ($\theta > 45°$), and that *incidence at an angle is much to be preferred for general surface portrayal.*

## APPENDIX II
## RESOLUTION OF REPLICA FILMS

### a) Incidence Shading

Figure 2 (a) shows, for the case of pure incidence shading, the replica film formed at the intersection of two surfacial planes with normals $\underline{n}_1$ and $\underline{n}_2$. The distance $d$ over which the thickness changes from $t_{e1}$ to $t_{e2}$ is

$$d = t_a \sin\theta_a. \tag{5}$$

The figure is drawn for the case $\varphi_2 = 0$, $\varphi_1 = \pi$. More generally, if $\beta$ is the azimuth of the line of intersection of the planes, it may be shown that

$$d = t_a \sin\theta_a \,|\sin\beta|. \tag{6}$$

The angle $\beta$ is easily measured on a micrograph.
In vector notation,

$$d = t_a \left| \frac{(\underline{n}_1 \times \underline{n}_2) \times i}{|(\underline{n}_1 \times \underline{n}_2) \times i|} \cdot \underline{a} \right|$$

Or, in polar coordinates,

$$d = t_a \sin\theta_a \frac{|\tan\theta_1 \cos\varphi_1 - \tan\theta_2 \cos\varphi_2|}{(\tan^2\theta_1 + \tan^2\theta_2 - 2\tan\theta_1 \tan\theta_2 \cos(\varphi_2 - \varphi_1))^{\frac{1}{2}}}$$

* Aluminum or other oxide replica films formed by surface oxidation presumably correspond closely to the case $\alpha = 0$, which does not occur in evaporated films.

Using eq. (1)

$$d = \frac{| t_{e1} - t_{e2} |}{(\tan^2 \theta_1 + \tan^2 \theta_2 - 2 \tan \theta_1 \tan \theta_2 \cos (\varphi_2 - \varphi_1))^{\frac{1}{2}}} \tag{7}$$

For a given replica film the *intrinsic resolution* $d$ varies simply as the sine of the azimuth of the line of intersection. For example, if $\theta_a = 30°$, $d$ varies from $\frac{1}{2}t_a$ down to zero. But *observable* resolution is closely connected with contrast, as is indicated by eq. (7). One could not hope to *observe* good resolution if present, unless the thickness difference were at least two or three times the minimum perceptible thickness difference. In micrographs, because of the greater *contrast* characteristic of the azimuth $\beta \simeq 90°$, the resolution always appears best in the vicinity of the plane of incidence of the atom beam. The average thickness in the direction of the electron beam is $\bar{t}_e = t_a \cos \theta_a$ and hence:

$$d = \bar{t}_e \tan \theta_a | \sin \beta | \tag{8}$$

For $\theta_a = 30°$, $\sin \beta = 1$, $d = 0.577 \bar{t}_e$. If an attempt is made to improve resolution by decreasing $\theta_a$, contrast is again reduced. It may be concluded that *the observable intrinsic resolution of an incidence-shaded replica film is not less than half the average thickness* $\bar{t}_e$.

b) Diffusion and Combined Shading

With pure diffusion shading,

$$d = t_n [\sin^2 \theta_1 + \sin^2 \theta_2 - 2 \sin \theta_1 \sin \theta_2 \cos (\varphi_2 - \varphi_1)]^{\frac{1}{2}} \tag{9}$$

Figure 2(b) illustrates this case for $\varphi_1 = \varphi_2 + \pi$. From eq. (4), for $\alpha = 0$, $| t_{e1} - t_{e2} | = t_n \left| \frac{1}{\cos \theta_1} - \frac{1}{\cos \theta_2} \right|$, since $t_n = t_a \cos \theta_a \frac{A_0}{A}$. Study of these two equations, for resolution and contrast respectively, shows that resolution is highly dependent on the relation of the local directions $\underline{n}_1$ and $\underline{n}_2$ to the incident electron beam: Values of $d$ observable from a contrast standpoint frequently exceed $\bar{t}_e$, and are less than $\frac{1}{2}\bar{t}_e$ only when the two intersecting planes are fairly steep ($\theta_1$ and $\theta_2 > 60°$) and are not too far apart in azimuth ($\varphi_2 \simeq \varphi_1$). *Hence the above conclusion also applies to pure diffusion shading, and consequently to combined shading such as that of a silica replica*, except possibly for special situations on steep slopes.

c) Shadow-edges

Shadow-edges are always due to the atoms that stick where they first strike, even though, as in the case of silica, part of the condensing material diffuses over the surface. Figure 17 illustrates the formation of a shadow-

edge by an obstacle of height $h$, whose profile is normal to the plane of the paper. From a point source, Fig. 17(a), the resolution is $d = l_a \sin \theta_a = \bar{l}_e \tan \theta_a$. More generally, if the profile makes an angle $\beta$ with the plane of the paper, one obtains $d = \bar{l}_e \tan \theta_a \mid \sin \beta \mid$, identical with eq. (8). But now, because the contrast is large, the factor $\mid \sin \beta \mid$ is effective in im-

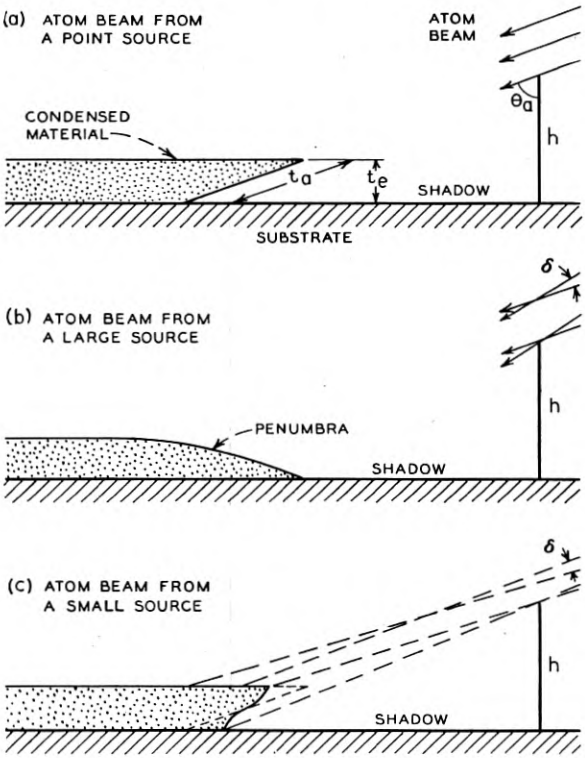



Fig. 17—Forms of shadow-edges when every atom sticks where it strikes.

proving the resolution. For example, one side of the shadow in the enlarged micrograph of Fig. 3 is very sharp.

Actual sources are of finite size, and in the cases illustrated in Fig. 17(b) and (c) the penumbra extends over a horizontal distance $p = h \sec^2 \theta_a \tan \delta$, where $\delta$ is the angle subtended by the source. If the source were of proper geometry and uniform intensity, one could set $p = d$ to obtain a vertical shadow-edge, whose intrinsic resolution would be zero. Actual sources are neither of proper geometry nor of uniform intensity, resulting in shadow-edges such as that of Fig. 17(c). If one takes into account a profile angle

$\beta$, and if $h$ is so large that $p > d$, then somewhere along the side of the profile there should be a very sharp length of shadow-edge. In Fig. 3 the sharpest length of shadow-edge is neither at top nor at bottom of the side, but in an intermediate position.

In replicas other than those of particles which are dispersed on flat substrate surfaces, the topography of the surface in the shadow region is usually not planar. In general the form of the shadow-edge is a complicated function of $\underline{n}$, of the finite source, and of the profile casting the shadow. Its intrinsic resolution may vary from large values down to nearly zero. This last fact is important since it implies that *the intrinsic resolution of the sharpest portions of shadow-edges may often be assumed to be so small that the observed resolution is that of the microscope itself.* These remarks also apply to shadow profiles, for which the special conditions required for extreme sharpness probably occur more frequently than for shadow-edges.

d) Total Resolution

The intrinsic resolution of the film is only part of the total resolution, which includes also the resolution of the electron microscope. This is determined theoretically by the wave-length of the electrons (about 0.05Å) and the numerical aperture of the microscope (about 0.003) as modified by lens aberrations. The theoretical value of 10 Å is seldom attained in practice because of lens imperfections associated with use of the microscope. Practically, microscope resolution usually lies in the range 30–50 Å, for the RCA type EMU microscope with which the micrographs were taken. In the $\times$30,000 enlargements shown, 33 Å becomes $10^{-2}$ cm, just the limit of resolution of the eye. In general the resolution of the micrographs is not this good, in part because the intrinsic resolution is larger but also because it is necessary to find suitable fine-scale features to test the resolution.

# A Gun for Starting Electrons Straight in a Magnetic Field

By J. R. PIERCE

In a simple electron gun consisting of a cathode and two apertured planes held at different potentials the apertures act as electron lenses. When the gun is immersed in a uniform axial magnetic field the aperture spacings and potentials can be chosen so that the emerging electrons have no radial velocities.

IN 1931 Davisson and Calbick showed[1] that a circular aperture in a conducting plate which separates regions with different electric gradients normal to the surfaces acts as an electron lens of focal length $F$ given by

$$F = \frac{4V}{V_2' - V_1'} \tag{1}$$

Here $V$ is the potential of the plate with respect to the cathode which supplies the electrons and $V_2'$ and $V_1'$ are the electric gradients on the far and near sides of the aperture respectively.

When an electron beam is produced by means of a plane cathode and an opposed plane positive apertured anode, the fields about the anode aperture form a diverging lens and cause the emerging beam to spread. Sometimes this is very undesirable. A strong uniform magnetic field parallel to the direction of electron flow may be used to reduce such spreading of the beam, as well as the spreading caused by space charge and by thermal velocities.

The magnetic field does not completely overcome the widening of the beam caused by the lens action of the anode aperture, for the radial velocities which the electrons have on emerging from the aperture cause them to spiral in the magnetic field, and the beam produced is alternately narrow and broad along its length.

This paper describes an electron gun consisting of a cathode and two apertured plates together with a uniform axial magnetic field. The gun is designed so that the net lens action is zero and the electrons emerge traveling parallel to the magnetic field.

The electrode system is shown in Fig. 1. The electrons travel from the plane cathode to the aperture in plane electrode $A_1$ in parallel lines. At $A_1$ they receive a radial velocity approximately $v_{r1}$, given by

$$v_{r1} = -\frac{r}{F_1} v_1 \tag{2}$$

[1] C. J. Davisson and C. J. Calbick, "Electron Lenses," *Phys. Rev.*, vol. 38, p. 585, Aug. 1931; vol. 42, p. 580, Nov. 1932.

Here $r$ is the radial position of the electron, $F_1$ is the focal length of the lens at $A_1$, and $v_1$ is the longitudinal velocity at $A_1$.

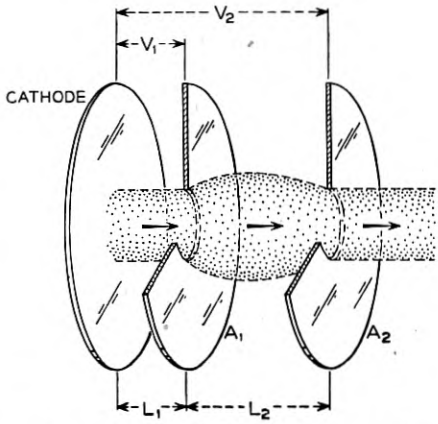



Fig. 1—The gun consists of a planar cathode and two apertured plane electrodes $A_1$ and $A_2$, with the spacings and the voltages with respect to cathode which are shown above.

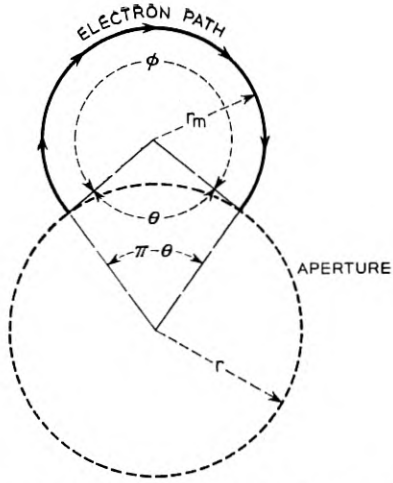



Fig. 2—Between electrodes $A_1$ and $A_2$ of Fig. 1, an electron path as seen looking parallel to the axis is a sector of a circle of angular extent $\Phi$.

The magnetic field strength is so adjusted as to return the electrons to the radius $r$ at $A_2$. Figure 2 shows the motion of an outer electron between $A_1$ and $A_2$, seen looking along the axis. Since there is no radial electric field

between $A_1$ and $A_2$, the electron will move in a circular arc of some radius $r_m$, and at $A_2$ the radial velocity will be equal and opposite to that at $r_1$; that is, it will be $-v_{r1}$.

The change in radial velocity of the electron in passing through the aperture in $A_2$, $v_{r2}$, is

$$v_{r2} = -\frac{r}{F_2} v_2 \tag{3}$$

where $F_2$ is the focal length of the lens at $A_2$ and $v_2$ is the longitudinal electron velocity at $A_2$. $F_2$ is made such that

$$v_{r2} = v_{r1} \tag{4}$$

Hence, the radial velocity $-v_{r1}$ of the approaching electrons is overcome in passing through the aperture in $A_2$ and the electrons move parallel to the axis to the right of $A_2$.

For temperature-limited emission and small space charge, we may assume a uniform gradient between the cathode and $A_1$, and between $A_1$ and $A_2$. Further, we may use the relation

$$v_2/v_1 = \sqrt{V_2/V_1} \tag{5}$$

From (1)–(5) we easily find that the required relation between $L_1$, the spacing from cathode to $A_1$, $L_2$, the spacing between $A_1$ and $A_2$, and $V_1$ and $V_2$, the potentials of $A_1$ and $A_2$ with respect to the cathode, is

$$L_2/L_1 = (\sqrt{V_1/V_2} + 1)(V_2/V_1 - 1) \tag{6}$$

In case of space-charge-limited emission, the space charge will cause the gradient to the left of $A_1$ to be $\frac{4}{3}$ times as great as in the absence of space charge. If space charge is taken into account in this region only, $L_2/L_1$ as obtained from (6) should be multiplied by $\frac{3}{4}$.

We have still to determine the magnetic field required to return the electrons leaving $A_1$ at a radius $r$ to the radius $r$ at $A_2$.

From Fig. 2 we see that the electrons turn through an angle $\Phi$. Since the angular velocity of electrons in a magnetic field is $(e/m)B$,

$$\Phi = (e/m)B\,\tau \tag{7}$$

where $\tau$ is the transit time between $A_1$ and $A_2$.

As the electron moves between $A_1$ and $A_2$ with a constant acceleration

$$\tau = \frac{2L_2}{v_1 + v_2}$$

$$\tau = \frac{2L_2}{\sqrt{2(e/m)V_2}\,(1 + \sqrt{V_1/V_2})} \tag{8}$$

Now, from Fig. 2 we see also that

$$r_m \sin(\theta/2) = r \sin(\pi/2 - \theta/2) = r \cos(\theta/2)$$
$$\tan(\theta/2) = r/r_m$$

Now

$$\theta = 2\pi - \Phi$$

so

$$\tan(\pi - (\Phi/2)) = r/r_m$$
$$\tan(\Phi/2) = -r/r_m \tag{9}$$

For circular motion with an angular velocity $(e/m)B$ and a circumferential speed $v = v_{r1} = v_{r2}$, the radius of motion $r_m$ is

$$r_m = v_{r2}/(e/m)B \tag{10}$$

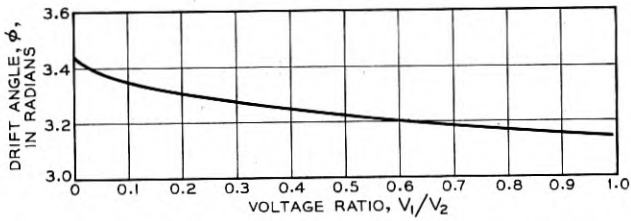



Fig. 3—The ratio $L_2/L_1$ of the electrode spacings shown in Fig. 1 should satisfy equation (6). When this is so, the angle $\Phi$, measured in radians, is a function of the voltage ratio $V_1/V_2$, and this function is shown above.

From (1), (3), (9) and (10) we obtain

$$\tan(\Phi/2) = -\left[\frac{(e/m)BL_2}{\sqrt{2\frac{e}{m}V_2(1 + \sqrt{V_1/V_2})}}\right]\frac{4}{(1 - \sqrt{V_1/V_2})}$$

From (7) and (8) we see that this may be written

$$\tan(\Phi/2) = -(\Phi/2)\frac{4}{(1 - \sqrt{V_1/V_2})}$$
$$V_1/V_2 = (1 + 4(\Phi/2)/\tan(\Phi/2))^2 \tag{11}$$

We note that $\Phi$ must lie in the third or fourth quadrant. In Fig. 3, $\Phi$ is plotted vs. $V_1/V_2$.

We now have both $L_2/L_1$ and $\Phi$ expressed in terms $V_1/V_2$, by (6) and (11). From $\Phi$ and $L_2$ we can obtain the proper value of $B$ from (7) and (8)

$$B = \Phi/(e/m)\tau$$

$$B = (\Phi\sqrt{V_2}/L_2\sqrt{2e/m})(1 + \sqrt{V_1/V_2}) \tag{12}$$

We see from Fig. 3 that there will be little error in assuming that $\Phi = \pi$.

If we assume complete space charge between the cathode and $A_1$ and neglect space charge between $A_1$ and $A_2$, nothing is altered save the ratio $L_2/L_1$; as was explained previously, this becomes $\frac{3}{4}$ times the value given by (6).

In the case of slits $L_2/L_1$ is the same function of $V_2/V_1$ as for apertures; the correction for space charge is the same, and (12) will give the correct magnetic field with $\Phi = \pi$.

# Electron Streams in a Diode*

By FRANK GRAY

A general solution of the electron stream equations is developed for a parallel plane diode, under the assumption that the electron velocity is single valued. This solution contains all particular solutions. It serves to unify the wave theory and the particle theory of electron flow, and it is an approximation for multi-velocity streams over a wide range of conditions.

## Introduction

THE theory of an electron stream flowing in a diode has received much attention;[1–13] because the tetrodes, pentodes and other modern tubes are cascade arrangements of individual diodes. The theory of the diode is the foundation for considering the circuit characteristics and the noise characteristics of these tubes. In earlier days when communication channels were confined to relatively low frequencies, an electron could traverse a diode in a short period of time compared to an oscillation of any electrical signal, and the theory could be developed rather simply from the known d-c equations. But in these days of high and ultra-high frequencies, the situation is quite different. A signal voltage may oscillate several times while an electron is traversing a diode, and the electron stream flows in bunches or waves. The present article is primarily concerned with this more complicated type of flow. It is confined to the case of parallel plane electrodes, and developed under the usual assumption that the electron velocity is a single valued function of space and time. It is shown to be an approximate solution for physical electron streams over a wide range of conditions.

Particular solutions for an electron stream under small signal conditions are given in various published articles. These theories approach the subject in two different manners. In one approach attention is confined to the motions of electrons as individual particles,[1–3] and the other approach may best be described as a wave theory of electron streams. But the two lines of approach have not hitherto given identical results, and the disagreement can probably be attributed to neglected factors in the wave theory.

The present article considers electron streams without regard to any other than a mathematical approach to the subject. The differential equations are linear in the derivatives, and they should therefore have a general solution that contains all particular solutions. The theory seeks and obtains

* The paper was presented at a meeting of the American Physical Society in Columbus, Ohio in 1945.

this general solution.† It involves a wave equation, and the results are in exact agreement with the small-signal calculations for the motions of electrons as individual particles. It is therefore believed that the general solution reconciles the two lines of approach to the theory of electron streams.

With this solution available, the situation is comparable to that encountered in two-dimensional potential theory; assignment of definite functions to two arbitrary functions gives a solution for a particular problem in electron streams, but it is then difficult to determine just what problem has been solved. In the case of small signals the general solution does not greatly shorten the calculations, and it probably should not be regarded as a labor saving tool in comparison to any particular solution when the latter is already known. It is more probable that the broader solution will serve as a guide for general reasonings about electron streams, and as a guide to approximations that can be used in particular problems.

## 1. The Parallel Plane Diode

The diode of this article is shown in Fig. 1. It is two parallel planes indicated as (a) and (b), and separated by a distance $l$. The first plane (a) may be a thermionic cathode that emits electrons, or it may be a grid through which a stream of electrons is injected into the diode. The second plane (b) may be a metallic plate that receives the electrons after they have traversed the diode space, or it may be a grid that permits the electron stream to pass out of the diode. The dimensions of the diode are assumed small compared to the electromagnetic wave-length at any frequency involved, that is, small compared to the velocity of light divided by the frequency; and the separation of the planes is assumed small compared to their lateral dimensions. Under these conditions the electric intensity is parallel to the $x$-axis, and the electrons move in that direction only.

The electron stream injected through the first plane may vary with time, both in charge density and electron velocity; and the voltages at the two planes may also vary with time. The total current flowing in the diode space is then the sum of two components: a conduction current resulting from the motion of electrons, and a displacement current arising from the time rate of change of electric intensity. The displacement current can flow even when there are no electrons in the diode space; it is then the familiar a-c. current, flowing between two plates of a condenser. But, when electrons are present, the two currents interact with each other and they both flow in a complicated manner.

† H. W. König also demonstrated the existence of a general solution; and he developed the solution for the particular case of a sinusoidal current.[5]

The determining conditions that can be measured in any physical circuit associated with the diode are: the total current, the conduction current at the first plane, and the electron velocity at that plane. Then, for conveniently considering the diode as a circuit element, it has been shown by others[8] that we should be able to calculate the conduction current at the second plane, the electron velocity at that plane, and the resultant voltage across the diode. From the viewpoint of circuit theory, these last three quantities may be considered as dependent variables whose solutions should be sought in terms of the initial conditions. But an electron stream flows according to its own nature, with little regard for circuit theory, its fundamental equa-

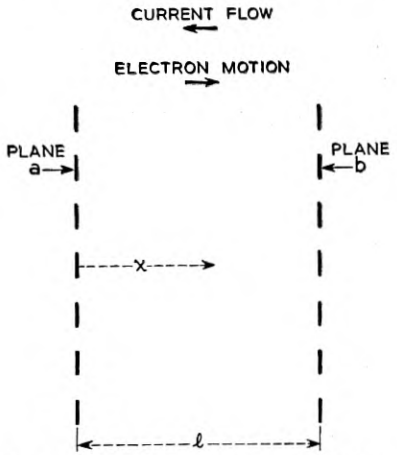



Fig. 1—Parallel plane diode with a first plane (a) and a second plane (b).

tions involve electric intensity and electron velocity as the dependent variables, and the general theory must therefore be developed in terms of these naturally occurring quantities. But it should be noted that the desired circuit relations can always be calculated from these fundamental variables.

### 1.1 Units and Symbols

The equations are written in practical electrical units, centimeters, grams and seconds. In this system of units, the permittivity $\epsilon$ of a vacuum is $\frac{10^{-11}}{36\pi}$, and the acceleration constant $\eta$ of an electron is approximately $1.77 \cdot 10^{15}$. To conform with circuit convention, the total current and the conduction current are measured in the negative $x$-direction, that is, opposite to the motion of the electrons; all other directed quantities are measured in the positive $x$-direction.

The symbols are introduced and defined as needed. The following partial list is included to give the reader a general idea of the symbolism:

*General Symbols*

| | |
|---|---|
| $a,b$ | Subscripts referring to the diode planes |
| $x$ | Distance from the first plane |
| $l$ | Length of diode space |
| $t$ | Time |
| $\tau$ | D-c. transit time to $x$ |
| $T$ | D-c. transit-time across the diode |
| $\omega$ | $2\pi$ Frequency |
| $j$ | $\sqrt{-1}$ |
| $\beta$ | $j\omega T$ |
| $\rho$ | Space-charge density |
| $\zeta$ | D-c. space-charge factor |
| $\epsilon$ | Permittivity of vacuum, $\frac{10^{-11}}{36\pi}$ |
| $\eta$ | Acceleration constant of an electron $1.77 \cdot 10^{15}$. |

*Symbols in Section 2—The General Solution*

| | |
|---|---|
| $V$ | Voltage |
| $E$ | Electric intensity |
| $U$ | Idealized electron velocity |
| $Q$ | Conduction current density |
| $I$ | Total current density |
| $I_D$ | D-c. component of $I$ |
| $I_A$ | Oscillating component of $I$ |
| $S$ | $\epsilon E + \int I\,dt$ |
| $S$ | $\epsilon E + \int I_A dt$ |
| $F_1(S), F_2(S)$ | Arbitrary functions of S |
| $A(S), B(S)$ | Finite arbitrary functions of $S$. |

*Symbols in Section 3—Small Signal Theory*

In this section the capital letters $V$, $E$, $U$, $Q$ and $I$ change their meaning and indicate only the d-c. components of their quantities; and the small letters $v$, $e$, $u$, $q$ and $i$ then indicate the amplitudes of the corresponding a-c. components. This section also uses the following special symbols:

| | |
|---|---|
| $A, B$ | Arbitrary constants |
| $A^* B^* \ldots I^*$ | Coefficients for the circuit theory of a diode. |

*Symbols in Section 4—Physical Electron Streams*

This section returns to the symbolism of the general theory; and the capital letters $E$, $U$, $Q$ and $I$ indicate total quantities. It also uses the following special symbols:

| | |
|---|---|
| $v$ | Actual electron velocity |
| $\overline{U}$ | Average of $v$ |
| $N$ | Mass density of electrons |
| $n$ | Partial density in a range $dv$ |
| $P$ | Momentum density of electrons |
| $K$ | Kinetic energy density of electrons. |

## 2. The General Solution for an Electron Stream

The present theory of electron streams is a solution of two partial differential equations, in which electron velocity $U$ and electric intensity $E$ appear as the dependent variables.

The equation for electron velocity $U$ is based on an idealism that is commonly used in vacuum tube theory. It assumes that the electron velocity is a single-valued function of space and time or, stated in other words, it assumes that all electrons in any plane normal to the $x$-axis have the same velocity. The variable $U$ may then be regarded as a continuous function of $x$ and $t$, which is everywhere equal to the velocities of the individual electrons. The differential equation for $U$ follows at once from the fundamental mechanics of electron motion, which states that for any individual electron

$$\frac{dU}{dt} = -\eta E \tag{1}$$

where $\eta$ is the acceleration constant of an electron, and the small relativity terms are neglected. Then, since $U$ is regarded as a continuous function of $x$, its total derivatives may be written in terms of partial derivatives, and

$$U\frac{\partial U}{\partial x} + \frac{\partial U}{\partial t} = -\eta E \tag{2}$$

which is here regarded as the fundamental equation for electron velocity. It is of course based on an idealizing assumption that imposes limitations on the general theory, and these limitations are discussed in a concluding section of the article, where it is shown that the idealized velocity is an approximation for the average velocity in physical electron streams.

The differential equation for the electric intensity $E$ is given by the theory of electromagnetism. It follows from this fundamental theory that

the total current density $I$ flowing in the diode is a function of time alone; it has the same value at all planes along the $x$-axis, and is given by

$$I = -\rho U - \epsilon \frac{\partial E}{\partial t} \tag{3}$$

The first term is the conduction current density, the second term is the displacement current density, and $I$ is measured according to circuit conventions in the direction opposite to the motion of the electrons. The charge density $\rho$ is

$$\rho = \epsilon \frac{\partial E}{\partial x} \tag{4}$$

and its substitution in (3) gives

$$U \frac{\partial E}{\partial x} + \frac{\partial E}{\partial t} = -\frac{I}{\epsilon} \tag{5}$$

which is the differential equation for electric intensity.

Before passing, it should be noted that the conduction current density $Q$, measured according to circuit conventions, is

$$Q = -\rho U = -\epsilon U \frac{\partial E}{\partial x} \tag{6}$$

The two differential equations for $U$ and $E$ are now repeated as a group

$$U \frac{\partial U}{\partial x} + \frac{\partial U}{\partial t} = -\eta E \tag{2}$$

$$U \frac{\partial E}{\partial x} + \frac{\partial E}{\partial t} = -\frac{I}{\epsilon} \tag{5}$$

and in this group the total current density $I$ may be regarded as any known or arbitrarily assigned function of time. These are the basic equations whose solution is sought in the present theory of electron streams. They are a description of the whole diode space, and they tell how $U$ and $E$ occur and vary with time throughout that whole space. They are first order equations, linear in their derivatives, and it is known from the theory of differential equations that their general solution is the complete solution, and that it will contain two arbitrary functions. So if we find a solution containing two arbitrary functions, we may be quite sure that it is the complete solution. The equations can be solved by the Lagrange method, as outlined in Appendix I. But that is a rather abstract operation, and the solution is here obtained by another method that has more physical meaning and is really equivalent to the Lagrange method.

For any individual electron, (2) and (5) may be written in the form of total differential equations (7) and (8)

$$\frac{dU}{dt} = -\eta E \tag{7}$$

$$\epsilon \frac{dE}{dt} = -I \tag{8}$$

where for that individual electron

$$U = \frac{dx}{dt} \tag{9}$$

and $x$ is the coordinate of the electron. This group of total differential equations describes $U$ and $E$ only in the immediate vicinity of the one moving electron, and it is therefore a restricted picture in comparison to the one given by the original partial differential equations. It should, however, be clearly understood that we are not seeking the solution of this group of total differential equations; we are merely using them as aids for solving the original equations.

Equation (8) may be written in the form

$$\frac{d}{dt}\left[\epsilon E + \int I\, dt\right] = 0 \tag{10}$$

the bracketed term is regarded as a new variable $S$, that is,

$$S = \epsilon E + \int I\, dt \tag{11}$$

and (10) says that $S$ is an invariant for any individual electron, it remains constant as the electron moves along. The solution of (10) is

$$S = C_1 \tag{12}$$

where $C_1$ is any arbitrary constant.

Turning now to (7) it may be written in the form

$$\frac{dU}{dt} = -\frac{\eta}{\epsilon}\left[S - \int I\, dt\right] \tag{13}$$

and its solution for any particular electron—remembering that $S$ is an invariant for that electron—is

$$U = C_2 - \frac{\eta}{\epsilon}\left[St - \iint I\, dt\right] \tag{14}$$

where $C_2$ is an arbitrary constant. [In repeated integrations with respect to time, the increment $dt$ is written only once, it being understood that $dt$ is repeated in each integration.] Now any arbitrary function $F_1$ of $S$ is a constant for the particular electron under consideration, so we may replace $C_2$ by $F_1(S)$ and write

$$U = F_1(S) - \frac{\eta}{\epsilon}\left[St - \iint I\,dt\right] \tag{15}$$

For the same electron, (9) may now be written in the form

$$\frac{dx}{dt} = F_1(S) - \frac{\eta}{\epsilon}\left[St - \iint I\,dt\right] \tag{16}$$

and its solution is

$$x = C_3 + F_1(S)t - \frac{\eta}{\epsilon}\left[\frac{St^2}{2} - \iiint I\,dt\right] \tag{17}$$

where $C_3$ is an arbitrary constant that may again be replaced by an arbitrary function $F_2$ of $S$, and

$$x = F_2(S) + F_1(S)t - \frac{\eta}{\epsilon}\left[\frac{St^2}{2} - \iiint I\,dt\right] \tag{18}$$

By considering one individual electron we have thus arrived at two general relations (15) and (18) which, taken together, describe $U$ and $E$ as functions of $x$ and $t$. Now the reader will probably be much surprised, as was the writer, to learn that these two equations when standing alone are not solutions of the group of total differential equations (7), (8) and (9). The solution of that group is (12), (15) and (18). In other words, the two general relations are solutions of the total differential equations only in the very special case of $S$ equal to a constant. But this constant may have any value, and the general relations therefore apply to all electrons in the diode space.

We are therefore practically forced to the conclusion that (15) and (18) are the solution of the broader group of partial differential equations (2) and (5), and this turns out to be true. This solution, which is here rewritten,

$$U = F_1(S) - \frac{\eta}{\epsilon}\left[St - \iint I\,dt\right] \tag{15}$$

$$x = F_2(S) + F_1(S)t - \frac{\eta}{\epsilon}\left[\frac{St^2}{2} - \iiint I\,dt\right] \tag{18}$$

$$S = \epsilon E + \int I\,dt$$

moreover contains two arbitrary functions, and it is therefore the general and complete solution for an idealized stream flowing in a diode.

As the solution stands, $I$ is an arbitrary function of time, and $F_1(S)$ and $F_2(S)$ are perfectly arbitrary. They correspond to all possible determining conditions: to all the d-c., a-c. and transient conditions that are possible in the idealized diode, and to all the purely mathematical conditions that cannot be realized in any physical sense.

With this complete solution available, the situation is analogous in many respects to that encountered in the solution of potential problems in two-dimensional space. We can find a particular solution by merely assigning definite functions to the three arbitrary functions $I(t)$, $F_1(S)$ and $F_2(S)$; but we then encounter the difficult task of finding out just what problem has been solved.

As a simple example of the general method, the reader may be interested in arbitrarily setting $I$, $F_1(S)$ and $F_2(S)$ equal to zero. He will then find that the resulting expressions, (15) and (18), are actual solutions of the partial differential equations, and that they represent a transient electron stream that can flow for a short period of time in a diode space.

### 2.1 The General Solution in the Presence of a Direct Current

In the majority of circuits that are of practical interest, there is a continuous direct current flowing in a diode, and the arbitrary functions then assume a more restricted form. In such cases the total current density $I$ may be considered as the sum of a d-c. component, which for the time being is indicated as $I_D$, and a transient or alternating component $I_A$.

Then we have the condition that

$$I_D > 0 \tag{19}$$

and also the condition that $U$ and $x$ must be finite in any physical tube. Now consider (15) for $U$ and note that

$$\begin{aligned} \int I\,dt &= I_D t + \int I_A\,dt \\ \iint I\,dt &= \frac{I_D t^2}{2} + \iint I_A\,dt \end{aligned} \tag{20}$$

The bracketed factor in (15) thus contains power terms in $t$, which becomes infinite as $t$ approaches infinity. The function $F_1(S)$ must therefore be of such form that it cancels these terms and causes $U$ to remain finite. Inspection shows that $F_1(S)$ must consequently be of the form

$$F_1(S) = A(S) + g + g_1 S + g_2 S^2 \tag{21}$$

where $A(S)$ is an arbitrary finite function of $S$, $g$ is an arbitrary constant, and the coefficients $g_1$ and $g_2$ have such values that the power terms in $t$ cancel out in (15). The finite function may, for example, be a sinusoidal function of $S$, or a series of such sinusoidal terms. The values of the coefficients are easily calculated, and when the resultant expression for $F_1(S)$ is substituted in (15), it may be written:

$$U = g + A(S) + \frac{\eta}{\epsilon}\left[\frac{\bar{S}^2}{2I_D} + \iint I_A\, dt\right] \tag{22}$$

where $\bar{S}$ is $S$ with the $I_D$ term omitted, that is,

$$\bar{S} = \epsilon E + \int I_A\, dt \tag{23}$$

In a similar manner it may be shown that, for $x$ to be finite, (18) assumes the form

$$x = k + B(S) - \frac{g + A(S)}{I_D}\bar{S} - \frac{\eta}{\epsilon}\left[\frac{\bar{S}^3}{6I_D^2} - \iiint I_A\, dt\right] \tag{24}$$

where $k$ is an arbitrary constant, and $B(S)$ is any arbitrary finite function of $S$. Then (22) and (24) constitute the general solution when a continuous direct current is flowing in the diode. They are mathematical means for shortening the calculations in the presence of the direct current.

It is believed that the general solution presented in this section will serve as a guide for reasoning about electron streams, and as a guide that can be used in particular problems. It should also be an aid for considering the large signal theory of electron streams. But it is here advisable to confine attention to a less ambitious program, and apply the method to the case of small signals. The results will not be entirely new, but they will bring out certain important features of the general solution.

### 3. The Small Signal Theory of Electron Streams

The small signal theory is developed as follows: The value of each dependent variable, in any plane normal to the $x$-axis, is regarded as the sum of two components: a value that does not vary with time and is therefore called the d-c. component, and a value that does vary with time and is called the a-c. component. All of these components may vary with $x$, that is, with the exception of the total current density which is a function of time alone. Corresponding to small signal circuit theory, it is also assumed that the a-c. quantities are small compared to the d-c. quantities, and that the squares and products of the a-c. quantities are negligible in comparison to their first order values. For such signals the circuit equivalent of a diode

is completely determined by its performance at single frequencies, and this permits the solution to be developed in terms of simple sinusoidal functions of time.

New symbols are needed for the small signal theory, and to avoid an undue number of subscripts they are introduced in the following manner: In the preceding general equations the dependent variables were indicated by capital letters; in the following small signal theory, the same capital letters are used to indicate the d-c. components, and the corresponding small letters then indicate the complex amplitudes of the a-c. components. This gives the following list of symbols:

| | DC Component | Amplitude of AC Component |
|---|---|---|
| Total current density | $I$ | $i$ |
| Conduction current density | $Q$ | $q$ |
| Voltage | $V$ | $v$ |
| Electric intensity | $E$ | $e$ |
| Electron velocity | $U$ | $u$ |

This symbolism has the disadvantage of using $e$ to indicate both electric intensity and the base of the natural logarithms, but the duplication causes no serious confusion, for the meaning of the symbol is always evident from the text. As examples of the new nomenclature, the conduction current density is now $Q + qe^{j\omega t}$, and the electron velocity is $U + ue^{j\omega t}$. It should be noted that the a-c. amplitude in each of these expressions is a complex space-varying amplitude, which is sometimes called the space part of the a-c. component.

Before passing it is well to write the following useful relations, which follow immediately from the fundamental equations (5) and (6):

$$\begin{aligned} q &= i + j\omega\epsilon e \\ e &= \frac{j(i - q)}{\epsilon\omega} \end{aligned} \tag{25}$$

their substitution in (11) and (23) give

$$\begin{aligned} S &= \epsilon E + It - \frac{jq}{\omega} e^{j\omega t} \\ \bar{S} &= \epsilon E - \frac{jq}{\omega} e^{j\omega t} \end{aligned} \tag{26}$$

With this introduction to the change in symbolism, we now express the general solution (22) and (24) in terms of the new symbols, and neglect all

second order terms in the oscillating components. Each part of the general solution then separates into two equations, one for the d-c. quantities, and another for the a-c. quantities. The resulting equations for the d-c. components are

$$U = g + \frac{\eta\epsilon E^2}{2I} \tag{27}$$

$$x = k - \frac{g\epsilon E}{I} - \frac{\eta\epsilon^2 E^3}{6I^2} \tag{28}$$

and the equations for the a-c. components are

$$ue^{j\omega t} = A(S) - \frac{\eta}{\epsilon}\left[\frac{j\epsilon E}{\omega I}\, q + \frac{i}{\omega^2}\right] e^{j\omega t} \tag{29}$$

$$0 = B(S) - \frac{\epsilon E}{I}\, A(S) + j\left[\frac{U}{\omega I}\, q + \frac{\eta}{\epsilon\omega^3}\, i\right] e^{j\omega t} \tag{30}$$

where in the last equation $g$ has been replaced by its value from (27).

### 3.1 The D-c. Components of the Electron Stream

We first consider the d-c. components in (27) and (28). It is easily shown that they obey the primitive differential equations

$$\begin{aligned} U\frac{\partial U}{\partial x} &= -\eta E \\ U\frac{\partial E}{\partial x} &= -I/\epsilon \end{aligned} \tag{31}$$

which are the static equations for a diode, when it is idling in the absence of an a-c. signal. Their solutions are given in various published articles, and they are available without further calculations.[7, 8] These d-c. components are involved in the subsequent development of the a-c. theory, and the latter requires certain d-c. relations. These relations are briefly presented without derivations as follows:

The current density $I$ and the d-c. voltages at the two diode planes are assumed to be known quantities. Then the d-c. velocities at those planes are also known quantities, because their values are given by the simple relations

$$U_a = \sqrt{2\eta V_a}\,, \qquad U_b = \sqrt{2\eta V_b} \tag{32}$$

where it is assumed that the original source of electrons is at zero voltage.

The d-c. transit time plays an important role in the small signal theory.

The transit time $\tau$ from the first plane to any coordinate $x$ is

$$\tau = \int_0^x \frac{dx}{U} \tag{33}$$

and

$$\frac{\partial \tau}{\partial x} = \frac{1}{U} \tag{34}$$

The total transit time $T$ across the diode space also plays an important role; it is usually expressed in terms of a so-called space charge factor $\zeta$, whose value is given by[8]

$$\zeta \left(1 - \frac{\zeta}{3}\right)^2 = \frac{4}{9} \frac{I}{I_m} \tag{35}$$

Here $I$ is the actual d-c. current; and $I_m$ is the maximum current that could be projected across the diode when its planes are at the voltages $V_a$ and $V_b$, that is, $I_m$ is the space charge limited current

$$I_m = \frac{2\epsilon}{9} \sqrt{\frac{2}{\eta}} \frac{(\sqrt{V_a} + \sqrt{V_b})^3}{l^2} \tag{36}$$

Then the total transit time $T$ is given by

$$T = \frac{2l}{\left(1 - \frac{\zeta}{3}\right)\left(U_a + U_b\right)} \tag{37}$$

It also follows that I can be expressed in the form

$$I = \frac{2\epsilon\zeta}{\eta T^2} (U_a + U_b) \tag{38}$$

Certain equations for the d-c. electric intensity are also required. They are

$$\begin{aligned} E &= E_a - \frac{I\tau}{\epsilon} \\ E_a &= \frac{1}{\eta T} (U_a - U_b) + \frac{IT}{2\epsilon} \\ E_b &= \frac{1}{\eta T} (U_a - U_b) - \frac{IT}{2\epsilon} \end{aligned} \tag{39}$$

They complete the list of d-c. relations required in the following small signal theory.

### 3.2 The A-c. Components of the Electron Stream

We now return to the a-c. equations for the electron stream, (29) and (30). In (29), the arbitrary function $A(S)$ must obviously involve an exponential function of $j\omega t$, and it must therefore be of the form

$$A(S) = Ae^{\frac{j\omega S}{I}} \tag{40}$$

where $A$ is an arbitrary constant. Then the substitution of (26) gives

$$A(S) = A \exp.\left[j\omega\left(t + \frac{\epsilon E}{I}\right) + \frac{q}{I}e^{j\omega t}\right] \tag{41}$$

The term in $q/I$ is a second-order term that may be neglected, and $\frac{\epsilon E}{I}$ can be replaced by its value from (39) that is,

$$\frac{\epsilon E}{I} = \frac{\epsilon Ea}{I} - \tau \tag{42}$$

where $\tau$ is the d-c. transit time to any coordinate $x$. The resultant exponential factor in $\frac{\epsilon Ea}{I}$ can then be included in the arbitrary constant $A$, and this gives

$$A(S) = Ae^{j\omega(t-\tau)} \tag{43}$$

The substitution of this function in (29) now gives the following relation for the amplitudes of the a-c. components

$$u = Ae^{-j\omega\tau} - \frac{j\eta E}{\omega I}q - \frac{\eta}{\epsilon\omega^2}i \tag{44}$$

The arbitrary function $B(S)$ may be treated in a similar manner, and (30) then gives the complex amplitude of the conduction current density

$$q = \frac{j\omega I}{U}\left(B - A\frac{\epsilon E}{I}\right)e^{-j\omega\tau} - \frac{\eta I}{\epsilon\omega^2 U}i \tag{45}$$

where B is another arbitrary constant.

The substitution of this value of $q$ in (25) and (44) also gives the amplitudes of electron velocity and electric intensity.

$$u = \left[B\frac{\eta E}{U} + A\left(1 - \frac{\eta\epsilon E^2}{IU}\right)\right]e^{-j\omega\tau} - \frac{\eta}{\epsilon\omega^2}\left[1 - \frac{j\eta E}{\omega U}\right]i \tag{46}$$

$$e = \frac{I}{\epsilon U}\left[B - A\frac{\epsilon E}{I}\right]e^{-j\omega\tau} + \frac{j}{\epsilon\omega}\left[1 + \frac{\eta I}{\epsilon\omega^2 U}\right]i \tag{47}$$

The amplitude of the a-c. voltage in the diode space is also required, and it is derived from its expression

$$v = v_a - \int_0^x e\, dx \tag{48}$$

where $e$ has its value (47), and the integration can be performed by remembering that $\frac{\partial \tau}{\partial x}$ is $1/U$. This gives

$$v = v_a + \frac{1}{\epsilon\omega}\left[\left(j\epsilon AE - A\frac{I}{\omega} - jBI\right)(e^{-j\omega\tau} - 1) - jAI\tau - j\left(x + \frac{\eta I\tau}{\epsilon\omega^2}\right)i\right] \tag{49}$$

We are now in a position to examine the character of the electron stream, and for this purpose we write the conduction current density in its complete form $qe^{j\omega t}$, that is,

$$qe^{j\omega t} = \frac{j\omega I}{U}\left(B - A\frac{\epsilon E}{I}\right)e^{j\omega(t-\tau)} - \frac{\eta I}{\epsilon\omega^2 U}\, ie^{j\omega t} \tag{50}$$

The phase angle of the first term involves the d-c.transit time $\tau$, which is a function of $x$, so this term is a wave traveling in the $x$-direction. Its amplitude involves the d-c. quantities $U$ and $E$, and its amplitude varies with $x$. The velocity of the wave is given by

$$\text{Wave velocity} = \left(\frac{\partial \tau}{\partial x}\right)^{-1} \tag{51}$$

and from (34)

$$\text{Wave Velocity} = U \tag{52}$$

That is the velocity of the conduction current wave is equal to the d-c. component of electron velocity.

The second term in (50) is an oscillation that has the same phase throughout the diode space, and its amplitude also varies with $x$. The a-c. conduction current is thus a wave of electric charge traveling at a finite velocity plus an oscillating charge that is in phase over the entire diode space.

An inspection of equations (46), (47) and (49) shows that the other a-c. components are of the same general form; each of them is a wave traveling in the $x$-direction plus an oscillation that is in phase over the entire diode. This clear-cut disclosure of the dual nature of an electron stream is an important contribution of the general theory.

The formal solution for small signals is really completed with the derivation of the preceding general equations for the a-c. amplitudes. But there still remains the rather tedious process of deriving the relations for circuit calculations as outlined in the following section, and they give a direct comparison with previous theories of electron streams.

### 3.3 Small Signal Equations for Circuit Calculations

Llewellyn[3] has shown that the treatment of a diode as a circuit element requires certain variables at the second plane to be expressed in terms of their values at the first plane; that is, the circuit theory requires three equations of the form

$$\begin{aligned} v_b - v_a &= A^* i + B^* q_a + C^* u_a \\ q_b &= D^* i + E^* q_a + F^* u_a \\ u_b &= G^* i + H^* q_a + I^* u_a \end{aligned} \tag{53}$$

where the starred coefficients are known functions of the d-c. components.

The derivation of these relations from the preceding general equations is outlined as follows: the first step is the evaluation of the arbitrary constants $A$ and $B$. This is done by substituting the values at the first plane in (44) and (45), and then solving for $A$ and $B$, which gives

$$A = u_a + \frac{j\eta E_a}{\omega I} q_a + \frac{\eta}{\epsilon\omega^2} i \tag{54}$$

$$B = \frac{\epsilon E_a}{I} A - \frac{jU_a}{\omega I} q_a - \frac{j\eta}{\epsilon\omega^3} i \tag{55}$$

These expressions, and the values at the second plane, are then substituted in the equations for the a-c. amplitudes (45), (46) and (49); and they immediately give the desired relations. They do, however, involve the inconvenient electric intensities $E_a$ and $E_b$, and these quantities are replaced by their values from (39).

These simple but rather tedious substitutions are illustrated by the following derivation of $q_b-$, which is brief enough to be included for that purpose. The first step is the substitution of the values at the second plane in (45); this gives

$$q_b = \frac{j\omega I}{U_b}\left(B - \frac{\epsilon E_b}{I} A\right) e^{-\beta} - \frac{\eta I}{\epsilon\omega^2 U_b} i \tag{56}$$

where $\beta$ is $j\omega T$. It is now advantageous to replace $B$ by its value from (55), and

$$q_b = j\omega\epsilon\left(\frac{E_a - E_b}{U_b}\right) A e^{-\beta} + \frac{U_a}{U_b} e^{-\beta} q_a + \frac{\eta I}{\epsilon\omega^2 U_b}(e^{-\beta} - 1) i \tag{57}$$

The inconvenient electric intensities are now easily eliminated by substitutions from the d-c. equations (39), which give

$$q_b = \frac{I\beta e^{-\beta}}{U_b} A + \frac{U_a}{U_b} e^{-\beta} q_a + \frac{\eta I}{\epsilon\omega^2 U_b} (e^{-\beta} - 1)i \tag{58}$$

The value of $A$ is now introduced from (54), with $E_a$ again eliminated by (39), and a grouping of terms then gives the final equation

$$q_b = \frac{\eta I}{\epsilon\omega^2 U_b} (\beta e^{-\beta} + e^{-\beta} - 1)i + \left(1 - \frac{\eta I T^2}{2\epsilon U_b}\right) e^{-\beta} q_a + \frac{I\beta e^{-\beta}}{U_b} u_a \tag{59}$$

This equation gives the following values of starred coefficients:

$$D^* = \frac{\eta I}{\epsilon\omega^2 U_b} (\beta e^{-\beta} + e^{-\beta} - 1) \tag{60}$$

$$E^* = \left(1 - \frac{\eta I T^2}{2\epsilon U_b}\right) e^{-\beta}$$

$$F^* = \frac{I\beta}{U_b} e^{-\beta}$$

These coefficients may now be rewritten in any desired form; and, to conform with previous articles on electron streams, we replace $\omega$ by its equivalent expression $\frac{-j\beta}{T}$; and we also express $I$ in terms of the space charge factor $\zeta$ from (38), that is,

$$I = \frac{2\epsilon\zeta}{\eta T^2} (U_a + U_b) \tag{61}$$

These substitutions then give the coefficients in the form

$$D^* = 2\zeta \left(\frac{U_a + U_b}{U_b}\right) \frac{1 - e^{-\beta} - \beta e^{-\beta}}{\beta^2}$$

$$E^* = \frac{1}{U_b} [U_b - \zeta(U_a + U_b)] e^{-\beta} \tag{62}$$

$$F^* = \frac{2\epsilon\zeta}{\eta T^2} \left(\frac{U_a + U_b}{U_b}\right) \beta e^{-\beta}$$

This is obviously a longer mode of expression, but it has two advantages: it is convenient for circuit calculations, and it permits a direct comparison with previous articles on electron streams.

The equations for $u_b$ and $(v_b - v_a)$ are obtained by similar substitutions in (46) and (49); and the three equations are then written in the symbolic

form (53), with the values of the starred coefficients abbreviated and assembled as follows:

$$\begin{aligned} \alpha_1 &= 1 - e^{-\beta} - \beta e^{-\beta} \\ \alpha_2 &= 1 - e^{-\beta} \\ \alpha_3 &= 2 - 2e^{-\beta} - \beta - \beta e^{-\beta} \end{aligned} \tag{63}$$

$$A^* = \frac{T^2}{2\epsilon\beta}(U_a + U_b)\left[1 - \frac{\zeta}{3}\left(1 - \frac{12\alpha_3}{\beta^3}\right)\right]$$

$$B^* = \frac{T^2}{\epsilon\beta^3}[(\alpha_1 - \beta\alpha_2)U_a - \alpha_1 U_b + \alpha_1\zeta(U_a + U_b)]$$

$$C^* = -\frac{2\alpha_1\zeta}{\eta\beta^2}(U_a + U_b)$$

$$D^* = \frac{2\alpha_1\zeta}{\beta^2}\left(\frac{U_a + U_b}{U_b}\right)$$

$$E^* = \frac{1}{U_b}[U_b - \zeta(U_a + U_b)]e^{-\beta} \tag{64}$$

$$F^* = \frac{2\epsilon\zeta}{\eta T^2}\left(\frac{U_a + U_b}{U_b}\right)\beta e^{-\beta}$$

$$G^* = -\frac{\eta T^2}{\epsilon\beta^3 U_b}[(\alpha_1 - \alpha_2\beta)U_b - \alpha_1 U_a + \alpha_1\zeta(U_a + U_b)]$$

$$H^* = -\frac{\eta T^2}{2\epsilon}\left(\frac{U_a + U_b}{U_b}\right)(1 - \zeta)\frac{e^{-\beta}}{\beta}$$

$$I^* = \frac{1}{U_b}[U_a - \zeta(U_a + U_b)]e^{-\beta}$$

With the exception of a difference in symbols, these coefficients are identically the same as those obtained by Llewellyn and Peterson[3, 8] from calculations on the motions of electrons as individual particles, and this correspondence apparently reconciles the wave theory and the particle theory of electron streams. The correspondence is largely the result of a new feature in the wave theory, that is, the expression of the electron stream as the sum of two components, a wave travelling with a finite velocity and an oscillation that is in phase over the entire diode space.

Llewellyn and Peterson have derived the circuit equivalents of electronic tubes from the values of the starred coefficients, and these equivalents are well known in the electronic art.[8] The present section confirms these relations, as derived for an idealized electron stream. The validity of this idealization is considered in the following section.

## 4. Physical Electron Streams

[This section returns to the symbolism of the general theory; and the capital letters, $V$, $E$, $U$, $Q$ and $I$ now indicate total values.]

The preceding general solution for an electron stream is based on idealism, namely, the assumption that the electron velocity is a single-valued function of space and time. The stream then obeys the differential equations (2) and (5), and the theory is a general solution of these fundamental equations. But it is well known that the velocity in a physical electron stream is not single valued:[10–13] Electrons are emitted from their original source with slightly different velocities; and some electrons acquire energy from the high-frequency electric field and overtake their slower neighbors. These factors cause the velocity to be a multi-valued function of space, and the electrons have various velocities in any plane normal to the axis of the diode. The present section derives the differential equations for a multi-velocity stream, and compares them with the idealized equations (2) and (5).

For this purpose, the actual velocity of an electron is indicated as $v$. It is also convenient to develop the equations in terms of mass, so we let $N$ equal the mass density of electrons at any coordinate $x$. The fractional mass density of electrons with velocities lying in any range from $v$ to $v + dv$ may likewise be expressed as $ndv$, where $n$ is a function of $v$, $x$ and $t$; and it follows that

$$N = \int_{-\infty}^{+\infty} ndv. \tag{65}$$

The momentum density P of the electrons is then given by

$$P = \int_{-\infty}^{+\infty} nvdv \tag{66}$$

and their kinetic energy density $K$ is

$$K = \int_{-\infty}^{+\infty} \frac{nv^2}{2} dv \tag{67}$$

It also follows that the average electron velocity $\bar{U}$ is given by

$$\bar{U} = \frac{P}{\bar{N}} \tag{68}$$

This is the new mechanical variable in the theory of physical electron streams.

The differential equation for the electric intensity is now easily derived from the fundamental electromagnetic equation

$$\epsilon \frac{\delta E}{\delta t} - Q = -I \tag{69}$$

The conduction current density $Q$ is

$$Q = -\frac{eP}{m} = -\bar{U}\frac{eN}{m} = -\bar{U}\epsilon\frac{\delta E}{\delta x} \tag{70}$$

and its substitution in (69) gives

$$\bar{U}\frac{\delta E}{\delta x} + \frac{\delta E}{\delta t} = -\frac{I}{\epsilon} \tag{71}$$

which is the analogue of (5).

The mechanical equation for the physical stream is obtained from the Liouville theorem. In the diode regions with which we shall be concerned, the individual electrons are so far apart that their microscopic forces are negligible, the electrons flow freely under the action of the macroscopic forces, and they therefore obey the Liouville theorem for particle motion. This theorem states that

$$\frac{dn}{dt} = 0 \tag{72}$$

that is, $n$ remains constant as we travel along with any particular electron. This equation may also be written in terms of partial derivatives of $n$

$$\frac{\delta n}{\delta x}\frac{dx}{dt} + \frac{\delta n}{\delta v}\frac{dv}{dt} + \frac{\delta n}{\delta t} = 0 \tag{73}$$

and the substitution of the values of the total derivatives then gives

$$v\frac{\delta n}{\delta x} - \eta E\frac{\delta n}{\delta v} + \frac{\delta n}{\delta t} = 0 \tag{74}$$

The mechanical relations are obtained by integrating this equation with respect to $v$. It is first multiplied by $dv$ and then integrated as follows:

$$\int_{-\infty}^{+\infty}\frac{\delta n}{\delta x}v\,dv - \eta E\int_{-\infty}^{+\infty}\frac{\delta n}{\delta v}\delta v + \int_{-\infty}^{+\infty}\frac{\delta n}{\delta t}dv = 0 \tag{75}$$

The second integral reduces to the difference in the values of $n$ at $v = +\infty$, and $v = -\infty$. It vanishes because there are no electrons with infinite velocities. The differential operators may also be moved outside the other integrals, to give

$$\frac{\delta}{\delta x}\int_{-\infty}^{+\infty} nv\,dv + \frac{\delta}{\delta t}\int_{-\infty}^{+\infty} n\,dv = 0 \tag{76}$$

then from (65) and (66)

$$\frac{\delta N}{\delta t} = -\frac{\delta P}{\delta x} \tag{77}$$

Equation (74) is next multiplied by $v dv$, and a similar integration gives

$$\frac{\delta P}{\delta t} = -\eta N E - 2\frac{\delta K}{\delta x} \tag{78}$$

With these relations we are now in a position to derive the differential equation for $\bar{U}$.

This mechanical equation is obtained by first writing the obvious equality

$$\bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{\delta \bar{U}}{\delta t} = \bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{\delta}{\delta t}\left(\frac{P}{\bar{N}}\right) \tag{79}$$

Then partial differentiation of the last term gives

$$\bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{\delta \bar{U}}{\delta t} = \bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{1}{\bar{N}}\frac{\delta P}{\delta t} - \frac{P}{\bar{N}^2}\frac{\delta N}{\delta t} \tag{80}$$

and, when the resultant time derivatives are replaced by (77) and (78)

$$\bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{\delta \bar{U}}{\delta t} = \bar{U}\frac{\delta \bar{U}}{\delta x} - \eta E - \frac{2}{\bar{N}}\frac{\delta K}{\delta x} + \frac{P}{\bar{N}^2}\frac{\delta P}{\delta x} \tag{81}$$

the substitution of $N\bar{U}$ for $P$ then gives the final differential equation for $\bar{U}$, which may be written in the form

$$\bar{U}\frac{\delta \bar{U}}{\delta x} + \frac{\delta \bar{U}}{\delta t} = -\eta E - \frac{2}{\bar{N}}\frac{\delta}{\delta x}\left[K - \frac{N\bar{U}^2}{2}\right] \tag{82}$$

It is the analogue of equation (2).

The two sets of equations are now assembled and written in a form suitable for comparison. The equations for the idealized stream are

$$U\frac{\delta E}{\delta x} + \frac{\delta E}{\delta t} = -\frac{I}{\epsilon} \tag{5}$$

$$\frac{1}{2}\frac{\delta U^2}{\delta x} + \frac{\delta U}{\delta t} = -\eta E \tag{2}$$

and the analogous equations for the physical stream are

$$\bar{U}\frac{\delta E}{\delta t} + \frac{\delta E}{\delta t} = -\frac{I}{\epsilon} \tag{71}$$

$$\frac{1}{2}\frac{\delta \bar{U}^2}{\delta x} + \frac{2}{\bar{N}}\frac{\delta}{\delta x}\left[K - \frac{N\bar{U}^2}{2}\right] + \frac{\delta \bar{U}}{\delta t} = -\eta E \tag{82}$$

When $U$ is set equal to $\bar{U}$, the first equations in the two sets are identical; and in this respect the theory of the idealized stream corresponds to that of the physical stream. But the second equation for the physical stream then differs from its analogue by the inclusion of an additional term

$$\frac{2}{N}\frac{\delta}{\delta x}\left[K - \frac{N\bar{U}^2}{2}\right] \tag{83}$$

The bracketed quantity in this term is the difference between the actual kinetic energy density and the kinetic energy density calculated as if the electrons were all moving with their average velocity $\bar{U}$. It is often a small term that can be neglected, and the physical stream is then approximately described by the idealized equations (2) and (5).

It is, however, rather obvious that there are cases in which this approximation cannot be made. It is invalid in the region between a thermionic cathode and its voltage minimum, where the electrons are traveling in both directions along the $x$-axis, and cause $K$ and $\frac{N\bar{U}^2}{2}$ to have appreciably different values. So, when the first plane of the diode is a space-change-limited cathode, the idealized theory can apply only in the region beyond the voltage minimum. This difficulty is usually overcome by considering the virtual cathode as the first plane of the diode. In all other regions the electrons are normally traveling in one direction only, and the idealized equations are then an approximation for the physical stream over a wide range of conditions.

The nature of this approximation is seen more clearly by considering the electrons to be uniformly distributed over a velocity range $s$, where $s$ is a function of $x$ and $t$. Then the mechanical equation (82) is

$$\frac{1}{2}\frac{\delta\bar{U}^2}{\delta x} + \frac{1}{8}\frac{\delta s^2}{\delta x} + \frac{\delta\bar{U}}{\delta t} = -\eta E \tag{84}$$

Under the usual conditions encountered in electronic tubes, $\frac{s^2}{8}$ is small compared to $\frac{U^2}{2}$, and its gradient may be neglected in comparison to that of $\frac{\bar{U}^2}{2}$.

The approximation can also be considered in a more rigorous manner as follows: The velocity spread $s$ may be expressed in electron volts by the relation

$$s^2 \doteq \frac{\eta\phi^2}{2V} \tag{85}$$

where $\phi$ is the spread measured in electron volts, and $V$ is the voltage in the stream. Then (84) may be written in the form

$$U\frac{\delta U}{\delta x}+\frac{\delta U}{\delta t}=\eta\frac{\delta V}{\delta x}-\frac{\eta}{8}\frac{\phi}{V}\frac{\delta\phi}{\delta x}+\frac{\eta}{16}\left(\frac{\phi}{V}\right)^2\frac{\delta V}{\delta x} \qquad (86)$$

The last term is small compared to $\eta\frac{\delta V}{\delta x}$ and may be neglected, and it follows that the idealized theory is an approximation for physical electron streams when

$$\frac{1}{8}\frac{\phi}{V}\frac{\delta\phi}{\delta x}\ll\frac{\delta V}{\delta x} \qquad (87)$$

it being understood that the inequality holds for the gradients of the d-c. components, and also for the gradients of the a-c. components of $\phi$ and $V$. This requirement is satisfied over a wide range of conditions, and the idealized equations are applicable in a corresponding manner.*

It is thought that these considerations explain why the single-velocity theory of electron streams is so successful in explaining the characteristics of electronic tubes.[8, 9]

## Conclusion

It is believed that the preceding pages serve to unify our theories of electron streams in some such manner as follows:

(1) They develop the general solution for a single velocity stream, and this solution contains all particular solutions.

(2) The small signal theory is considered in detail as a special case of the general solution, and the a-c. stream is shown to be the sum of two components: a wave traveling with a finite velocity plus an oscillation that is in phase over the entire diode space.

(3) The wave expression gives identically the same results as previous calculations based on the motions of electrons as individual particles.

(4) The idealized stream is shown to be an approximation for multi-velocity streams over a wide range of conditions, and this correspondence explains why the single velocity theory is so successful in describing the characteristics of electronic tubes.

## Acknowledgments

The writer wishes to thank F. B. Llewellyn and L. C. Peterson for their assistance in the study of electron streams, and to thank R. K. Potter for his encouragement in writing the article at this time.

* It should be noted that this requirement is not satisfied by a velocity-modulated stream of small current density in a long, field-free drift space.

## APPENDIX I—THE LAGRANGE SOLUTION

Lagrange has shown that any partial differential equation of the first order, linear in its derivatives, is equivalent to a group of total differential equations. The Lagrange equations corresponding to (2) and (5) are

$$\frac{dx}{U} = \frac{dt}{1} = \frac{dU}{-\eta E} \tag{88}$$

$$\frac{dx}{U} = \frac{dt}{1} = \frac{\epsilon dE}{-I} \tag{89}$$

or, taken together,

$$\frac{dx}{U} = \frac{dt}{1} = \frac{dU}{-\eta E} = \frac{\epsilon dE}{-I} \tag{90}$$

Now we can find three independent solutions of this group. One solution is

$$\epsilon E + \int I dt = c_1 \tag{91}$$

The first member of this solution is indicated as $S$; then the other solutions are

$$U + \frac{\eta S t}{\epsilon} - \frac{\eta}{\epsilon} \iint I dt = c_2 \tag{92}$$

$$x - Ut - \frac{\eta S t^2}{2\epsilon} + \frac{\eta}{\epsilon}\left[ t \iint I dt - \iiint I dt \right] = c_3 \tag{93}$$

Since each of these quantities is a constant, we may set any one of them equal to an arbitrary function of another, and the resulting equation is also a solution of (90). We can, however, obtain only two independent solutions in this manner, and we naturally choose the two simplest combinations, that is,

$$U + \frac{\eta S}{\epsilon} t - \frac{\eta}{\epsilon} \iint I dt = F_1(S) \tag{94}$$

$$x - Ut - \frac{\eta S}{2\epsilon} t^2 + \frac{\eta}{\epsilon}\left[ t \iint I dt - \iiint I dt \right] = F_2(S) \tag{95}$$

where $F_1(S)$ and $F_2(S)$ are arbitrary functions of $S$. These equations contain two arbitrary functions; they are solutions of the Lagrange equations (88) and (89), and they therefore constitute the general and complete solution of the partial differential equations (2) and (5). With the exception of a slight

difference in form, this solution is identically the same as the one given in Section 2.

## References


1. "Theory of the Internal Action of Thermionic Systems at Moderately High Frequencies," W. E. Benham, *Phil. Mag.*, *5*, 641, 1928 and *11*, 457, 1931.
2. "Electronenschwingungen im Hockvacuum", J. Muller, *Hockfreguenztech u. Electroakustik*, *41*, 156, 1933; and *43*, 195, 1934.
3. "Electron Inertia Effects," F. B. Llewellyn, Cambridge University Press, 1941.
4. "Space Charge and Field Waves in an Electron Beam," S. Ramo, *Phys. Rev.*, *56*, 276, 1939.
5. "On the Behavior of Electron Currents in Longitudinal Electric Fields," H. W. König, *Hockfreguenztech u. Electroakustik*, *62*, 76, 1943.
6. "Theory of Parallel Plane Diode" A. H. Traub and Nelson Wax, *Jl. Applied Physics*, *21*, 974, 1950.
7. "On the Theory of Space Charge Between Parallel Plane Electrodes," C. E. Fay, A. L. Samuel and W. Shockley, *B.S.T.J.*, *17*, 49, 1938.
8. "Vacuum Tube Networks," F. B. Llewellyn and L. C. Peterson, *Proc. I.R.E.*, *32*, 144, 1944.
9. "Space Charge and Transit-Time Effects on Signal and Noise in Microwave Tetrodes." L. C. Peterson, *Proc. I.R.E.*, *35*, 1264, 1947.
10. "Theory of Space Charge Effects," P. S. Epstein, *Verk. d. Deut. Phys. Ges.*, *21*, 85, 1919.
11. "Thermionic Current between Parallel Plane Electrodes: Velocities of Emission Distributed According to Maxwell's Law," T. C. Fry, *Phys. Rev.*, *17*, 441, 1921; and *22*, 445, 1923.
12. "The Effect of Space Charge and Initial Velocities on the Potential Distribution and Thermionic Current between Parallel Plane Electrodes," I. Langmuir, *Phys. Rev.*, *21*, 419, 1923.
13. "On the Velocity—Dependent Characteristics of High Frequency Tubes," J. K. Knipp, *Jl. Applied Physics*, *20*, 425, 1949.

# The Davisson Cathode Ray Television Tube Using Deflection Modulation

By A. G. JENSEN

The paper describes a cathode ray television receiving tube incorporating several unique features. The tube was designed and constructed by Dr. C. J. Davisson and was used in some of the early demonstrations of television transmission over the coaxial cable.

THE present day coaxial cable broad-band transmission system was developed during the early 1930's, and was originally conceived as a means for multi-channel telephone transmission. During the same period the rapidly developing television art was producing video signals requiring wider and wider frequency bands. It was very soon realized, therefore, that this coaxial system would also lend itself admirably to the transmission of such wide band television signals. The early development culminated in the installation of a coaxial cable route from New York to Philadelphia. This system was designed to provide 240 telephone channels or a single 800 kc television channel, and both types of transmission were successfully accomplished during a series of demonstrations in 1937.[1]

The scanning equipment used for producing the television signals for these demonstrations was developed under the direction of Dr. H. E. Ives at the Bell Telephone Laboratories. It was designed to scan standard 35 mm motion picture film and consisted of a six-foot steel disk rotating at 1440 rpm and having 240 lenses mounted around the periphery. It thus produced a television signal of 240 lines and 24 frames per second, occupying a bandwidth of about 800 kilocycles.

From this same period came the well known work of Dr. Davisson in the field of electron diffraction.[2] This work had resulted also in important advances in electron optics and in the development of the sharply focussed, well defined electron beam. It was natural, therefore, that Dr. Ives should discuss with Dr. Davisson the possibility of designing a cathode ray tube capable of adequately displaying a picture from the television signals specified above. The outcome of these discussions was that Dr. Davisson, with the close and able collaboration of C. J. Calbick, undertook to design and construct the tube described in the following pages. In this connection it should also be mentioned that the first experimental samples of the tube were built by G. E. Reitter, while the later engineering for limited production was carried out by H. W. Weinhart.

The disk scanner was a linear transmitter, since the amplitude of the signal was directly proportional to the film brightness. The fundamental requirement for a receiving tube was therefore as stated by Dr. Davisson in an early memorandum:

"If screen brightness is proportional to beam current, as for most screens it is, then beam current in the receiving tube should be proportional to signal voltage; the modulations of beam current by signal voltage should be linear. Failure to meet this requirement results in falsification of tone values in reproduction, and when departures from linearity are marked [it leads] to unsatisfactory pictures."

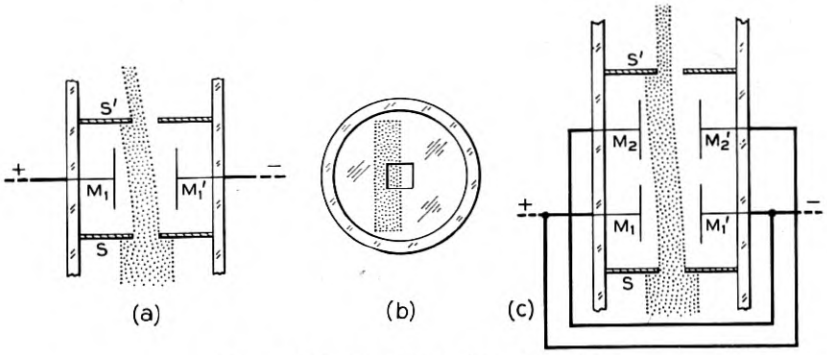



Fig. 1—Principle of deflection modulation.

It was this fundamental concept of a beam current directly proportional to input voltage, which led to the development of a tube employing deflection modulation. This is a type of modulation in which the modulating voltage causes the electron beam to be deflected across a defining aperture in such a manner that increasing modulating voltage will cause a larger area of the beam to pass through the aperture and thus increase the brightness of the screen proportionally to the modulating voltage. The principle of this type of modulation is illustrated in Fig. 1.

Figure 1(a) shows a narrow beam of electrons passing through the slit $S$ (perpendicular to the plane of the paper) and arranged to form a sharp image of the slit in the plane of the square aperture $S'$. The relation of the slit image to the square aperture is shown in Fig. 1(b). A bias voltage is applied across the modulator plates $M_1$ and $M_1'$ so that the slit image falls just off the square aperture for no signal voltage. As signal voltage is applied across the modulator plates the slit image will move across the aperture in such a manner that the cross-sectional area of the beam beyond the aperture is proportional to signal voltage (for small angles of deflection, such as used here).

Beyond the aperture $S'$ the electron beam enters a projection lens system designed to project an enlarged electron image of $S'$ onto the fluorescent screen of the tube. In order to avoid further modulation of the beam

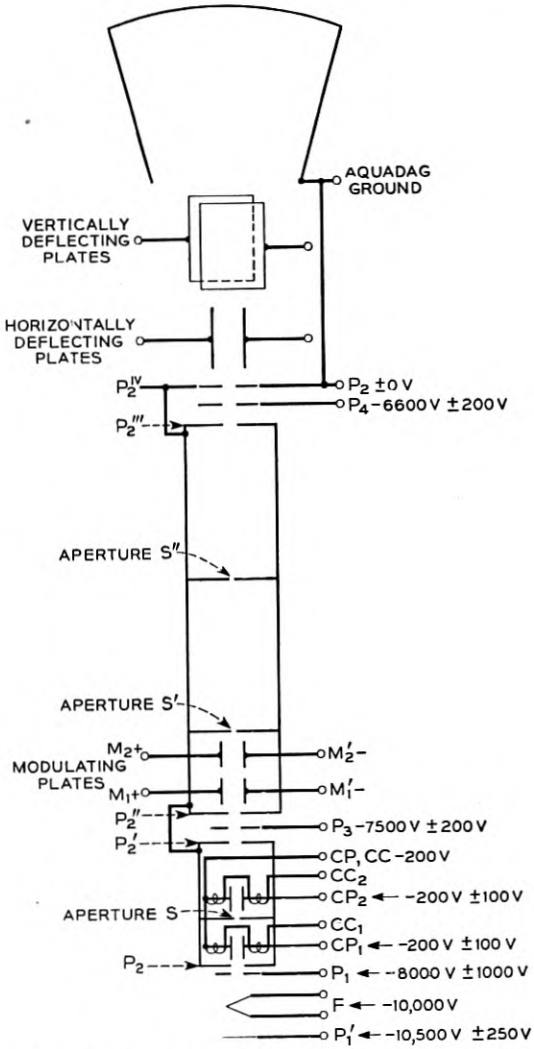



Fig. 2—Schematic diagram of electron optical system.

in the projection lens system it is essential that the beam be maintained parallel with the tube axis and not be deflected by modulation as in Fig. 1(b). To accomplish this a second pair of modulating plates $M_2$ and $M_2'$

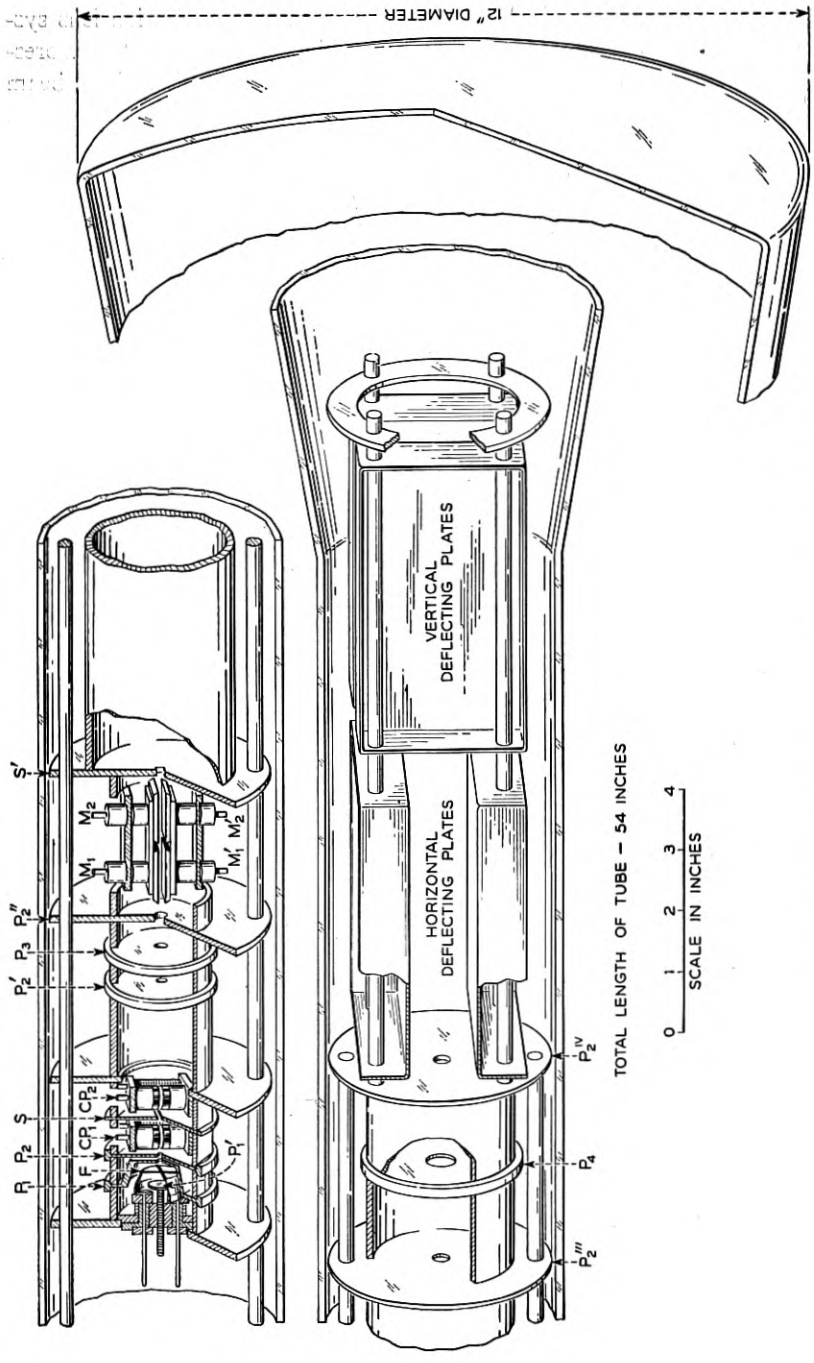



Fig. 3—Dimensional cross-section of electron optical system.

is added and cross connected to $M_1$ and $M_1'$ as indicated in Fig. 1(c). Now when signal voltage is applied, the electron beam is displaced, rather than deflected, across the aperture $S'$, and the portion of the beam entering the projection lens system is maintained in proper alignment with the tube axis.

The complete electron-optical system of the tube is indicated schematically in Fig. 2. A dimensional cross section of the tube is shown in Fig. 3,

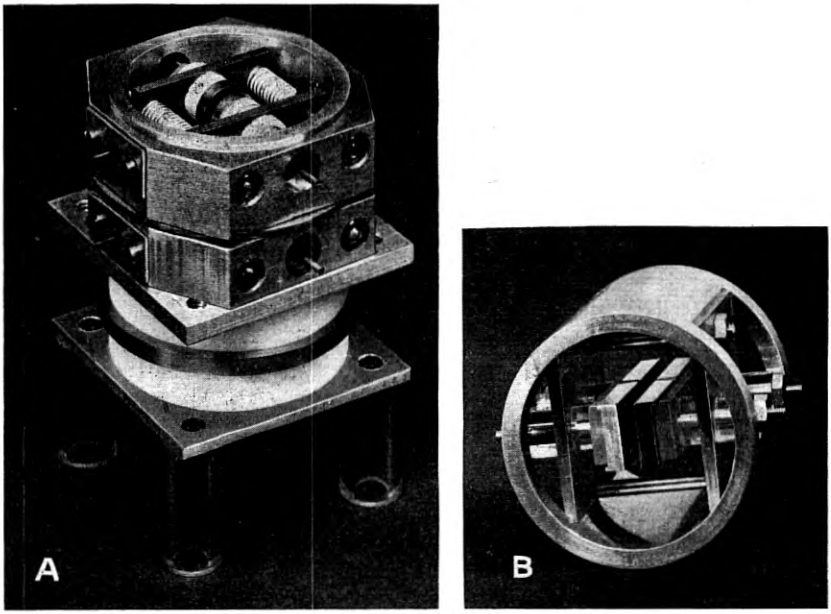



Fig. 4—Photograph of tube details.
(a). Photograph of collimating unit.
(b). Photograph of modulating unit.

while Figs. 4 and 5 show mechanical details of the assembly. Figure 6 shows the finished tube, held by Mr. Calbick.

Referring to Fig. 2*, the plates $P_2$, $P_2'$, $P_2''$, $P_2'''$ and $P_2^{IV}$ are metallically connected together with separating metallic cylinders to form the structural "frame" of the entire electrode assembly. They are connected to the internal Aquadag coating and held at ground potential.

The backing plate $P_1'$, filament F, and circular aperture plates $P_1$ and $P_2$ constitute the electron source and condensing lens system whose function

* This diagram actually shows the electrode voltages used in a later model for 441 line pictures. In the 240 line tube the anode voltage was 5000 volts.

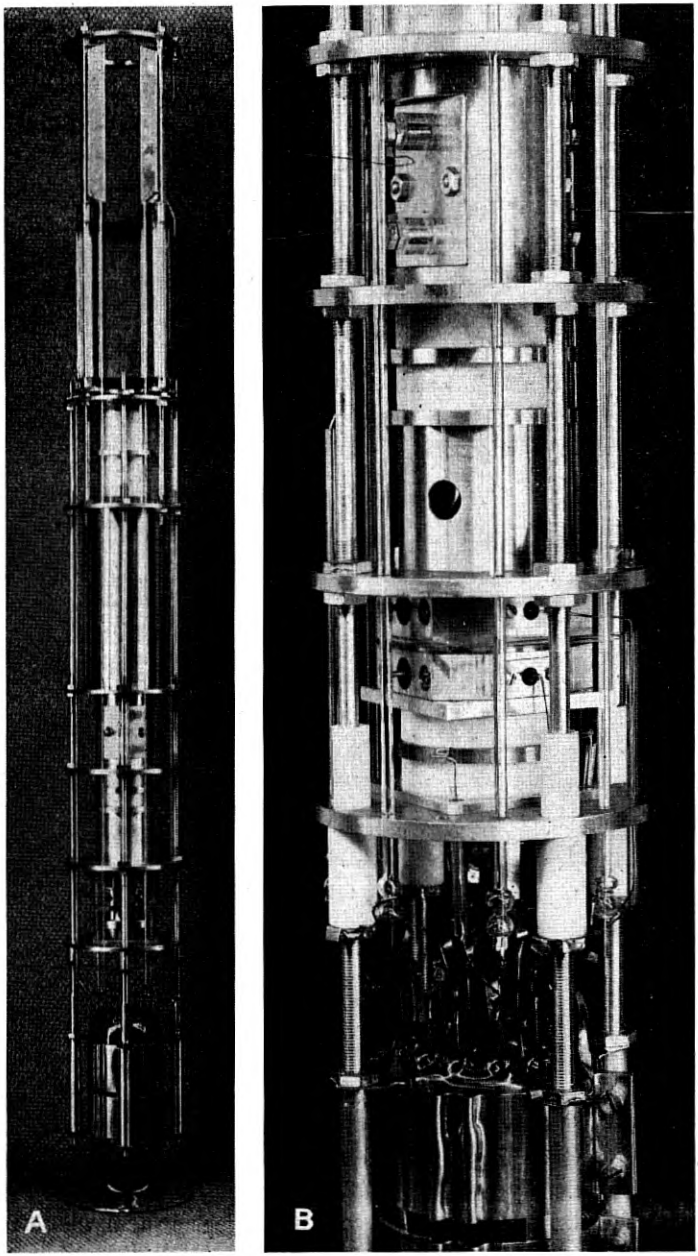



Fig. 5—Photograph of assembly.
(a). Photograph of over-all mechanical assembly.
(b). Photograph of assembly details.

is to produce an intense "focal spot" of electrons in the plane of the slit aperture $S$ (Fig. 1).

The filament F is made of tungsten foil in the shape of a cross (Fig. 3) and is fed by direct current with opposite ends of the cross connected together. This construction insures a nearly uniform temperature over the center area of the cross and also minimizes magnetic fields set up by the filament current, which would otherwise tend to disturb the electron optical symmetry. The filament resistance is approximately 1 ohm and the current 10 amperes. The high emission current density from a tungsten filament was expected to result in a more intense focal spot than would be obtainable with an oxide coated cathode. Also the large metallic structure inside the tube might tend to cause deactivation or poisoning of an oxide cathode.

If the mechanical alignment were perfect the "focal spot" of electrons would fall directly on the slit aperture $S$. Because of unavoidable small misalignments and particularly due to the fact that the filament cross is not perfectly flat, such perfect symmetry cannot be insured without some corrections. These corrections are supplied by a so-called collimating unit $CP_1$ ; $CC_1$ . The unit consists of an electromagnet with insulated pole-pieces (Fig. 4), and by applying small correcting voltages and currents to the unit the "focal spot" is shifted until its most intense part falls on the center portion of the slit aperture $S$. A second identical collimating unit $CP_2$ ; $CC_2$ is mounted after the slit, in order to center the slit image on the square aperture $S'$.

The three circular aperture plates $P_2'$ , $P_2''$ , and $P_3$ constitute the so-called modulator lens system and serve to form an electron image of the slit $S$ on the square aperture $S'$, with a magnification of 1:1. Accurate focussing is accomplished by adjusting the potential of the plate $P_3$ .

The function of the modulator plates $M_1M_1'$ and $M_2M_2'$ has already been described. The photograph in Fig. 4 shows the modulator plate assembly.

The set of three circular aperture plates $P_2'''$ , $P_4$ and $P_2^{IV}$ comprise the projection lens system, which forms an electron image of the square aperture $S'$ upon the fluorescent screen with a magnification of 5:1. Proper focussing is accomplished by adjusting the potential of plate $P_4$.

Two sets of coils, mounted in the lateral corners of the tube housing, were used to neutralize the earth's field. Each coil produces a field directed at an angle of 45° from the vertical and, by properly adjusting the coil current, it is possible to produce a practically uniform magnetic field, which was used to center the beam on the screen.

Because of the complicated mechanical assembly inside the tube there was originally some difficulty in properly de-gassing the tubes and maintaining vacuum. The earlier models were therefore continuously pumped,

Fig. 6—Photograph of finished tube.

and a tantalum filament manometer was used for checking pressure. The tubes used during demonstrations were sealed off, but the manometer was retained as shown in the photograph in Fig. 6. In this case the hot tantalum filament acted as a getter which successfully kept the pressure down to about $10^{-6}$ mm Hg.

The aperture $S'$ was .006″ square. With a 5:1 magnification this resulted in a scanning line height on the screen corresponding to a 240 line picture of about 7 x 8 inches.

A stationary spot viewed on the screen showed a sharply defined rectangular cross-section of approximately uniform brightness. Adjusted for no

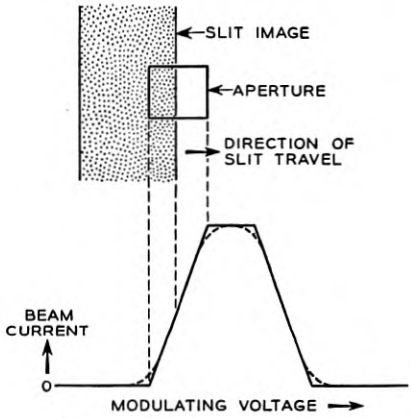



Fig. 7—Schematic diagram of modulation curve.

overlap this resulted in a flat field with only a faint indication of line structure. If the image of the slit $S$ falling on the square aperture $S'$ is perfectly focussed, with edges parallel to the sides of the aperture; if no stray electrons are present due to secondary emission or other causes, and if the electrons in the beam all have the same thermal emission energy, then the curve of beam current versus modulator voltage would be as shown in Fig. 7. The current is zero until the leading edge of the slit enters the aperture. The current then increases linearly until it reaches a maximum when the slit fills the aperture. If the slit is wider than the aperture the curve will have a flat top and will then decrease linearly as the trailing edge of the slit travels across the aperture.

The actual modulation curve did not show sharp corners, but was rounded both at top and bottom as indicated by the dotted lines in Fig. 7. The dispersion of thermal velocity of the electrons causes "chromatic" aberration of the condensing lens system ($P_1$ ; $P_2$) which therefore forms a

"focal spot" of non-uniform density. This in turn results in rounded corners of the modulation curve. Even so, the linear portion of the modulation curve corresponded to a beam current ratio of about 10:1.

As will be seen from the curve, the tube may be used equally well for either positive or negative modulation, but in the demonstrations positive modulation was employed.

An actual modulation curve for the tube later used for 441 line pictures is shown in Fig. 8. The dotted line indicates the modulation characteristic without any modification, while the solid curve shows the improvement

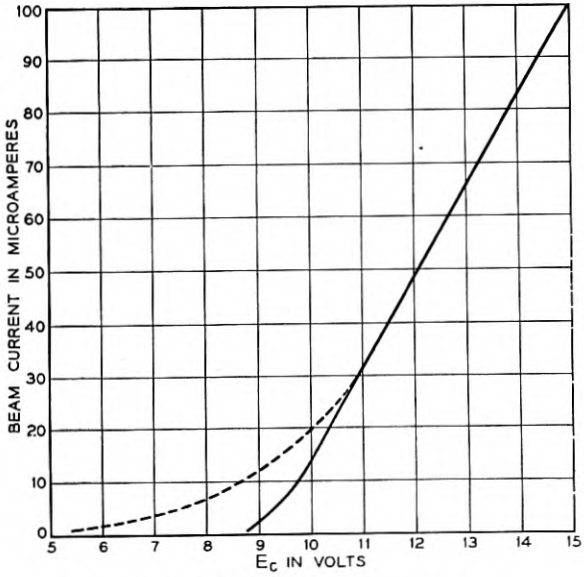



Fig. 8—Measured modulation curve.

near cut-off obtained by incorporating a non-linear circuit in the output stage of the video amplifier. With this circuit, linear modulation was obtained over a brightness range of nearly 100 to 1.

Due to the variable width of the scanning spot it is obvious that conventional aperture equalization is not applicable. The width of the rectangular spot changes from maximum at full brightness to zero or nearly zero in the deep shadows. In other words, the correct aperture equalization would be a function of brightness. Some theoretical computations indicated that the effective horizontal resolution, without any aperture equalization, might be well above the vertical resolution determined by 240 lines. That this was actually so was indicated by the fact that, by unbalancing the

coaxial terminal equipment so as to allow a small amount of 2500 kc carrier leak to come through, the resulting vertical stripes were clearly visible, in spite of the fact that the highest frequency in the video signal was only about 800 kc.

For the reasons given above no aperture equalization was employed for the receiver; in fact, the excessive horizontal resolution was later on traded for higher brightness as mentioned below.

A slightly modified version of the Davisson tube was used later in the 1941 demonstrations of the transmission of 441 line, 30 frame television signals over the 2.7 megacycle coaxial cable from New York to Philadelphia.† In this later tube the anode voltage was raised to 10,000 volts instead of 5000 volts as used in the 1937 demonstrations. Furthermore, the square aperture $S'$ was changed to a rectangular aperture with the horizontal side twice as long as the vertical. This alone of course doubled the highlight brightness obtainable, and in the 1941 demonstrations the received pictures had a highlight brightness of about 20 foot-lamberts, using an aperture size of .0036″ x .0072″.

Before concluding the description of the Davisson tube and its performance, one more item should be mentioned. It was found that the glass thickness of the tube's end wall gave rise to some halation due to internal reflection from the outer surface, and this in turn resulted in a somewhat degraded contrast range. A very much thicker glass wall would increase the diameter of the circle corresponding to total reflection to a point where the halation effect would be very much diluted and therefore negligible. The effect of such a thick wall was obtained in the following manner.[4] A plate glass disk was mounted between one and two inches in front of the tube fact and cemented to this by means of an airtight gasket. The intervening space was then filled with Nujol, which has approximately the same index of refraction as the glass. The resulting increase in contrast range was very noticeable.

It is interesting to note that at the time it was also proposed to add a small amount of dark dye to the Nujol in order to further decrease halation effects, and also to decrease the effect of ambient light. This is, of course, the same principle that is now widely employed in present day "dark glass" kinescopes.

Dr. Davisson designed this tube on the basis of his knowledge of electron optics. At no stage did he depart from a design which would allow him accurately to predict the performance. This accounts for the "thin" lenses used in the different focussing systems, for the small deflection angles em-

† The transmitting equipment for these demonstrations was a film scanner employing a Farnsworth image dissector and producing a 4 mc video signal. See reference[3].

ployed to insure sharp focus all over the screen, and for the extreme care with which the deflection plate system was made to avoid either "pin cushion" or "barrel" distortion. Apart from the rounded corners of the modulation curve, the entire system was indeed "calculable." It resulted in a very long tube (about 5 feet) and an unusually complex assembly of precision mechanical parts. It also resulted in an actual performance very close to the predicted performance and markedly superior to that of other television receiving tubes of the same period.


REFERENCES

1. "Coaxial Cable System for Television Transmission," M. E. Strieby—*Bell Sys. Tech. Jl.*, Vol. 17, pp. 438–457, July 1938.
2. K. K. Darrow's article in this issue.
3. "Film Scanner for Use in Television Transmission Tests," A. G. Jensen—*I. R. E., Proc.*, Vol. 29, pp. 243–249, May 1941.
4. *U. S. Patent No.* 2,312,206, issued to C. J. Calbick.

# Electron Transmission Through Thin Metal Sections with Application to Self-Recovery in Cold Worked Aluminum

By R. D. HEIDENREICH

Features of the dynamical or wave mechanical theory of electron diffraction pertinent to the interpretation of electron images of crystalline materials are briefly discussed. It is shown that the type of image obtained depends upon the local bending of the crystal, the coherent crystal size and the state of internal strain. The absence of extinction contours seen in annealed aluminum is an indication of the presence of internal strains.

New data concerning the effect of temperature on the polygon or domain size in cold worked aluminum is presented. These data indicate that self recovery occurs rapidly at temperatures as low as −196°C leading to the conclusion that the process must have a low activation energy. The mechanism by which dislocations leave the slip bands and redistribute and annihilate themselves during recovery is obscure.

## Introduction

DIRECT experimental evidence for the wave nature of the electron first demonstrated by Davisson and Germer was based on a property unique to wave phenomena; namely, interference or diffraction. The ability of a regular or periodic array of atoms to diffract electrons (just as a ruled grating diffracts light) has led to the construction of quite general theories of the behavior of matter, the band theory of solids being a notable example. The forbidden energy zones in a crystal are simply a result of diffraction of the valence electrons by the periodic structure. On the other hand, the results of straight forward diffraction experiments are a consequence of the zone or band structure of the crystal thus illustrating an interesting closure or completion of the cycle.

This paper is concerned with the interpretation of electron interference phenomena occurring in thin metal sections particularly as it pertains to structural changes accompanying plastic deformation. Diffraction effects are observed not only in the usual electron diffraction methods but in electron microscope images as well. The chief difference is that in the former the diffracted rays are of primary interest while in the latter the regions of the crystal in which diffraction occurs are imaged with the diffracted beams removed by the objective aperture. Electron microscope images of crystalline materials thus offer a high resolution method of studying variations in diffracting power. This information can then be interpreted in terms of structural features.

The structural changes accompanying plastic deformation of metals are of great interest and can be profitably investigated by electron interference

techniques. In particular the loss of work hardening in pure metals by the process of self-recovery is well suited to study using the electron microscope and thin sections, as will be seen. In a previous paper[1] the technique of preparing thin aluminum sections and the interpretation of the diffraction features were discussed in detail. The dynamical theory of electron diffraction has been applied to this particular problem and also extended to more general cases.[2] It was shown[1] that, immediately after cold working, high purity aluminum exhibits a sub-grain or domain structure of the order of 1–2 microms in size. These domains are slightly misoriented and become visible in electron images through small variations in diffracted intensity in passing from one domain to the next. The domains have been identified with self-recovery following plastic deformation but their properties and origin have not been investigated in detail. The experiments to be described here include the effect of electron bombardment, and the effect of temperature on the size of the domains.

It should be pointed out that the term "recovery domains" has been applied by the writer to the sub-grains or units appearing in cold worked aluminum. It appears now that the recovery domains are an early stage of the process of "polygonization" commonly used in the literature. Polygonization was first applied to the formation of larger blocks or polygons in bent crystals that were annealed at an elevated temperature. The polygons were made visible in a light microscope by the use of an etchant which produced etch pits in the polygon boundaries.[3] Either term, polygonization or recovery domains, would be suitable although the latter is more specific as to the process involved and will be used in this paper.

## Dynamical Theory Applied to Electron Images

The wave mechanical or dynamical theory of electron diffraction is essential to the interpretation of electron images of crystals. The kinematic theory, a simpler approximation, is not adequate. Consequently, it is advisable at this point to review briefly the salient features of the dynamical theory as applied to electron images.*

Suppose a polycrystalline film (such as a thin metal section) is mounted in the conjugate focal plane of the objective lens of an electron microscope as depicted in Fig. 1. Let the incident electron beam be taken as monochromatic and plane-parallel. Regions of the specimen film oriented such that a set of net planes satisfies the Bragg condition $n\lambda = 2d \sin \theta$ will

[1] R. D. Heidenreich, *Jl. App. Phys. 20*, 993 (1949).
[2] R. D. Heidenreich, *Phys. Rev. 77*, 271 (1950).
[3] P. Lacombe and L. Beaujard, "Report of a Conference on Strength of Solids," p. 91 (Physical Society, London, 1948).
* Detailed treatments are given in references 1 and 2.

diffract the incident beam of wave length $\lambda$ through an angle $2\theta$. $d$ is the interplanar spacing. The numerical aperture of the electron lens is such that in general $2\theta > \alpha_{ob}$ with the result that the diffracted beams are removed from the optical system and not focused in the image plane. In Fig. 1, crystals 1 and 3 are suitably oriented to diffract and so will appear dark in the final image since electrons have been removed from these regions by Bragg reflection. The calculations of the intensities for this case (the Laue case) were first made by Bethe.[4] A similar treatment[1,2] employing the zone theory of crystals can be given which yields the same final results. The procedure consists in solving the Schrödinger equation for an electron moving in the

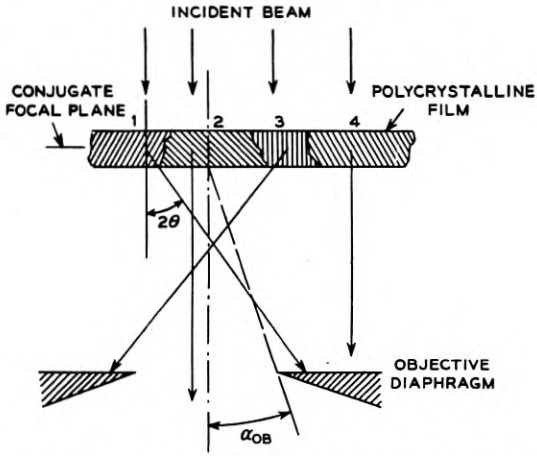



Fig. 1—Diffraction of electrons outside objective aperture by suitably oriented crystals in polycrystalline film. Crystals 1 and 3 would appear dark in final image.

periodic potential inside the crystal and then fitting the solutions so obtained to the plane wave solution found for the vacuum incident and diffracted waves. The first result of the solution inside the crystal is that the total energy of the electron, $E$, is not a continuous function of the wave number $K$ ($|K| = \frac{2\pi}{\lambda}$) as it is in field-free space, but exhibits discontinuities as illustrated in Fig. 2. These discontinuities occur whenever the Bragg or Laue condition is satisfied; i.e., if $g$ is a vector of the reciprocal lattice, then discontinuities in the $E$ vs $K$ curve occur when $|K| = |K + 2\pi g|$ which is equivalent to the Bragg formula. The magnitude of the energy

[4] H. A. Bethe, *Ann. d. Physik 87*, 55 (1928). This treatment was intended to explain the results of Davisson and Germer which were published about a year before. Had Bethe examined the behavior of the total energy in his solution of the Schrödinger equation he would have discovered the band theory of crystals.

gap[5] is $\Delta E = 2|Vg|$ where $Vg$ is the Fourier coefficient of potential. The discontinuities in energy are the Brillouin zone boundaries and form a family of polyhedra in reciprocal or $K$ space. For example, a simple square lattice gives rise to a square reciprocal lattice as seen in Fig. 3. One reciprocal lattice point is taken as the origin, 0, and the remainder of the lattice is generated by the vector $g$ where $|g| = \frac{1}{a}$, the reciprocal of the cell constant of the original, direct lattice. The Brillouin zone boundaries are the perpendicular bisectors of the reciprocal lattice vectors[6] and define the series of zones shown in Fig. 3a. Whenever the incident electron wave vector, $K$,

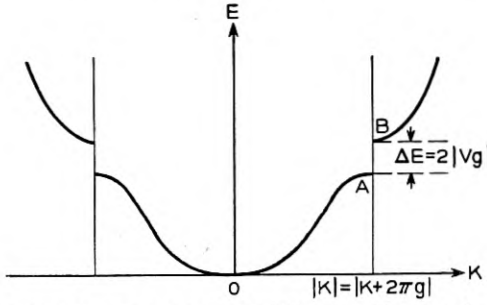



Fig. 2—Plot of energy, $E$, vs. wave number $|K|$, along $K$ vector showing discontinuity when $|K| = |K + 2\pi g|$ or when Bragg condition is realized.

terminates on a Brillouin zone boundary, a diffracted wave is possible. However, when the boundary conditions at the surfaces of the crystal are applied, it turns out that for a fixed total energy, $E$, there are two incident crystal wave vectors $K_0^0$ and $K_1^1$ which must be considered. Consequently there are also two diffracted wave vectors $K_g^0$ and $K_g^1$ with $K_g^0 = K_0^0 + 2\pi g$ and $K_g^1 = K_0^1 + 2\pi g$ as shown in Fig. 3b. $K_g^1$ and $K_g^0$ are related by a beat wave vector $\Delta K$ or $K_g^1 = K_g^0 + \Delta K$. The net result is two waves of slightly different wave length traveling nearly parallel which may undergo interference. This beating of the diffracted waves makes itself known by passing the energy back and forth between the incident and diffracted beams. This is the motivation for the name "dynamical" theory.

The intensity of a diffracted beam for the case when the incident wave vector terminates near a Brillouin zone boundary but far from an edge or corner is found to be[1]:

[5] This treatment is the case of loose binding which is applicable for fast electrons. It turns out that for the valence electrons in a crystal, this approximation is not very good and that the value $\Delta E = 2|Vg|$ is not correct.

[6] L. Brillouin, "Wave Propagation in Periodic Structures," McGraw-Hill Book Company, Inc., New York (1946).

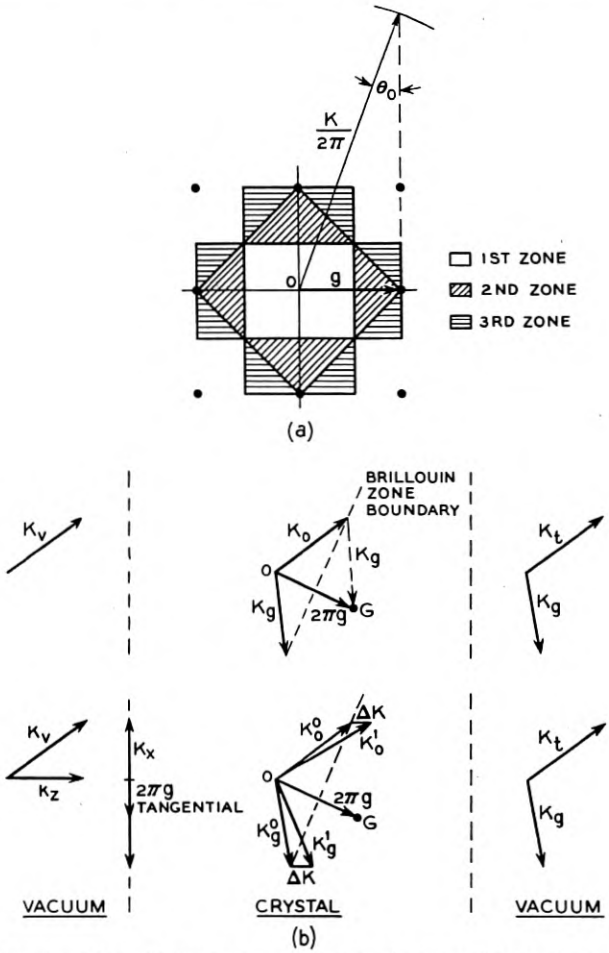



Fig. 3—(a). First three Brillouin zones for a simple square lattice. In an actual case, $|g|$ is of the order of $0.5A^{-1}$ and $\left|\frac{K}{2\pi}\right|$ about $19A^{-1}$.

(b). Relation among wave vectors in reciprocal space for the kinematic and for the dynamical theories. The wave vectors must satisfy the Bragg condition and the boundary conditions at the crystal faces. The dynamical theory introduces the beat wave vector $\Delta K$. $K_v$ is the vacuum incident wave vector.

$$Ig = \frac{|Vg|^2}{\left(\frac{\Delta g}{2}\right)^2 + |Vg|^2} \sin^2 \tfrac{1}{2}\Delta Kz \tag{1}$$

where $Vg$ = Fourier coefficient of potential

$\Delta g$ = deviation from the Laue condition in volts

$= 2E\Delta\theta \sin 2\theta_0$

$E$ = total energy of incident electrons in volts

$\theta_0$ = Bragg angle

$\Delta\theta$ = angular deviation from Bragg condition

$\Delta K$ = beat wave vector

$$= \frac{2}{\lambda E}\left(\left(\frac{\Delta g}{2}\right)^2 + |Vg|^2\right)^{\frac{1}{2}}$$

$z$ = penetration measured normal to surface of crystal

$D$ = thickness of crystal

It is evident from equation (1) that the intensity of the diffracted beam is periodic with penetration in the crystal and with the deviation, $\Delta g$. This dependence of intensity upon thickness and deviation accounts for most of the image detail seen in electron micrographs of thin crystalline sections. Inelastic scattering and crystal imperfections are neglected in the derivation of equation (1).

Experience with thin sections of pure aluminum has indicated that it is nearly impossible to prepare and handle them without introducing some bending or rumpling of the thin area. This bending in conjunction with thickness variations gives rise to the major features of the electron images in the form of intensity maxima and minima called "extinction contours." Those arising through bending of the section are of chief interest in the uniform area where thickness changes are very gradual. The extinction contours are determined by the maxima of equation (1) or where $\Delta KD = n\pi$ to give

$$\Delta g = \pm\left(\frac{(n\lambda E)^2}{D} - 4|Vg|^2\right)^{\frac{1}{2}} \qquad n = 1, 3, 5, \ldots \tag{2}$$

Equation (2) predicts that for a bent crystal offering a continuous range of $\Delta g$ a series of intensity maxima or fringes will be observed. In an electron image of a bent crystal the spacing of the fringes is the only quantity which can be measured other than relative intensity. The central fringe corresponds to $\Delta g = 0$ with subsidiary maxima occurring at a distance $s$ from the central fringe given by[1]

$$s = \left(R + \frac{D}{2}\right)\frac{\Delta g}{2E \sin 2\theta_o} \tag{3}$$

where $R$ is the radius of curvature of the bending.

If two crystallites in a thin section differ only slightly in orientation and the bending is favorable, then a series of fringes will occur in the two crystals with a displacement at the boundary as sketched in Fig. 4.

The displacement $l$ is related to the orientation difference $\Delta\alpha$ and the

radius of curvature $R$ by

$$\Delta\alpha = \frac{l}{R} \tag{4}$$

If the situation is such that $R$ can be determined from equation (3), then the misorientation $\Delta\alpha$ can be calculated from (4). This combination of circumstances is at present one of chance and does not occur frequently in images of thin sections. An example will be shown.

The image contrast between adjacent regions of a thin metal section composed of small misoriented domains as depicted in Fig. 1 is determined by the rocking curve of equation (1). This is simply a plot of the intensity given by (1) against $\Delta g$ or $\Delta\theta$. As an example, a rocking curve for the (200)

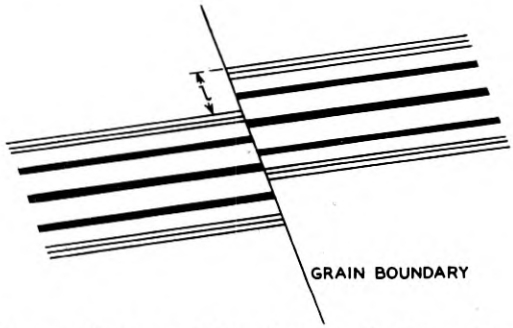



Fig. 4—Displacement ($l$) at a grain boundary of extinction contours due to bending as predicted by equations (3) and (4).

reflection of aluminum is shown in Fig. 5. The Fourier coefficient $Vg$ is computed from the relation.*

$$V_{hkl} = 300\frac{ed^2_{hkl}}{\pi\Omega}\sum_j (Z_j - f_j)e^{2\pi i}(hu_j + kv_j + lw_j) \text{ volts} \tag{5}$$

with $e$ = charge on the electron = $4.80 \times 10^{-10}$ *esu*
$\Omega$ = volume of the unit cell = 70.4 $A^3$ for aluminum
$d_{hkl}$ = interplanar spacing
$Z_j$ = atomic number of atom species $j$
$f_j$ = atom form factor
$(u_j, v_j, w_j)$ = atomic coordinates of atom species $j$

For aluminum, $V_{(200)} = 5.13$ volts. Using 50KV electrons: $E = 5 \times 10^4$, $\lambda = 0.055$A, $\sin 2\theta_0 = 0.0272$ radian and a reasonable value of $D = 250$A, the intensity can be computed from (1) as a function of $\Delta g$ as shown in Fig. 5.

* Reference 1, Appendix I.

It will be realized that the detailed shapes and location of the maxima are quite sensitive to thickness and more often than not a calculation from the fringe spacings cannot be made with certainty. The limiting value of the fringe separation is found from (2) to be given by $\frac{d_{hkl}}{D}$ for large values of the integer $n$. The fringe pattern indicated in Fig. 4 is simply a rocking curve for the crystallites with a displacement due to their difference in orientation.

The absence of extinction contours from the electron image of a crystal may indicate that one or more of the following conditions exists:

(1) No bending or thickness changes.

(2) The thickness is sufficiently small that the argument of the $\sin^2$ in

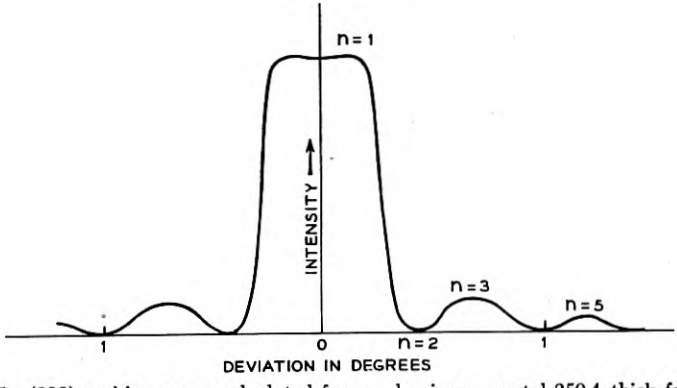



Fig. 5—(200) rocking curve calculated for an aluminum crystal 250$A$ thick for 50$KV$ electrons.

(1) can replace the sine function thus suppressing the periodic features; this can occur for $D$ less than about 100A.

(3) The crystal is sufficiently distorted that the assumption of a periodic potential function in the Schrödinger equation is not valid.

Of these causes for the absence of extinction contours, the first is very unlikely as mentioned previously. The second is quite obvious since a section too thin to produce contours would give a very high transmitted intensity and would be immediately apparent. The third is the most likely reason and is thought to be the case in all the thin sections examined to date. This is particularly important here since crystals that have been subjected to plastic deformation are of primary interest. The incorporation of strains or lattice disortion into the potential function for the Schrödinger equation appears to be a formidable task and will not even be attempted. A more or less semi-quantitative approach to the effect of lattice distortion

can be had by considering the expression for the diffracted intensity (equation (1)). If dislocations are introduced into a crystal the effect is that of producing a mosaic or block structure[7] the units of which will scatter incoherently. If there are a sufficient number of dislocations to reduce the coherent penetration path $z$ to a value $z'$ such that $\Delta K z'$ is small, then equation (1) will become

$$Ig \simeq \frac{|V_g|^2}{\left(\frac{\Delta g}{2}\right)^2 + |V_g|^2} \tfrac{1}{4} (\Delta K z')^2 = \frac{\pi^2}{\lambda^2 E^2} |V_g|^2 (z')^2 \tag{6}$$

The important part in considering equation (6) is the disappearance of the periodic, dynamic term. If the model obtained by introducing dislocations into the crystal even roughly approximates the actual situation then it would be expected from equation (6) that the extinction contours will vanish. Actually, starting with a perfect crystal and adding dislocations, the dynamical effects will not be noticeably effected until the coherent penetration $z'$ becomes much smaller than $\frac{\lambda E}{V_g}$. For 50 KV electrons in aluminum, $z'$ would have to be something of the order of 150A or less before the dynamical effects would disappear. If the model is carried still further, it can be speculated that $z'$ is the mean separation of dislocations in the crystal indicating that a distance of separation of the order of 150A is required to extinguish the extinction contours. This would correspond to a dislocation density of about $5 \times 10^{11}$ lines/cm$^2$. This is admittedly a very crude approximation and, although the dislocation density is of the right order of magnitude, too much significance should not be attached to it. It does seem fairly safe to conclude that the extinction contours will disappear when the dislocation density reaches a high enough value. This fact in itself greatly broadens the interpretation of the electron micrographs to be presented.

## Preparation of Thin Aluminum Sections

The details of the preparation of the thin sections used for electron microscopy have been published[1] and are not vital to the discussion. Suffice it to say that the sections are produced from 0.005″ sheet by an electropolishing technique using a special holder. The central portion of the metal disc is thinned down to several hundred Angstroms or less while maintaining a smooth surface. A rinsing procedure is necessary to prevent the formation of corrosion layers.

[7] R. D. Heidenreich and W. Shockley, Report of a Conference on Strength of Solids, p. 57 (Physical Society, London, 1948).

The metal used in all this work was 99.993% French aluminum rolled into 0.005″ sheet.

## Recovery of Cold Worked Aluminum

Having briefly discussed the essential phenomena in interpreting electron images of crystals, the application of the thin section method to self-recovery in deformed aluminum can be demonstrated. This type of investigation is based to a considerable extent upon comparison of images of the metal

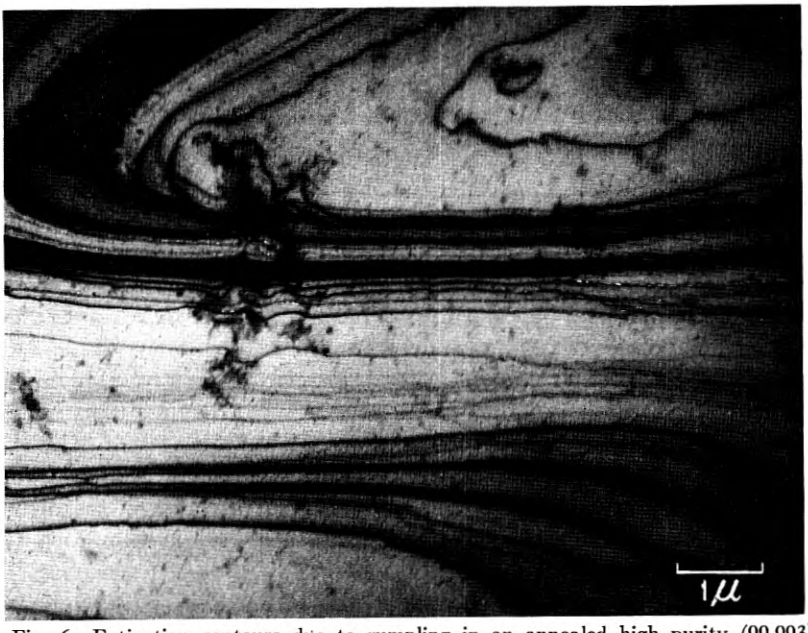



Fig. 6—Extinction contours due to rumpling in an annealed high purity (99.993) aluminum section.

under various conditions of anneal and plastic deformation. The standard state for comparison is a well annealed specimen in which the crystals are reasonably perfect. The bending or rumpling produces the extinction contour patterns quite unique to the annealed condition. The chief characteristics of the contours for an annealed crystal are their general continuity and extension over relatively large areas. Figure 6 illustrates a contour pattern with an unusually high density of lines obtained from an aluminum section annealed 30 min. at 335°C. The dark regions are those of electron deficiency.

At a grain boundary in an annealed section the contours end abruptly

as seen in Fig. 7a. Figure 7b illustrates a case in which the contour family can be identified on either side of a grain boundary with the displacement very much in evidence at the boundary. This situation was anticipated in Fig. 4.

Proof that the dark contour lines seen in Fig. 6 are due to diffraction from those regions was given in reference (1) where both the transmitted and diffracted beams were imaged in an electron shadow microscope. The usual transmission electron diffraction patterns from annealed sections generally exhibit an array of spots characteristic of a single crystal. Sometimes weak, broad Kikuchi lines are obtained but generally the bending of the section and the area of the incident electron beam are such as not to favor Kikuchi lines.

The effect of cold working on the images of the sections was studied by lightly pounding the center region of a $\frac{1}{8}''$ diameter disc (0.005″ thick) with a small, rounded and polished steel rod against an anvil. The disc was then electro-thinned and examined in the electron microscope. Originally it was hoped that further information regarding laminar slip[7] might be obtained in this manner. However, no details of slip have been observed in the cold worked sections, the general appearance being that seen in Fig. 8(a). Figure 8 was obtained from a section cold worked by pounding at room temperature and shows the recovery domains or early stage of polygonization.[8] These domains are not made visible by etching a polished surface and are observed only by electron transmission. The domains are slightly disoriented one with respect to the other and are made visible by the differences in diffracted intensity. Since the extinction contours are absent in Fig. 8a it is concluded that there is considerable internal strain in the domains as previously mentioned. Figure 8b is a transmission electron diffraction pattern of this section and shows arced rings made up of discrete spots. It is concluded that each spot on an arc corresponds to a domain. Insufficient domains are included in the primary beam to produce continuous rings. The electron diffraction pattern of Fig. 8b is very similar to microbeam x-ray patterns published by Kellar[9] et al and the domain size of about $2\mu$ from the electron micrographs is in excellent agreement with the results of Kellar for pure aluminum.

That domains sufficiently free of strain to yield extinction contours can be obtained is illustrated in Fig. 9. Figure 9a is from a section prepared 36 hours after a block of the high purity aluminum had been rolled to 0.005″ sheet with some annealing between passes. The contours are very much in

[8] Recovery domains are not found in aluminum deformed by simple extension. Apparently inhomogeneous strain is necessary. Extinction contours are observed in specimens deformed in tension but the slip bands are not in evidence.

[9] J. N. Kellar, P. H. Hirsch and J. S. Thorp, *Nature 165*, 554 (1950).

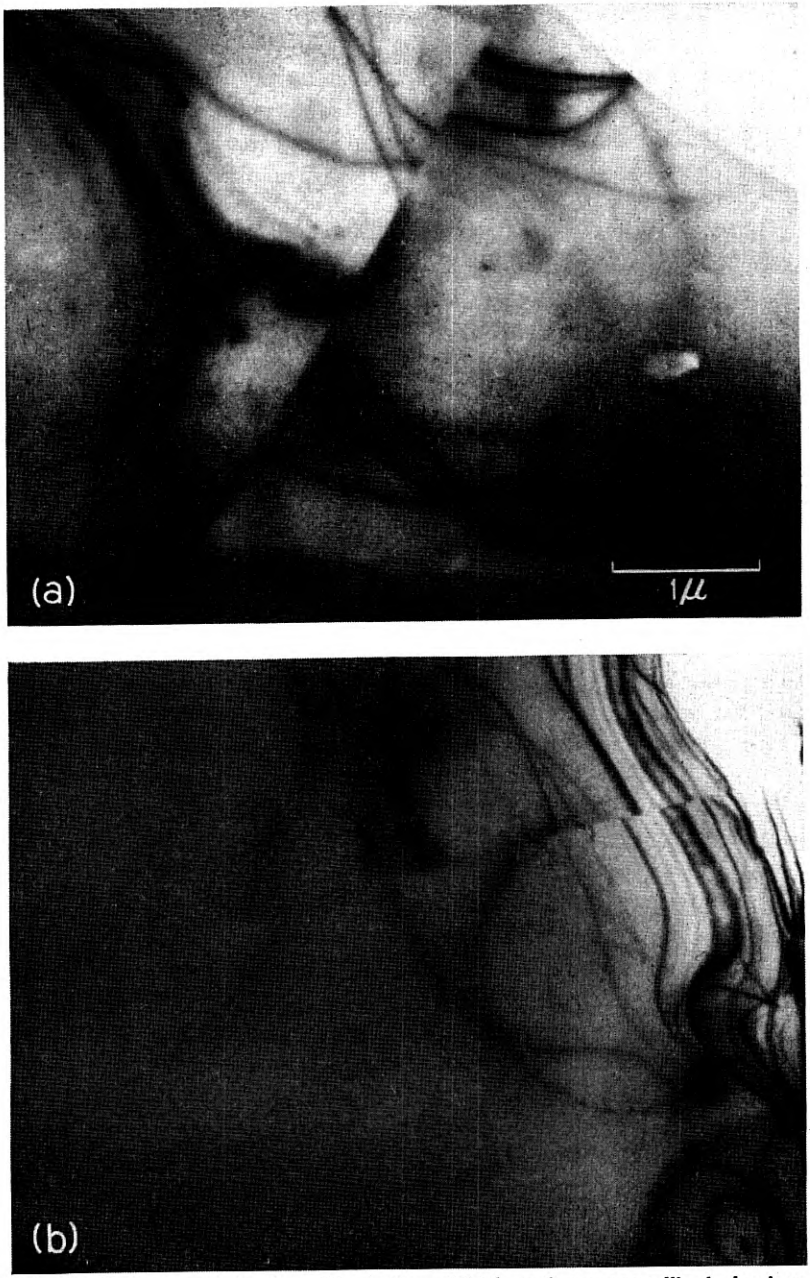



Fig. 7—Grain boundaries in a thin section of high purity, recrystallized aluminum. The displacement of a family of contours at the boundary is evident in (b).

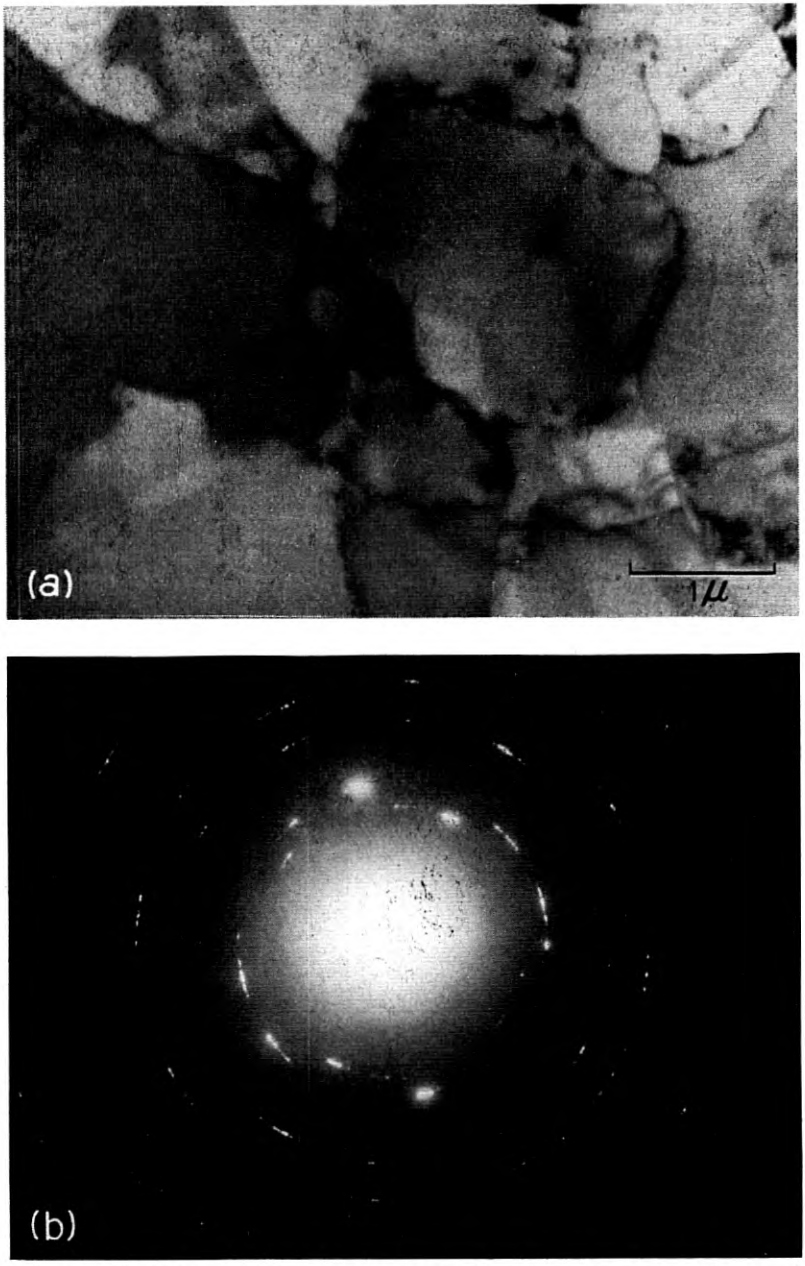



Fig. 8—Thin section of high purity aluminum cold worked by pounding. The recovery domains are evident in (a). (b) is an ordinary electron diffraction pattern (transmission) of the section and shows the discrete spots on the arced rings.

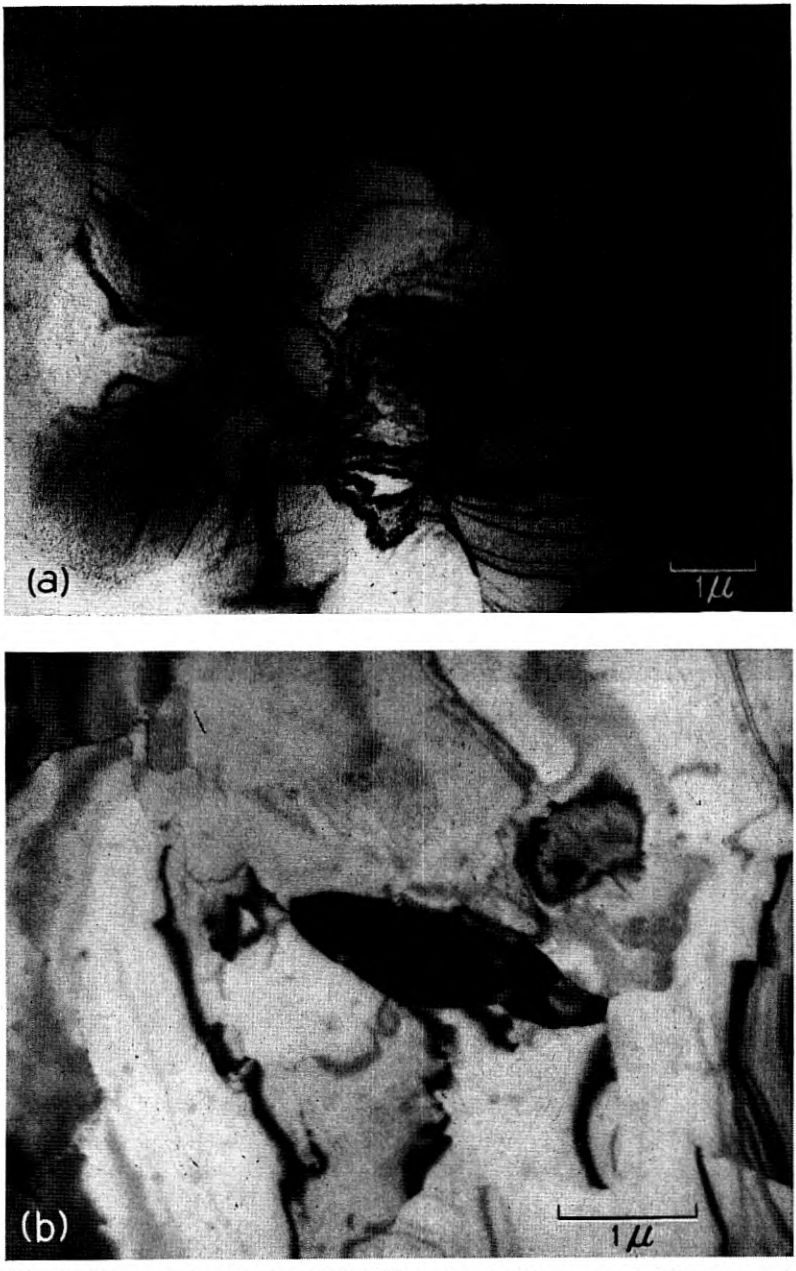



Fig. 9—Recovery domains in rolled, high purity aluminum annealed between passes through the rolls.

(a). 36 hrs. after rolling showing extinction contours in the domains.

(b). 1 year after rolling. The domain boundaries are evident now only through the discontinuities of the extinction contours.

evidence indicating that the internal strain is less than in Fig. 8. Another section was prepared after the rolled sheet had stood for about one year with the result shown in Fig. 9b. The domain boundaries in this section are made visible chiefly through the discontinuities in the extinction contours rather than overall contrast between domains. Apparently the domains slowly relieve their internal strains and become less distinct as recovery proceeds at room temperature.

A possible complication in the study of thin sections in the electron microscope is the effect of rather intense electron bombardment. There are several possible ways in which the sections might be changed by bombardment. One of these is simply annealing due to heating by bombardment. However, the metal is a good conductor of heat and is in contact with the heavy brass specimen holder so that high local temperatures would not be expected as with thermal insulators or isolated particles. Another phenomenon that is quite common is the deposition of a carbonaceous layer on the areas exposed to the electron beam. This is generally due to the residual hydrocarbon atmosphere in the vacuum system and is visible to the naked eye as a black deposit. The remaining possibility is that of producing lattice defects or vacancies by collision with the incident electrons. The cross-section for this process is not known but it would be expected that for 50 KV electrons it would be quite small.

Many thin sections of aluminum have been examined in the RCA EMU instrument at moderate intensities using the biased electron gun with little evidence for any changes occurring over the normal times required for obtaining pictures. However, if the peak intensity attainable with the biased is used, quite significant changes occur as illustrated in Fig. 10. These images are taken from a sequence and show the effect of time of bombardment on the recovery domains. The loss of contrast and irreversible changes in details with time of bombardment are evident. Part of the effect is due to heating and part to deposition but in general the behavior is not understood.

An outstanding feature of the domains in cold worked, high purity aluminum has been the relatively uniform size exhibited over a great many samples prepared at room temperature. The deformations have ranged from the order of about 30% to several hundred percent with the domain size consistently in the neighborhood of $2\mu$. At low deformations of the order of a few percent the domains are not found. No growth or change in size of any consequence has been found after months at room temperature. The relief of internal strains seems to be the only significant change with time. A short anneal at or above the recrystallization temperature removes the domains and gives rise to new crystals which exhibit extinction contours. Observations such as this tie the domain structure quite firmly to recovery.

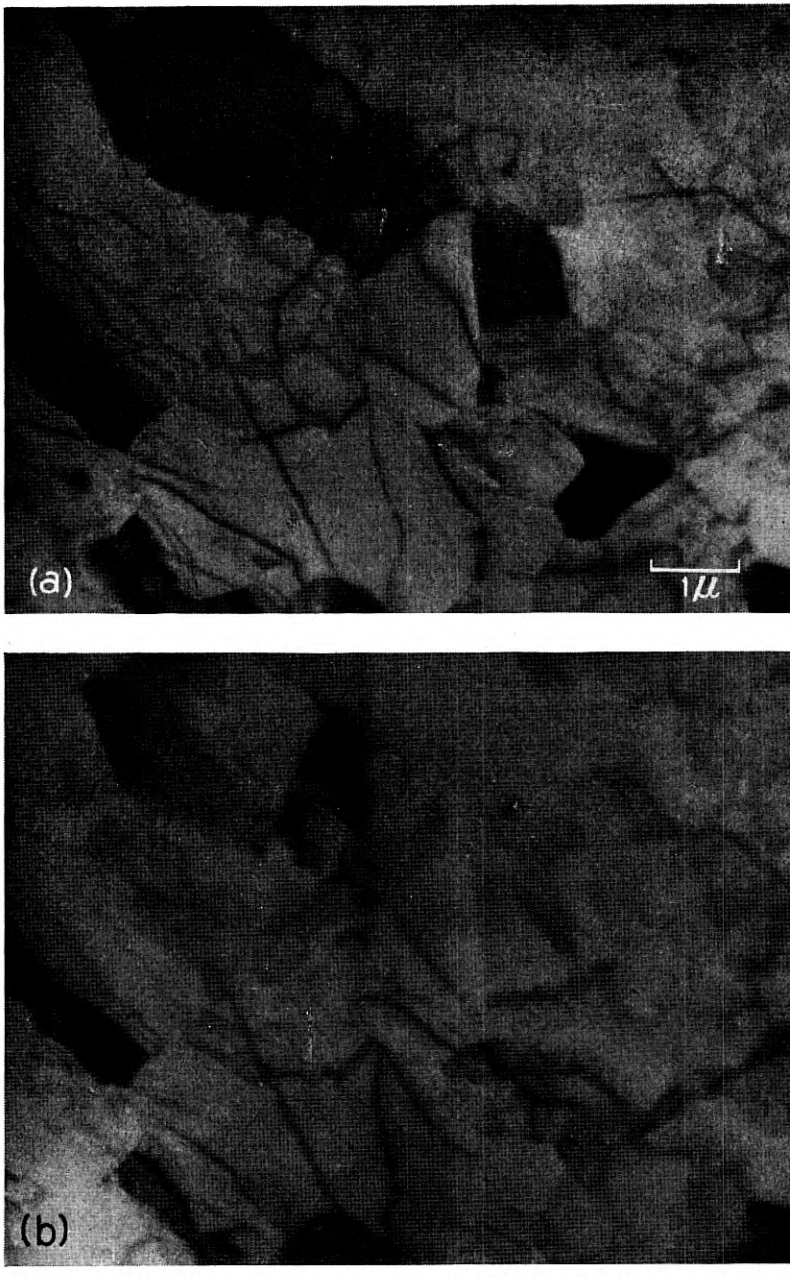



Fig. 10—Effect of intense electron bombardment (5$OKV$ electrons) on the domain structure.
(a) 45 seconds bombardment
(b). 125 seconds bombardment

As previously pointed out, the formation of recovery domains upon cold working represents an early stage of polygonization with a much smaller size than that observed at low deformation.[10] This suggests a nucleation and growth process[11] for recovery during and following plastic deformation. Cahn[12] has published a detailed nucleation theory to account for recrystallization grain size. He considers the nuclei to be regions in the crystal with large curvature brought about by cold working. It would be expected that both the nucleation and growth rates would be effected by changes in temperature and by additions of alloying elements which reduce the rate of diffusion of dislocations. The latter was tried first by adding about 4% copper to the high purity aluminum and rolling the alloy to 0.005″ sheet. A section prepared from the rolled sheet with no anneal gave the results shown in Figure 11. It will be noted in Fig. 11a that the large, well defined domains seen in Figs. 8 and 9 are not present in the alloy. The structure is much smaller and is strung out in the direction of rolling. The electron diffraction pattern, (Fig. 11(b)) exhibits arced rings with the arcs being continuous rather than showing the discrete spots seen for the pure metal (Fig. 8b). The number of recovery domains produced in the alloy is thus much greater than in the base metal which checks with the much smaller recovery exhibited by cold worked aluminum alloys as compared to pure aluminum. It is concluded that the addition of copper has produced "knots" in the aluminum lattice which impede the rate of growth of domains.[13]

The effect of temperature is of much interest since nucleation processes generally involve a temperature dependent term of the form $e - \frac{A}{KT}$ where $A$ is an activation energy[14] A plot of nucleation rate against temperature should yield a curve exhibiting a maximum at some temperature $T_c$. For $T > T_c$, only a portion of the embryos are able to exceed the critical size, the smaller ones dissociating. For $T < T_c$, the thermal diffusion rates are sufficiently low to impede embryo formation. The maximum nucleation rate is thus a balance between the diffusion rate and the number of embryos able to exceed the critical size and grow. In order to investigate the effect of temperature, high purity aluminum specimens were cold worked by pounding at −78°C (dry ice) and at −196°C (liquid nitrogen) and then allowing the specimen to warm up slowly to room temperature. It was thought that if the working was done at a temperature below that at which

[10] A. Guinier and J. Tennevin, C. R. Acad. of Sci., Paris *226*, 1530 (1948).
[11] R. D. Heidenreich, "Cold Working of Metals," page 57 (American Society for Metals, Cleveland, 1949.
[12] R. W. Cahn, *Proc. Phys. Soc. A 63*, 323 (1950).
[13] If this alloy is given a 10 min. anneal at 300°C, recovery domains very similar to Fig. 8 are obtained.
[14] D. Turnbull, *A.I.M.M.E. Gech. Pub.* #2365 (1948).

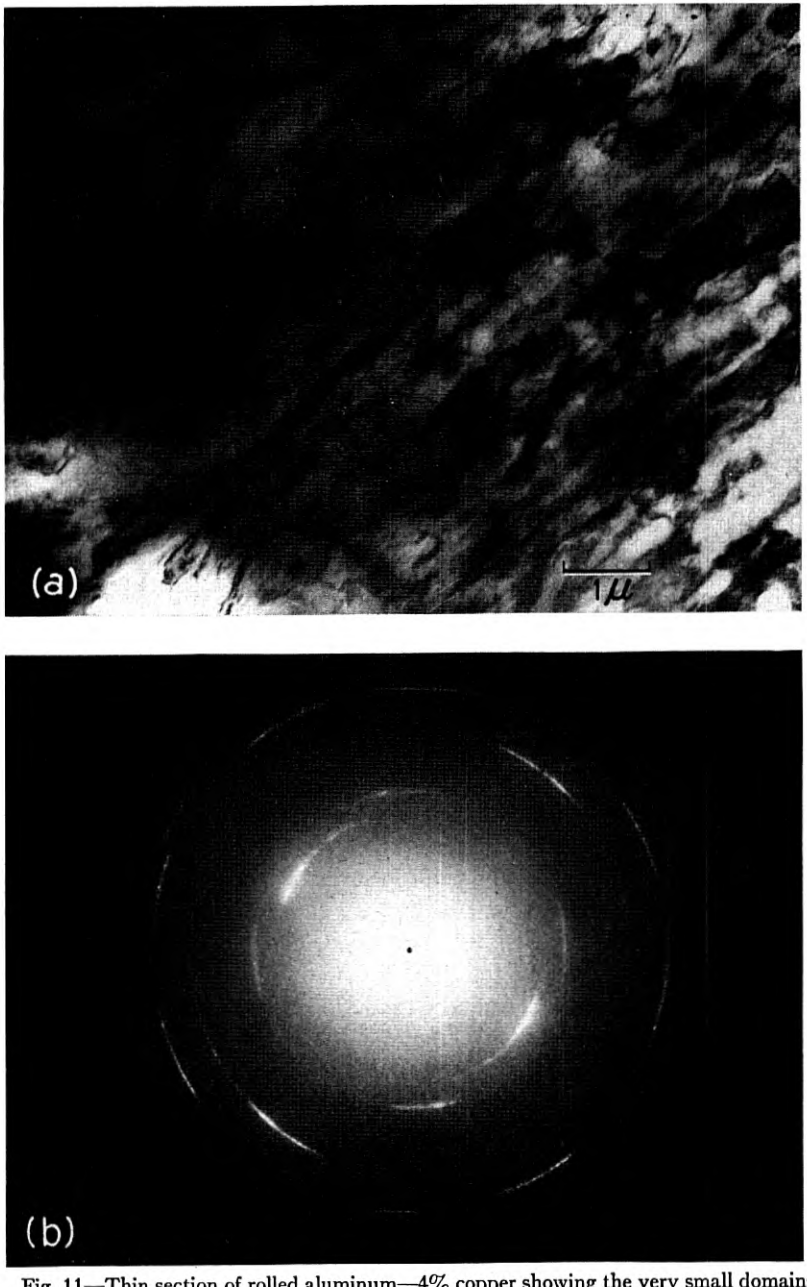



Fig. 11—Thin section of rolled aluminum—4% copper showing the very small domains. (b) is a transmission electron diffraction pattern. Compare with Fig. 8.

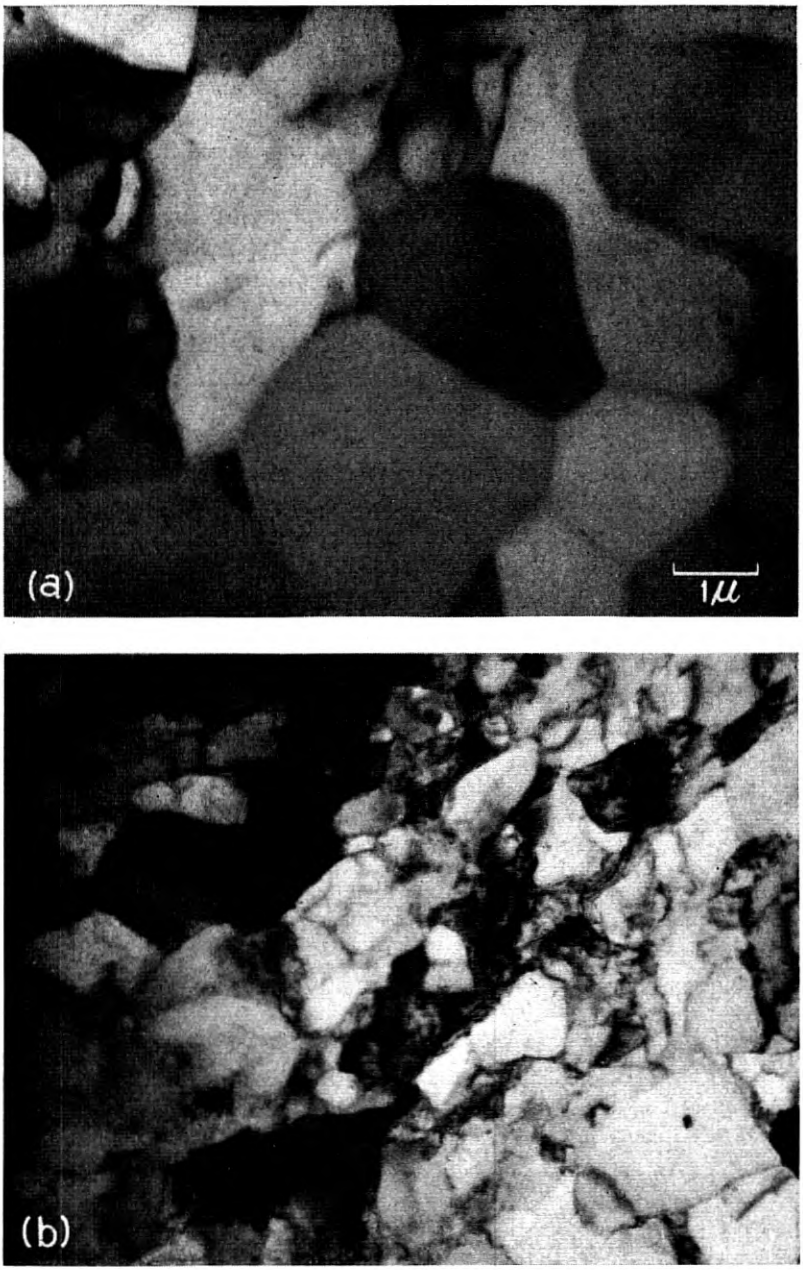



Fig. 12—Effect of temperature at which recovery proceeds on domain size in high purity aluminum.
(a). Dry ice (−78°C)
(b). Liquid nitrogen (−196°C)

the nucleation rate is a maximum, then as the specimen warmed slowly it would pass through the maximum and result in a larger number of nuclei. The time for recovery at the low temperature was varied from 15 minutes to several hours before bringing the specimen to room temperature but the results did not reflect any dependence on the time at low temperature. Even so, a primary weakness in the experiment lies in the fact that the structure could not be observed at the temperature of working. The results of cold working at −78°C and −196°C are shown in Fig. 12. Surprisingly enough, the domain size after cold working at −78°C is practically the same as at room temperature as seen in Fig. 12a. It was thought that this simply meant that the structure had reverted to the room temperature configuration until the results were obtained at −196°. The domain size resulting from the −196°C treatment is slightly less than half that obtained from the −76°C treatment, indicating that the effect of recovery temperature can be seen after bringing the specimen up to room temperature.[15] This is consistent with low temperature rolling experiments[16] on pure copper performed by W. C. Ellis and E. Greiner of Bell Telephone Laboratories in which the amount of work hardening was considerably increased over that obtained by rolling at room temperature. Thus, at present it appears that the recovery domains seen in Fig. 12 are a fair approximation to those produced at the low temperatures. More work on low temperatures is certainly justified since it appears that the recovery mechanism involves a process with a very low activation energy, at least in the case of pure aluminum.

## General Remarks

The conclusions drawn from the electron images of cold worked aluminum are, in review,

(1) During and immediately following cold working of pure aluminum self-recovery takes place by the formation of recovery domains about $2\mu$ in size.
(2) The recovery domains produced at room temperature and below at first possess sufficient internal strains to prohibit extinction contours. These strains are slowly relieved at room temperature.
(3) The addition of copper to the aluminum inhibits the growth of recovery domains resulting in a recovery domain size much smaller than for the pure metal. Thus, aluminum-copper work hardens to a far greater extent than does pure aluminum.
(4) Recovery in pure aluminum is reduced only by going to relatively

[15] The microbeam x-ray technique (reference 9) should be invaluable in checking this point since the entire experiment could be done at the low temperature.
[16] To be published.

low temperatures; i.e., of the order of that of liquid nitrogen (−196 °C).

(5) The recovery domains do not show (at least not readily) on etched surfaces. The overall rate of etching is much higher than in the annealed state, however.

(6) Deformation by simple extension does not produce the recovery domains such as seen in Fig. 8. Neither domains nor slip bands are visible.

These conclusions and observations suggest explanations for several well known phenomena in cold worked metals. One is the fact that slip lines are not visible on a surface etched *after* cold working, which has always been rather puzzling. It seems clear now that this is simply due to an immediate rearrangement giving rise to recovery so that the slip band exists as such *only during the actual deformation process*. The traces left on polished surfaces are actually only traces of the displacements that occurred during deformation and do not indicate where the energy of cold work resides when slipping has stopped but only where the energy was introduced. The energy would reside in the slip bands only if no recovery whatever took place.

Another point of interest is that of recrystallization. In one sense the recovery domains constitute recrystallization on a smaller scale than is usually meant. However, in view of the fact that the domains do not etch preferentially (probably due to internal strains) and that they disappear at the recrystallization temperature, it would seem more accurate to view the recovery domains as distinct from recrystallization. It would seem logical to consider the recovery domains as *embryos* for recrystallization. When the temperature is raised sufficiently those recovery domains which most rapidly relieve their internal strains would serve as nuclei for new grains and consume the surrounding embryos or domains. In a sense, then, it is the least strained material from which new grains spring. However, the embryo for the new grain very probably sprang from a region that was very highly strained. Between the actual slip process and final recrystallization grains there are actually two nucleation and growth processes.

The author is grateful to Mr. W. T. Read and Dr. W. Shockley for valuable criticisms and discussions of the subject matter presented in this paper.

# On the Reflection of Electrons by Metallic Crystals

By L. A. MacCOLL

This paper gives the results of some calculations of the reflection coefficient for electrons incident normally on a plane face of a metallic crystal. The physical situation is treated as being one-dimensional; and it is assumed that the potential energy of an electron is a sinusoidal function of distance inside the crystal, and obeys the classical image force law outside the crystal. The reflection coefficient is computed as a function of the energy of the incident electrons, over the range from 0 to 20 electron volts, for a variety of values of the parameters which define the model of the crystal.

## 1. Foreword

THE work which is presented in this paper was undertaken as a result of conversations had with Dr. C. J. Davisson at various times during the years 1938 and 1939, when he was investigating the reflection of electrons impinging on the surface of a metallic crystal. The results for a simple special case of the general problem were published in 1939[1]. Thereafter the work on the general problem continued intermittently, and it was almost completed by the early part of 1942, when it was brought to a halt by the onset of wartime activities. Since then nothing has been done on the problem, and the results already obtained have never been published *in extenso*. However, C. Herring and M. H. Nichols have included an illuminating discussion of some of the more significant of the results in their recent monograph on thermionic emission[2].

Although the intervening years, by bringing new problems in physics to the fore, have caused this work to lose some of the interest which it possessed at the time it was being done, it still seems to be worth while to put the full results upon record. The present occasion, when his friends and former colleagues are celebrating Dr. Davisson's seventieth birthday, is an especially appropriate one for this purpose.

## 2. Formulation of the Problem

We consider electrons moving with energy $E$ and impinging on a plane face of a metallic crystal. (Fig. 1.) According to quantum mechanics there is certain probability $R$, generally neither 0 nor 1, that an electron will be reflected by the crystal, and caused to move backwards toward the source; and there is the complementary probability $1 - R$ that the electron will

[1] *Physical Review*, v. 56, pp. 699–702. This paper will be referred to henceforth as [LAM, 1939].

[2] *Reviews of Modern Physics*, v. 21, pp. 185–270 (1949).

penetrate the crystal, and flow away through the remainder of the circuit. We call $R$ the *reflection coefficient*, and we can define it alternatively as the ratio of the intensity of the reflected electron beam to the intensity of the incident beam.

We wish to calculate $R$ as a function of $E$. In order to be able to do this effectively, it is necessary to idealize the actual physical situation quite drastically. (However, the idealization which we shall use preserves what seem to be the most important features of the physical situation.) On the other hand, once the idealization has been set up, the mathematical calculations themselves will be carried through without approximations[3]. Hence,

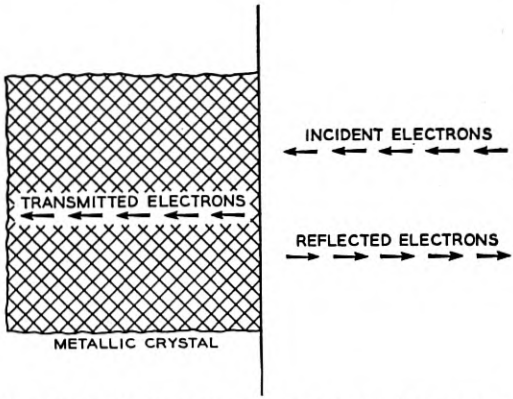



Fig. 1—Reflection of an electron beam by a crystal. (Schematic representation).

any discrepancies between the theoretical results and the results of experiment are to be attributed to the inadequacy of the model, and not to illegitimate steps in the mathematical work.

Our idealization of the physical situation can be described in the form of the three following assumptions:

*Assumption* I. The problem may be treated as one concerning one-dimensional motion of electrons. Thus, we set up a rectangular coordinate system in space; and we assume that the crystal occupies the half-space $x < 0$, and that all of the point functions with which we are concerned depend solely upon the coordinate $x$.

*Assumption* II. There exists a function $V(x)$, such that an electron at the point $x$ has potential energy $V(x)$; and the behavior of an electron is governed by the Schrödinger wave equation

[3] Except, of course, simple arithmetical approximations, such as are involved in almost all calculations.

$$\frac{d^2\psi}{dx^2} + k^2 [E - V(x)]\psi = 0. \tag{1}$$

(Here $k^2 = 8\pi^2 m/h^2$, where $h$ is Planck's constant, and $m$ is the mass of an electron.) This assumption deals in a summary way with various complicated processes involving electrons in crystals. Discussions of the validity of the assumption are to be found in various works on the electron theory of metals.

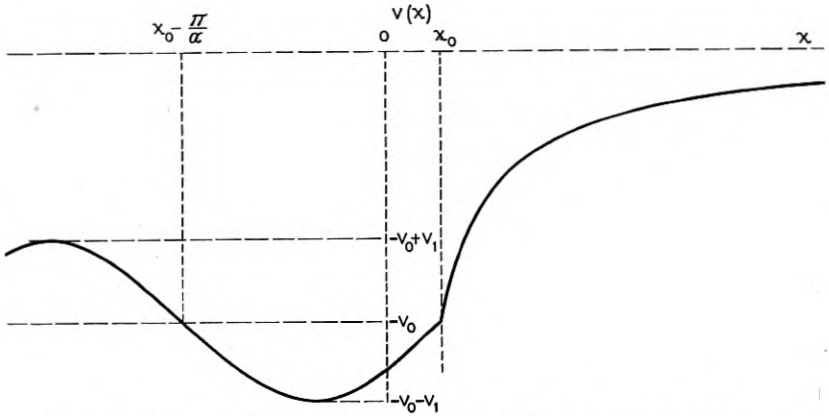



Fig. 2—Assumed potential energy as a function of the coordinate $x$.

*Assumption* III. Specifically, the function $V(x)$ is given by the formulae

$$\begin{aligned} V(x) &= -V_0 + V_1 \sin \alpha(x - x_0), & x \leq x_0 \\ &= -\epsilon^2/(4x), & x \geq x_0 \\ x_0 &= \epsilon^2/(4V_0), \end{aligned} \tag{2}$$

where $\epsilon$ is the absolute value of the electronic charge, and $V_0$, $V_1$ and $\alpha$ are suitable non-negative constants. (A graph of this function $V(x)$ is shown in Fig. 2.) According to this assumption, an electron in the region $x > x_0$ is subjected to the classical image force. This is known to be in good agreement with the facts, at least if $x$ is not too small.* Also, according to the assumption, the potential energy of an electron in the depths of the crystal is a periodic point function with a negative mean value. This part of the assumption is as correct as any assumption can be which attempts to account for the complicated actual processes in terms of a potential energy function. However, our particular choice of a periodic function is based largely upon mere considerations of mathematical convenience. Finally, we

* See Herring and Nichols, footnote 2, p. 245 et seq.

observe that our $V(x)$ is continuous, as physical considerations indicate that it should be.

We can now state the mathematical problem before us in the following terms:

$V(x)$ being defined by (2), we are to obtain a solution $\psi(x)$ of (1) satisfying the following conditions:

(a) In the region $x > x_0$ the function $\psi(x) \exp(-2\pi iEt/h)$ represents an incident beam of electrons moving toward the left, and a reflected beam of electrons moving toward the right.

(b) In the region $x < x_0$ the average electron flow, if it is not zero, is directed toward the left.

(c) The function $\psi(x)$ and its derivative $\psi'(x)$ are everywhere continuous.

Having obtained such a solution $\psi(x)$, we are then to compute the ratio of the intensity of the reflected electron beam to the intensity of the incident beam. In particular, we are to study the dependence of this ratio upon the quantities $E$, $V_0$, and $V_1$.

The paper [LAM, 1939] already referred to dealt with the special case in which $V_1 = 0$, i.e. the case in which $V(x)$ is assumed to be constant in the region $x \leq x_0$. Consequently, we are now concerned chiefly with the cases in which $V_1 > 0$.

## 3. Generalities Concerning the Calculation of $R$

In the region $x \geq x_0$ the wave equation (1) takes the form

$$\frac{d^2\psi}{dx^2} + k^2\left[E + \frac{\epsilon^2}{4x}\right]\psi = 0.$$

The general solution of this equation is of the form

$$\psi(x) = A\psi_1(x) + B\psi_2(x),$$

where $A$ and $B$ are arbitrary constants, and $\psi_1(x)$ and $\psi_2(x)$ are two particular solutions which we choose so that the functions $\psi_1(x) \exp(-2\pi iEt/h)$ and $\psi_2(x) \exp(-2\pi iEt/h)$ represent beams of electrons, of unit intensity, moving to the left and right, respectively.

In the region $x \leq x_0$ the wave equation takes the form

$$\frac{d^2\psi}{dx^2} + k^2[E + V_0 - V_1 \sin \alpha(x - x_0)]\psi = 0. \tag{3}$$

We are concerned with a solution of this equation of the form

$$\psi(x) = C\psi_3(x),$$

where $C$ is a constant, and $\psi_3(x)$ is a particular solution such that the func-

tion $\psi_3(x)$ exp $(-2\pi iEl/h)$ represents a state in which the average flow of electrons in the crystal either vanishes or is directed toward the left.

The actual forms of the functions $\psi_1(x)$, $\psi_2(x)$, $\psi_3(x)$ will be discussed presently.

Now the continuity of the functions $\psi(x)$ and $\psi'(x)$ gives us the system of equations

$$A\psi_1(x_0) + B\psi_2(x_0) = C\psi_3(x_0)$$
$$A\psi_1'(x_0) + B\psi_2'(x_0) = C\psi_3'(x_0),$$

from which we can calculate the ratio $B/A$ in terms of the $\psi_i(x_0)$, $\psi_i'(x_0)$. Our required reflection coefficient $R$ is $|B/A|^2$, and so we obtain the formula

$$R = \left| \frac{\psi_1'(x_0) - \dfrac{\psi_3'(x_0)}{\psi_3(x_0)}\psi_1(x_0)}{\psi_2'(x_0) - \dfrac{\psi_3'(x_0)}{\psi_3(x_0)}\psi_2(x_0)} \right|^2. \tag{4}$$

It was shown in [LAM, 1939] that the functions $\psi_1(x)$ and $\psi_2(x)$ are given by the formulae

$$\psi_1(x) = W_{\lambda,\frac{1}{4}}(\xi), \qquad \psi_2(x) = W_{-\lambda,\frac{1}{4}}(-\xi),$$

where

$$\xi = 2ikxE^{1/2}, \qquad \lambda = -ik\epsilon^2/(8E^{1/2}),$$

and the symbols $W_{\lambda,\frac{1}{4}}(\xi)$, $W_{-\lambda,\frac{1}{4}}(-\xi)$ denote the usual functions occurring in the theory of the confluent hypergeometric functions[4]. The earlier work gives us all the information concerning $\psi_1(x)$ and $\psi_2(x)$ that we shall require. Hence, in order to calculate $R$, we have, in effect, only to identify a suitable solution $\psi_3(x)$ of equation (3), and then to calculate $\psi_3'(x_0)/\psi_3(x_0)$.

## 4. The Solution of Equation (3)

In order to facilitate the use of known results, it is convenient to write

$$\alpha(x - x_0) = 2z - \frac{\pi}{2}, \qquad \frac{4k^2}{\alpha^2}(E + V_0) = \theta_0^2, \qquad \frac{4k^2}{\alpha^2}V_1 = -2\theta_1.$$

Then equation (3) takes the form

$$\frac{d^2\psi}{dz^2} + (\theta_0^2 + 2\theta_1 \cos 2z)\psi = 0. \tag{3'}$$

This is one of the canonical forms of Mathieu's differential equation, for

[4] E. T. Whittaker and G. N. Watson, "Modern Analysis" (Chapter XVI), Cambridge Univ. Press, 4th Ed., 1927.

which an extensive theory exists. We shall recall a few of the chief facts brought out in this theory.†

Unless the constants $\theta_0$ and $\theta_1$ satisfy some one of certain special relations, the general solution of equation (3′) is of the form

$$\psi = K_1 e^{\mu z} f(z) + K_2 e^{-\mu z} f(-z),$$

where $\mu$ is a constant determined by $\theta_0$ and $\theta_1$, $f(z)$ is a function which is periodic with the period $\pi$, and the $K$'s are constants of integration.

In certain ranges of values of the $\theta$'s, the constant $\mu$ is real, and in other ranges it is pure imaginary. When $\mu$ is real we can obviously take it to be positive; and then, in order that $\psi_3(x)$ may be bounded in the range $x < x_0$, we must choose $\psi_3(x)$ to be the function $e^{\mu z} f(z)$. When $\mu$ is pure imaginary, we can take it to be $i \mid \mu \mid$; and then, in order that $\psi_3(x)$ shall represent a state in which the flow of electrons is to the left in the crystal, we must choose $\psi_3(x)$ to be the function $e^{-\mu z} f(-z)$.

When $\mu$ is pure imaginary we have a non-vanishing flow of electrons to the left in the crystal. Consequently, the intensity of the reflected beam must be less than the intensity of the incident beam. Hence, under this condition we must have $R < 1$. On the other hand, when $\mu$ is real there is no average electron flow in the crystal. Consequently, under this condition the intensities of the incident and reflected beams must be equal, so that $R = 1$. These considerations point to the importance of discussing, first of all, the conditions under which $\mu$ is real or pure imaginary.

Figure 3 shows a well known diagram, modified slightly to suit our present purposes[5]. Here $\theta_0^2$ and $\theta_1$ are taken to be rectangular coordinates of a point in a plane, and the plane is divided into regions of two kinds (shaded and unshaded) by a system of curves. If the point $(\theta_0^2, \theta_1)$ is in the interior of one of the shaded regions, the above $\mu$ is real; if the point is in the interior of one of the unshaded regions, $\mu$ is pure imaginary. (If $(\theta_0^2, \theta_1)$ lies exactly on the boundary of one of the regions, we have a somewhat more complicated situation, which we do not need to consider here.) This diagram enables us easily to determine, for any fixed values of $V_0$ and $V_1$, the ranges of values of $E$ in which we have $R = 1$. We shall call these ranges of values of $E$ the *diffraction bands*.

Now our problem has been reduced to that of computing $R$ for values of $E$ which do not lie in diffraction bands. In treating this phase of the subject we shall follow the course of the actual calculations, without any examination of ways in which the work might have been done more efficiently.

† See, for instance, E. T. Whittaker and G. N. Watson, footnote 4, Chapter XIX.

[5] See, for instance, N. W. McLachlan, "Theory and Application of Mathieu Functions" (p. 40), Oxford University Press, 1947.

Of the many methods which have been devised for finding solutions of Mathieu's differential equation, the one which is conceptually simplest is that due to Bruns. This method can be described as follows:

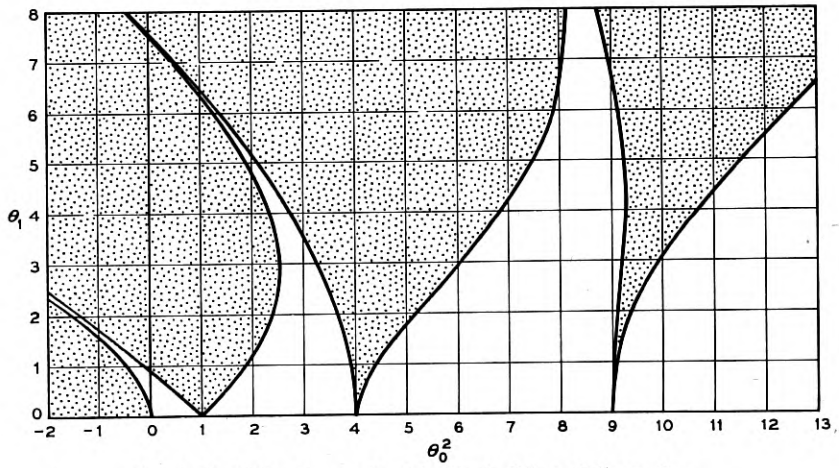



Fig. 3—Stability diagram for Mathieu's differential equation.

Under the transformation

$$\psi = \exp \int_{z_0}^{z} \varphi(z)dz$$

the Mathieu equation (3′) goes over into the Riccati equation

$$\frac{d\varphi}{dz} + \varphi^2 + \theta_0^2 + 2\theta_1 \cos 2z = 0. \tag{5}$$

We seek a solution of this equation in the form of a power series in the parameter $\theta_1$, say

$$\varphi(z) = \varphi_0(z) + \theta_1\varphi_1(z) + \theta_1^2\,\varphi_2(z) + \cdots,$$

and we easily find that the functions $\varphi_0(z)$, $\varphi_1(z)$, $\varphi_2(z)$, ... must satisfy the differential equations

$$\begin{aligned}
&\varphi_0' + \varphi_0^2 + \theta_0^2 = 0,\\
&\varphi_1' + 2\varphi_0\varphi_1 + 2\cos 2z = 0,\\
&\varphi_2' + 2\varphi_0\varphi_2 + \varphi_1^2 = 0,\\
&\varphi_3' + 2\varphi_0\varphi_3 + 2\varphi_1\varphi_2 = 0,\\
&\varphi_4' + 2\varphi_0\varphi_4 + 2\varphi_1\varphi_3 + \varphi_2^2 = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{6}$$

Obviously, in order that we shall get the solution which we require, $\varphi(z)$ must reduce to $-i\theta_0$ when $V_1 = 0$. Also $\varphi(z)$ must be periodic with the period $\pi$. These conditions, together with the equations (6), suffice to determine $\varphi_0(z)$, $\varphi_1(z)$, $\varphi_2(z)$, $\cdots$ successively and uniquely. We easily obtain the results:

$$\varphi_0(z) = -i\theta_0,$$

$$\varphi_1(z) = -\frac{1}{2i}\left[\frac{e^{2iz}}{1-\theta_0} - \frac{e^{-2iz}}{1+\theta_0}\right],$$

$$\varphi_2(z) = \frac{1}{8i}\left[\frac{e^{4iz}}{(1-\theta_0)^2(2-\theta_0)} + \frac{2}{(1-\theta_0^2)\theta_0} - \frac{e^{-4iz}}{(1+\theta_0)^2(2+\theta_0)}\right]$$

$$\varphi_3(z) = -\frac{1}{16i}\left[\frac{e^{6iz}}{(1-\theta_0)^3(2-\theta_0)(3-\theta_0)} + \frac{(4-3\theta_0)e^{2iz}}{(1-\theta_0)^3(1+\theta_0)\theta_0(2-\theta_0)}\right.$$
$$\left. + \frac{(4+3\theta_0)e^{-2iz}}{(1+\theta_0)^3(1-\theta_0)\theta_0(2+\theta_0)} - \frac{e^{-6iz}}{(1+\theta_0)^3(2+\theta_0)(3+\theta_0)}\right],$$

$$\varphi_4(z) = \frac{1}{128i}\left[\frac{(11-5\theta_0)e^{8iz}}{(1-\theta_0)^4(2-\theta_0)^2(3-\theta_0)(4-\theta_0)}\right.$$
$$+ \frac{4(15-18\theta_0+5\theta_0^2)e^{4iz}}{(1-\theta_0)^4(1+\theta_0)(2-\theta_0)^2(3-\theta_0)\theta_0} - \frac{2(8-35\theta_0^2+15\theta_0^4)}{(1-\theta_0^2)^3(4-\theta_0^2)\theta_0^3}$$
$$+ \frac{4(15+18\theta_0+5\theta_0^2)e^{-4iz}}{(1+\theta_0)^4(1-\theta_9)(2+\theta_0)^2(3+\theta_0)\theta_0}$$
$$\left. - \frac{(11+5\theta_0)e^{-8iz}}{(1+\theta_0)^4(2+\theta_0)^2(3+\theta_0)(4+\theta_0)}\right].$$

The functions $\varphi_5(z)$ and $\varphi_6(z)$ have been computed also; but, because of their complexity, they will not be exhibited here.

It is easily found that the expression $\psi_3'(x_0)/\psi_3(x_0)$ appearing in equation (4) has the value $\varphi(\pi/4)(\alpha/2)$ in terms of our present notation[6].

In principle we now have all the information we need to calculate the reflection coefficient $R$. Furthermore, our experience showed that it is quite easy to compute $R$ by this method, provided that the value of $E$ does not lie too near an edge of a diffraction band. However, Brun's method proved to be unsuitable for calculating values of $R$ for values of $E$ in the neighborhood of a diffraction band edge, and we were forced to seek another method to obtain these values. After some tentative work with other methods, we

[6] We must take account of the fact that the symbols $\psi(x)$ and $\psi(z)$, which we are using for convenience, do not represent merely the same function with different arguments. In fact, after the change of the independent variable from $x$ to $z$ we have the following relation between the old and new $\psi$'s: $[\psi(x)]_{\text{old}} = [\psi(x_0 + 2z/\alpha - \pi/2\alpha)]_{\text{old}} = [\psi(z)]_{\text{new}}$.

settled upon a method due to Whittaker, and this was used to complete the calculations.

Whittaker's method is described briefly in Whittaker and Watson's "Modern Analysis," 4th Edition, p. 424. The method is developed more fully in papers by Ince[7]. We shall confine ourselves here to some summary indications of the nature of the method.

The method leads to representations of the solutions of Mathieu's equation by formulae which differ in structure depending upon the part of the $(\theta_0^2, \theta_1)$ plane in which we are working. We shall give the formulae suitable for use in the neighborhood of the point $\theta_0^2 = 1, \theta_1 = 0$; the formulae for use in other parts of the plane are given by Ince.

Given the values of $\theta_0^2$ and $\theta_1$, we first determine a number $\sigma$ by means of the implicit equation

$$\theta_0^2 = 1 + \theta_1 \cos 2\sigma + \frac{\theta_1^2}{8}(-2 + \cos 4\sigma) - \frac{\theta_1^3}{64}\cos 2\sigma + \frac{\theta_1^4}{512}\left(\frac{32}{3} - 11\cos 4\sigma\right) + \cdots.$$

Then we seek a solution of equation (3′) in the form

$$\psi = e^{\mu z}\sum_{n=0}^{\infty}\{a_{2n+1}\cos[(2n+1)z - \sigma] + b_{2n+1}\sin[(2n+1)z - \sigma]\},$$

where $\mu$, the $a$'s, and the $b$'s are constants. We substitute the expression for $\psi$ into the differential equation, and determine values of the constants by imposing the condition that the resulting relation shall be an identity in $z$. After some rather intricate algebraic manipulations we finally arrive at the following results:

$$\mu = \frac{\theta_1}{2}\sin 2\sigma - \frac{3\theta_1^3}{128}\sin 2\sigma - \frac{3\theta_1^4}{1024}\sin 4\sigma + \cdots$$

$$a_1 = 0, \qquad b_1 = 1$$

$$a_3 = \frac{3\theta_1^2}{64}\sin 2\sigma + \frac{3\theta_1^3}{512}\sin 4\sigma + \frac{\theta_1^4}{8^4}\left(-\frac{274}{9}\sin 2\sigma + 9\sin 6\sigma\right) + \cdots$$

$$a_5 = \frac{14\theta_1^3}{(8^3)(9)}\sin 2\sigma + \frac{44\theta_1^4}{(27)(8^4)}\sin 4\sigma + \cdots$$

$$a_7 = \frac{35\theta_1^4}{(108)(8^4)}\sin 2\sigma + \cdots$$

[7] *Monthly Notices, Royal Astronomical Society of London*: v. 75, pp. 436–448; v. 76, pp. 431–442.

$$a_9 = O(\theta_1^5)$$

$$b_3 = \frac{\theta_1}{8} + \frac{\theta_1^2}{64}\cos 2\sigma + \frac{\theta_1^3}{8^3}\left(-\frac{14}{3} + 5\cos 4\sigma\right) + \frac{\theta_1^4}{8^4}\left(-\frac{74}{9}\cos 2\sigma + 7\cos 6\sigma\right) + \cdots$$

$$b_5 = \frac{\theta_1^2}{192} + \frac{4\theta_1^3}{(9)(8^3)}\cos 2\sigma + \frac{\theta_1^4}{(3)(8^4)}\left(\frac{82}{9}\cos 4\sigma - \frac{155}{18}\right) + \cdots$$

$$b_7 = \frac{\theta_1^3}{(18)(8^3)} + \frac{\theta_1^4}{(12)(8^4)}\cos 2\sigma + \cdots$$

$$b_9 = \frac{\theta_1^4}{(180)(8^4)} + \cdots$$

. . . . . .

The calculated terms which are exhibited here enable us to calculate the solution $\psi(z)$ to a certain accuracy, and this accuracy proved to be sufficient for our purposes.

Although this method is very complicated analytically, it was found to be quite convenient for purposes of numerical calculation.

## 5. The Reflection Coefficient for Large Values of $E$

This work is concerned chiefly with the reflection coefficient for small values of $E$ (actually up to 20 electron volts). However, it is interesting that we can obtain a simple approximate formula for $R$ for indefinitely large values of $E$ in the intervals between the diffraction bands.

For this purpose we go back to Brun's method, and we write the dependent variable in equation (5) in the form $\varphi = -i\theta_0 + \omega$. We find that the new dependent variable $\omega$ satisfies the equation

$$\frac{d\omega}{dz} - 2i\theta_0\omega + \omega^2 + 2\theta_1\cos 2z = 0,$$

and we seek a solution of this equation in the form

$$\omega = \omega_0(z) + \frac{\omega_1(z)}{\theta_0} + \frac{\omega_2(z)}{\theta_0^2} + \cdots.$$

The functions $\omega_n(z)$ are easily computed, and we finally arrive, in an entirely straight-forward way, at the result

$$\varphi(\pi/4) = -i\theta_0 + \frac{\theta_1}{\theta_0^2} + \frac{\theta_1}{\theta_0^4} + \cdots. \tag{7}$$

In [LAM, 1939] we derived an approximate formula for $R$ when $E$ is large for the case in which $V_1 = 0$. The work involved in that derivation, together with the equation (7), enables us to obtain the desired formula by a simple calculation. The result obtained is the following:

$$R \doteq \frac{V_0^4}{4\epsilon^4 k^2 E^3}\left[1 - \frac{\alpha\epsilon^2 V_1}{4V_0^2}\right]^2. \tag{8}$$

The range of validity of the approximate formula (8) has not been determined. The nature of the derivation, and also the form of the result, leads us to suspect that the approximation is good only so long as the ratio $V_1/V_0$ does not exceed some bound depending upon the other quantities entering into the expression for $R$. The approximation certainly breaks down when $V_1/V_0$ reaches the value $4V_0/(\alpha\epsilon^2)$. However, this value is well above any of the values with which we deal with in this work. Consequently, we suspect that the formula can be used, to extrapolate our calculations of $R$ to higher values of $E$, without serious danger of error in the cases which we consider here.

## 6. The Calculated Results

The reflection coefficient depends upon the independent variable $E$, and upon the three parameters $V_0$, $V_1$, and $\alpha$. The effects upon $R$ of taking various values of $V_0$ and $V_1$ seemed to be of greater interest than the effect of taking various values of $\alpha$; and, consequently, we confined ourselves in the calculations to a single value of $\alpha$, namely, $\alpha = \pi \times 10^{-8}$ cm$^{-1}$. This value of $\alpha$ makes the period of $V(x)$ in the crystal equal to $2 \times 10^{-8}$ cm.

We took six values of $V_0$, proceeding in equal steps from 10 electron volts to 20 electron volts inclusive. These values adequately cover the range which is of interest in connection with actual metals.

Including the calculations reported in [LAM, 1939], we have taken, for each of the values of $V_0$, five values of $V_1$, proceeding in equal steps from 0 to 0.4 $V_0$. Although it is somewhat difficult to say just what value of $V_1$ is most appropriate to the case of a specific actual metal, it appears that these values cover the range of values of interest adequately.

The results of the calculations are shown in a self-explanatory form by the curves given in Figs. 4 to 9 inclusive. For the sake of unity, we have included the results which were previously published in [LAM, 1939].

## 7. Physical Discussion of the Results

The results do not call for much discussion, especially in view of the discussion which Herring and Nichols have given in the paper already referred to. However, there are a few observations which should be made.

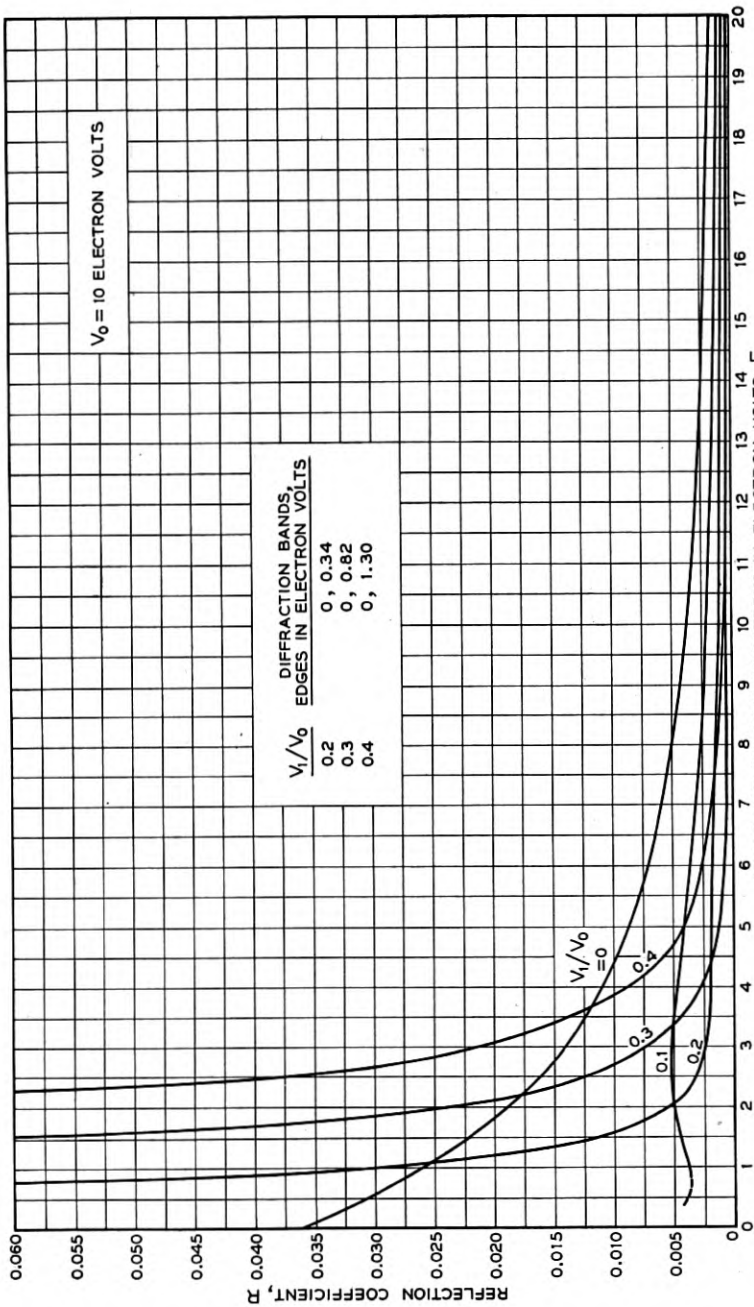



Fig. 4—Reflection coefficient as a function of energy. $V_0$ = 10 electron volts.

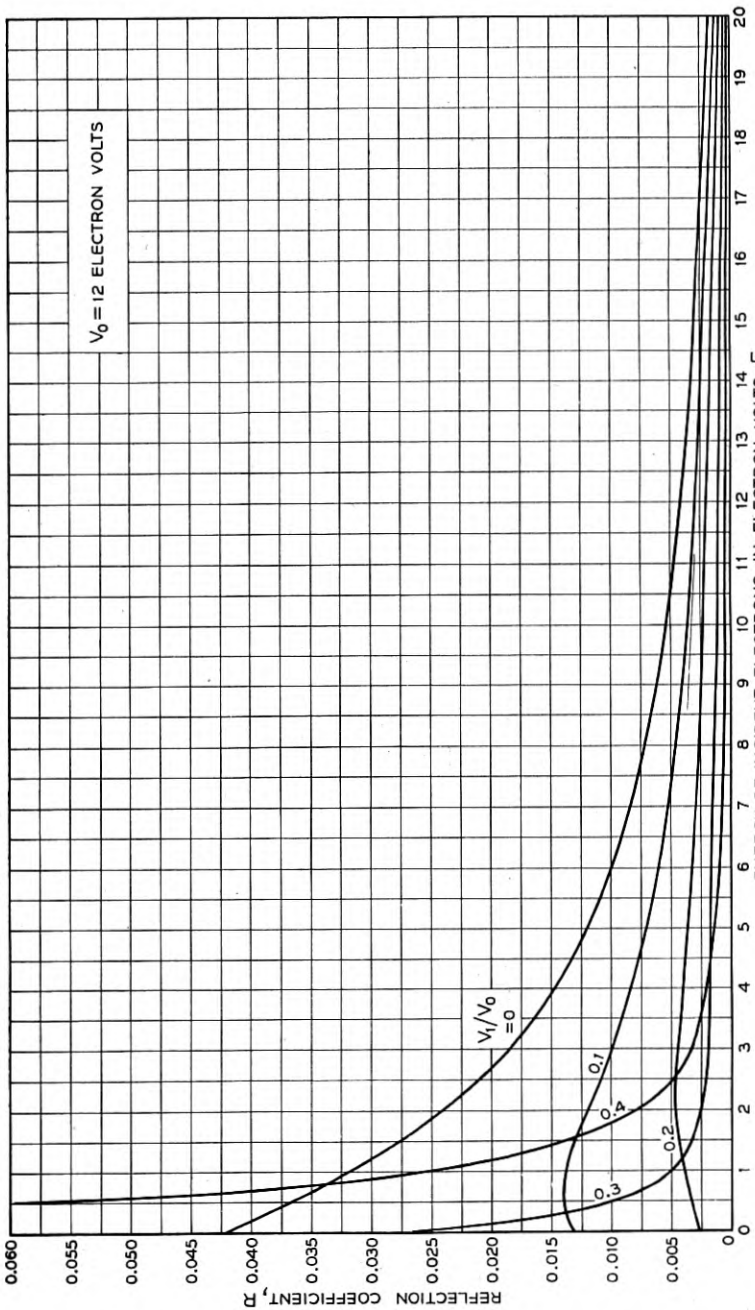



Fig. 5—Reflection coefficient as a function of energy. $V_0 = 12$ electron volts.

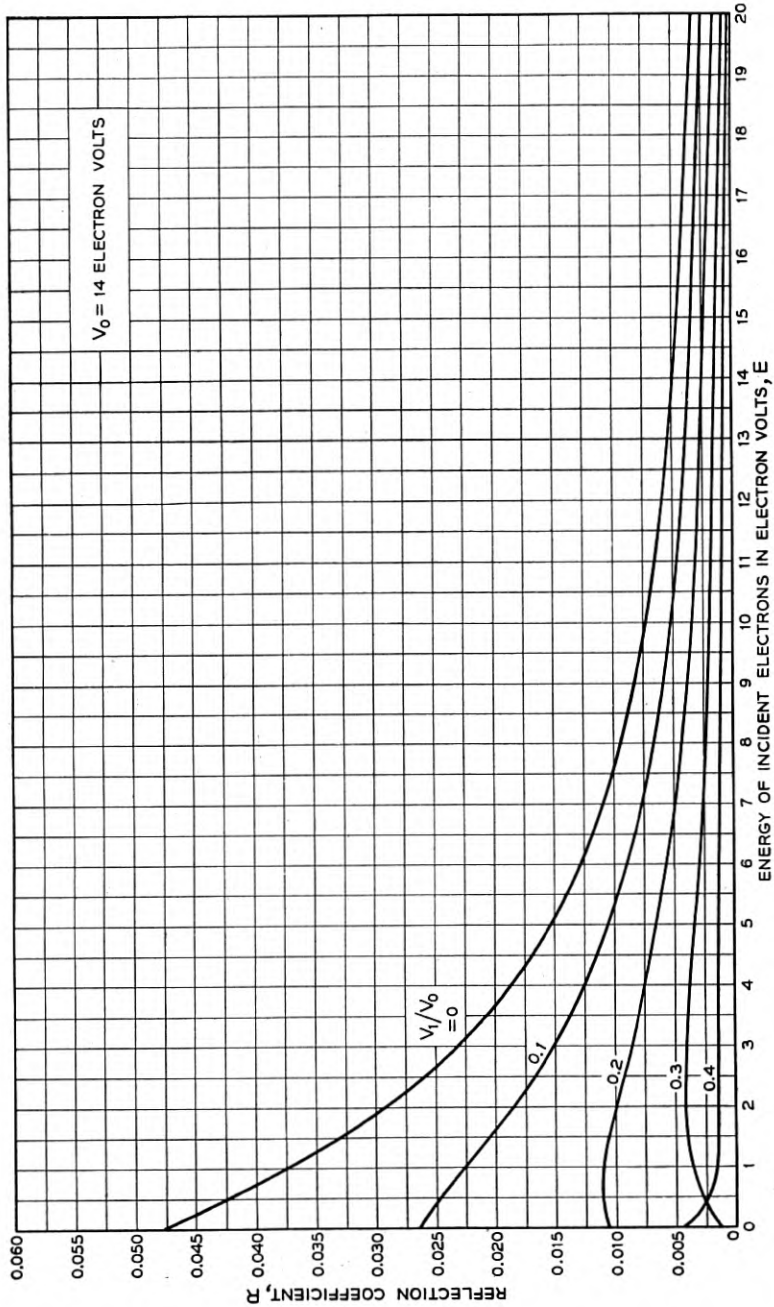



Fig. 6—Reflection coefficient as a function of energy. $V_0$ = 14 electron volts.

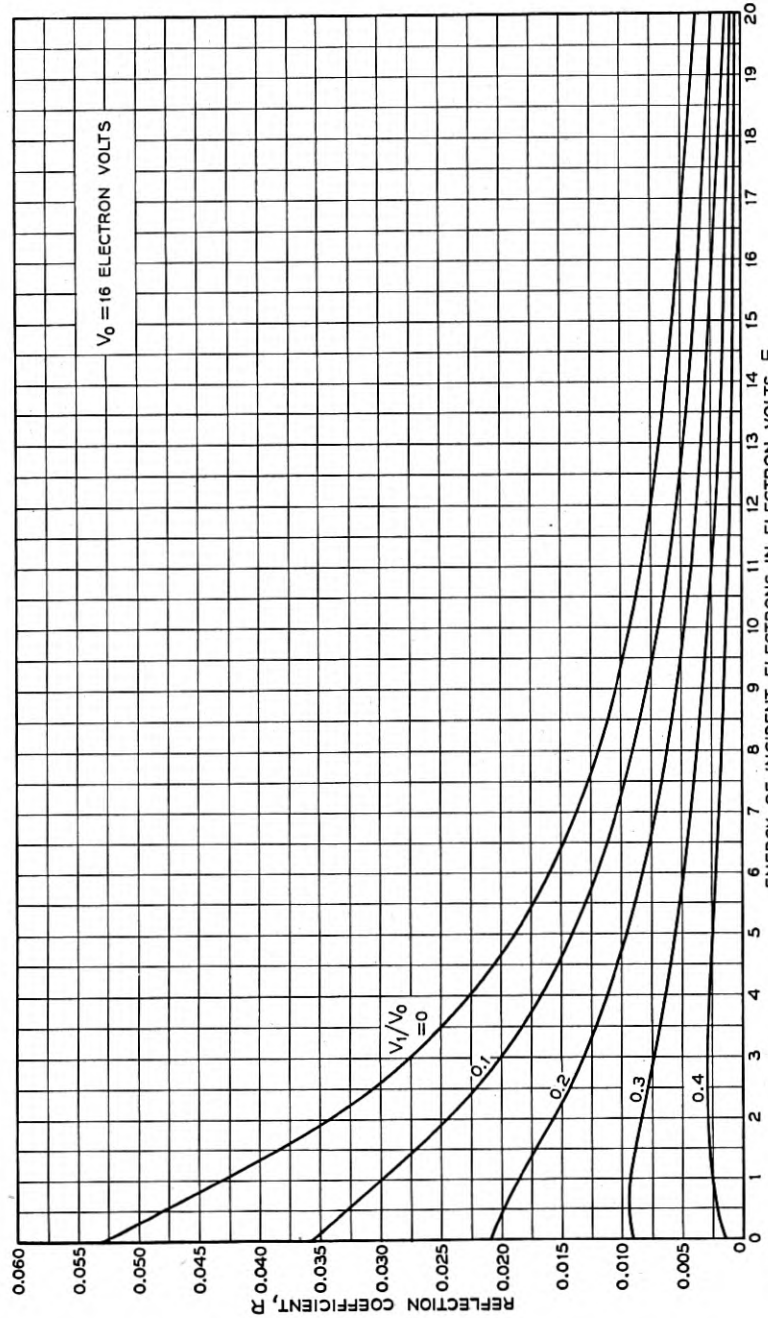



Fig. 7—Reflection coefficient as a function of energy. $V_0$ = 16 electron volts.

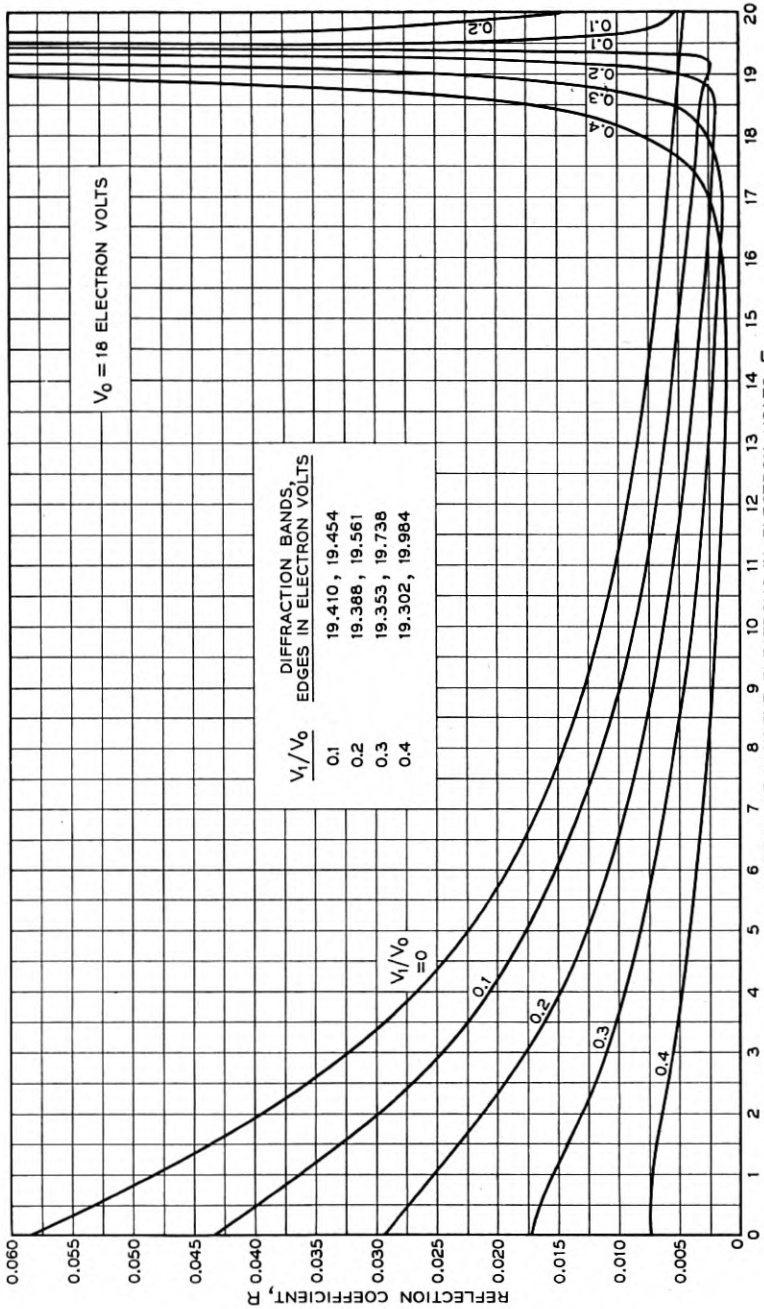



Fig. 8—Reflection coefficient as a function of energy. $V_0 = 18$ electron volts.

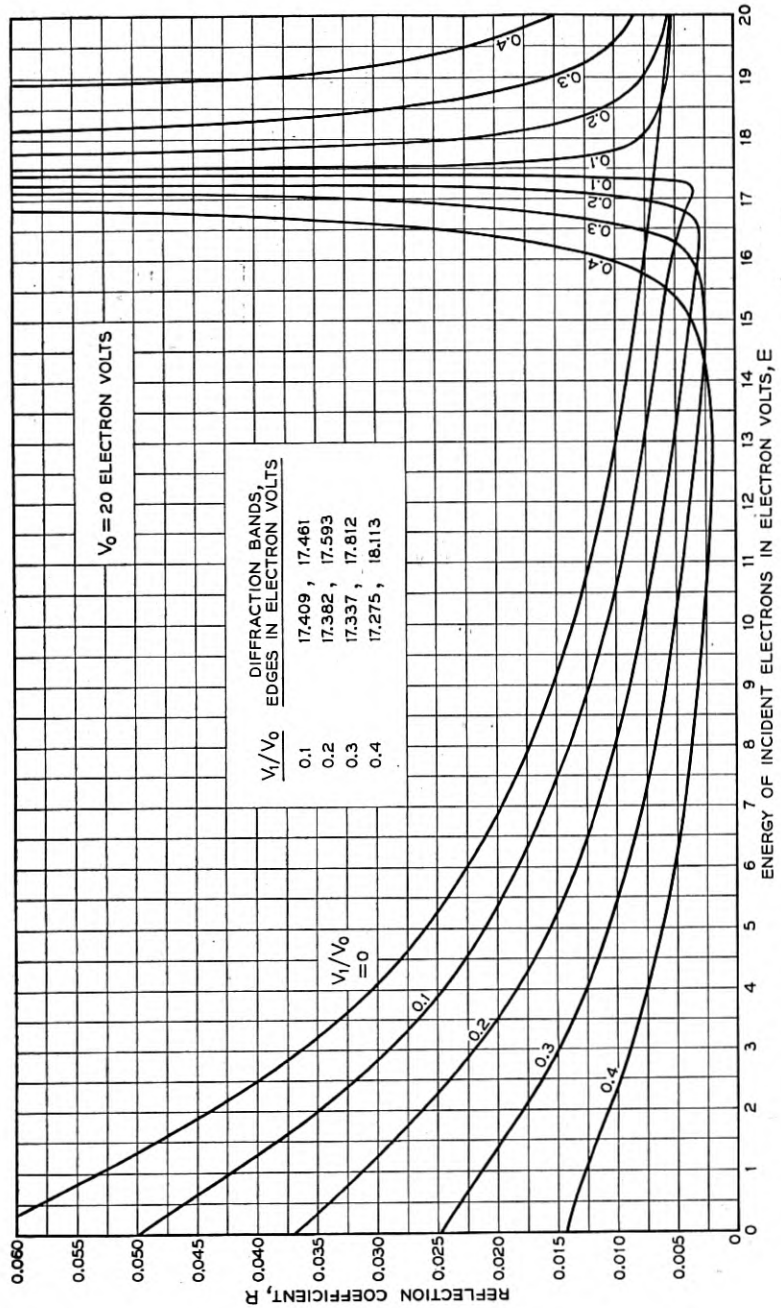



Fig. 9—Reflection coefficient as a function of energy. $V_0 = 20$ electron volts.

For fixed values of $V_0$ and $V_1$, the reflection coefficient tends to decrease with increasing $E$ over most of the range between any two successive diffraction bands. For fixed values of $V_0$ and $E$, $E$ being in a range between two diffraction bands, $R$ tends to decrease with increasing $V_1$. This is a result which was not anticipated when the work was begun. As was expected, we find a tendency for $R$ to increase with $V_0$ when $E$ and $V_1/V_0$ are held fixed.

The most interesting feature of the results is the behavior of $R$ in the neighborhoods of the edges of the diffraction bands. Unfortunately, the range of values of $E$ considered is not great enough to reveal this behavior very completely. (The failure to consider a greater range of values of $E$ was the result of our reluctance to embark again on the difficult numerical computations relating to the $W_{\pm\lambda,\frac{1}{3}}(\pm\xi)$ functions. Since the necessary computations had been performed earlier for values of $E$ up to 20 electron volts, we decided, unfortunately as it now appears, to confine ourselves to this range.)

The behavior of the curves near the edges of the diffraction bands which occur in the neighborhood of $E = 19$ electron volts (when $V_0$ is 18 or 20 electron volts) does not require much discussion. The reader will observe a small dip in the curves for $V_1 = V_0/10$ just below these diffraction bands. The accuracy of the computations is believed to be high enough that we are justified in taking this dip to be entirely genuine.

When $V_0$ is 10, 12, or 14 electron volts the behavior of the curves for values of $E$ in the neighborhood of zero is rather complicated. Herring and Nichols consider this behavior to be one of the most significant features of the results, and they have given a full discussion of it from the physical point of view.* In view of the availability of their discussion, we may confine ourselves here to a few brief remarks.

In some of the cases which we are referring to now there are diffraction bands extending from $E = 0$ to certain positive values of $E$. (These diffraction bands are shown in a self-explanatory way in the figures.) When such a diffraction band exists the complicated behavior of $R$ for small values of $E$ is a result of the existence of the diffraction band, and it is comparable with the behavior of $R$ in the upper part of the range of values of $E$ in the case in which $V_0 = 20$ electron volts.

In the cases in which we do not have diffraction bands extending upward from $E = 0$ we have to explain the complicated behavior of $R$ in somewhat different terms.

Assuming that we have such a case, let us momentarily ignore the fact that the physically significant values of $E$ are non-negative, and consider $E$ as an unrestricted real parameter. Under this convention concerning $E$,

* Herring and Nichols, footnote 2, pp. 248–249.

we find that there is a diffraction band lying in the range of negative values of $E$, and extending more or less near to the point $E = 0$[8]. We shall call this a *fictitious diffraction band*. Now it is immediately clear that the complicated behavior of our curves arises from the fact that the small positive (and hence physically significant) values of $E$ concerned are near the upper edge of the fictitious diffraction band.

[8] Specifically, what is meant is that there is such a range of values of $E$ in which the exponent $\mu$ is real.

# The Use of the Field Emission Electron Microscope in Adsorption Studies of W on W and Ba on W

By J. A. BECKER

The chief conclusions from these studies are given in the Introduction. Table II summarizes the adsorption properties of *W* on *W* and *Ba* on *W*. These properties vary with the crystal plane and are given for five planes. The extent to which these planes develop depends on *T* and the applied field, *F*. The temperature at which *W* atoms migrate on *W* at detectable rates depends on the plane and on *F*, and varies from 800 to 1200°K.

The adsorption properties of *Ba* on *W* are quite different for the first layer than they are for subsequent layers. In the first layer for which $\theta \leq 1$, *Ba* forms two phases: a condensed phase in which the *Ba* forms clusters or islands having a median diameter of $100 \times 10^{-8}$ cm, and a dispersed phase consisting of individual atoms. The temperature at which *Ba* migrates at detectable rates varies from 370 to 800°K from the 110 to the 100 plane. The evaporation rate depends on $\theta$. At $\theta$ near 1.0 it is detectable at 1050°K. At 1600°K practically all the *Ba* is evaporated.

For more than one layer of *Ba* on *W*, the *Ba* forms crystallites which grow outward from the *W* surface even at room *T*. Their median diameter is about $400 \times 10^{-8}$ cm and they disappear between 600 and 800°K.

## Introduction and Conclusions

E. W. MÜLLER,[1] in 1936, described a tube in which the field emission electrons from a very sharp tungsten point were made to impinge on a fluorescent screen and there portray a magnified image of the variation in emission density from different regions on the point. He showed that magnifications approaching a million fold could be obtained. In subsequent papers[2] he showed how such a tube can yield direct and striking information on the surface structure and on the effects of adsorbed films. Jenkins,[3] in 1943, summarized the progress to that date and showed that fields of the order of $10^7$ volts/cm produced pronounced changes in the surface configuration. More recently F. Ashworth[4] has reviewed the field emission from clean metallic surfaces.

In Fig. 2, (a) and (b) are two examples of photographs of the screen when field emission electrons are drawn from a single crystal of tungsten. The bright and dark regions are caused by variations in the intensity of electron emission from different regions of the tungsten surface. From such photographs it is possible to deduce how the electron work function varies for different crystallographic planes, how adsorbed atoms change this work function, and how the surface deviates from a smooth hemisphere when the tungsten is subjected to a range of temperatures and fields.

It is quite apparent that this new and powerful tool will reveal, on an

almost atomic scale, the nature of adsorption phenomena which are basic to thermionic, photoelectric and secondary electron emission, to catalysis and also to biological processes. Unfortunately, in most of the early work the residual gas pressure in the tube was such that the surface was contaminated in a few minutes and hence the results were probably affected by this contamination. In the present investigation the vacuum conditions were improved to such an extent that the residual gas produced only barely detectable effects after about one week. Under such conditions the following observations and conclusions have been made:

(1) When a sharp point of a single crystal of tungsten is held at 2400°K until a steady state is reached, most of the surface is approximately hemispherical or more precisely paraboloidal. About 20% of the surface consists of three atomic planes which in decreasing order of size are 110, 211, and 100 planes. This is also the order of increasing intensity of field emission. The greatest intensity of emission is from the 611 direction and other directions surrounding the 100 direction. The next greatest intensity comes from the 111 direction and neighboring regions.

(2) For temperatures $>$ 2400°K the area of the 110, 211, and 100 planes decreases; between 2400 and 1050°K the 211 and 100 planes increase steadily in size; below 1050°K the rate of change of area is so slow that no changes are observed in one hour. These changes are ascribed to migration of $W$ atoms on $W$.

(3) From 1050 to 1200°K, $W$ atoms migrate most easily in the 111 direction on the 211 plane. In this direction the atoms in the outermost layer contact their nearest neighbors but the rows of atoms are separated by 1.635 atom diameters. The migrated $W$ atoms are deposited on the hemispherical surface adjoining the 211 plane and form a crescent shaped mound resulting in an abnormally high field and enhanced emission. In the region between the 211 and 110 planes, $W$ atoms are also mobile in the 111 direction and form a series of step-like planes. In other regions the $W$ atoms show no large scale migration. Above 1200°K, $W$ atoms are mobile everywhere.

(4) When fields of the order of 40 million volts/cm are applied to the surface the rate of change of the surface configuration is greatly increased and migration of $W$ atoms can be observed in one hour on the 211 planes and near-by regions at 800°K. These changes are the same for electron accelerating and electron retarding fields. The rate of change increases rapidly with the strength of the field, perhaps as the square or cube of the field. At $T =$ 1400°K and for fields of $40 \times 10^6$ volts/cm applied for hours, over half of the surface consists of planes: the 211 planes almost meet the 110 planes, and 111 and 310 planes develop. Subsequent glowing without an applied field undoes the effects produced by the field.

(5) When Ba is deposited on clean $W$, the average work function $\varphi$ decreases from about 4.4 volts to about 2.1 volts when an optimum amount is reached at somewhere near a monomolecular layer. Further deposition increases $\varphi$ to that of bulk Ba for which $\varphi$ is 2.5 volts. For convenience we define the average coverage, $\theta$, as the Ba concentration divided by the concentration when $\varphi$ is a minimum.

(6) For $\theta$ from 0 to 1.0, the emission comes largely from aggregates or clusters of Ba, approximately circular in shape with diameters ranging from 40 to 200 Å and a median diameter of about 100 Å. Between 600 and 900°K these clusters are in violent agitation with the centers of a cluster appearing to shift about half a diameter. Sometimes one cluster may disappear and another one near by appear. We propose that this means that the Ba forms two phases on the tungsten surface: a condensed phase of clusters and a gaseous phase of individual Ba atoms. We propose that the centers of these clusters are irregularities on the tungsten surface where small atomic planes or facets meet to form a valley. Even a clean tungsten surface shows evidence of such irregularities whose distribution in numbers and sizes is about the same as for Ba on $W$ but in which the variation in emission density is much less pronounced.

(7) For $\theta > 1.0$, the emission comes mostly from larger aggregates which range in size from 200 to 600 Å or more. They produce spots which are intensely bright and are in continuous agitation of flicker even at room temperature. We associate these larger bright spots with crystallites because we believe them to be caused by Ba crystals which grow out normal to the tungsten surface and thus produce extra large local fields and hence enhanced local emission. These crystallites disappear, presumably due to migration or evaporation, at temperatures from 400 to 600°K.

(8) For $\theta < 1.0$ and $T$ between 600 and 1000°K, the chief effects are due to migration of Ba from one region to another. From 600 to 700°K this migration is restricted to the 211 planes and adjoining regions in the 111 zones; the regions near 100 do not yet show migration. In any region migration starts when the Ba clusters show noticeable agitation. At 800°K cluster agitation and migration occur in all regions and Ba atoms migrate from one side of the point to the other side for a distance of 3000 Å in about 5 minutes. At 900°K the migration rate is more rapid.

(9) For $\theta < 1.0$ and $T$ between 1050 and 1600°K the chief effect is that of evaporation. This is deduced from the fact that, as the temperature is increased progressively in about 100° steps and maintained at each $T$ for about 5 min., the voltage or field required to obtain an emission of say 10 microamps becomes progressively higher, presumably because $\theta$ decreases and $\varphi$ increases. At any one temperature, the rate of evaporation is at first quite rapid but decreases as $\theta$ decreases. After five minutes the rate is much

less than it was in the first minute, and after about 20 minutes the rate of evaporation is so small that $\theta$ has nearly reached a steady state. To reduce $\theta$ still more it is necessary to increase $T$. At 1600°K nearly all the Ba has evaporated and the emission pattern looks like that of clean $W$. Subsequent glowing at still higher temperatures produces only small increases in $\varphi$ and small changes in the emission pattern.

(10) For $\theta$ = about .18 and $T$ = 800°K we have observed a marked redistribution of Ba when a high field is applied: some Ba leaves the 211 and 111 regions and accumulates near the 100 regions. If the surface is then held at 800°K with zero field, some Ba leaves the 100 region and returns to

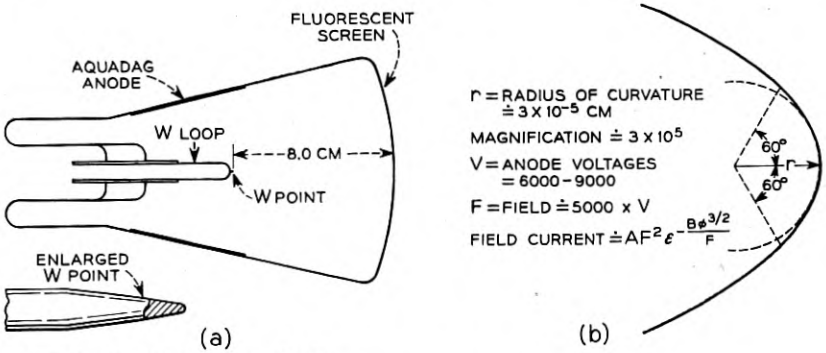



Fig. 1—(a). Cross-section of tube.
(b). Enlarged cross-section of tip of $W$ point and values of constants.

the 211 and 111 regions. This indicates that the adsorption forces can be modified appreciably by high applied fields.

## PART I

### Description of Field Emission Microscope Tube

Figure 1(a) is a cross-section of the tube. The loop which was used to heat the point consisted of $W$ wire 5.7 cm long and .0165 cm diameter. The $W$ point projected about 1 mm beyond the loop. It was formed by repeatedly heating the $W$ wire in a gas flame to oxidize it and removing the oxide with sodium nitrite. Two tantalum wire loops of $2.0 \times 10^{-2}$ cm diameter wire are not shown. They were used to evaporate a tantalum film over most of the glass surface in order to adsorb residual gases and thus decrease the pressure. This proved to be very effective and estimated pressures of $10^{-12}$ to $10^{-14}$ mm Hg. were obtained. The tube also contained a source of Ba which could be deposited on part of the $W$ point and loop. It consisted of a coil of .020 tantalum wire heavily coated with BaO + BeO. The coil con-

sisted of 8 turns, .3 cm diameter. The center of the coil was 2 cm from the point, and the axis of the coil formed an angle of about 30° with the axis of the tube. Since the composition of the source is essentially that of Batalum getters which are known to evaporate Ba, it is assumed that nearly pure Ba evaporated from it. There is, however, the possibility that some BaO evaporated with the Ba. The tube has been in an operable condition for about 12 years.

The tube was baked at 400°C for one hour. Then all the parts were glowed or heated to outgas them. It was rebaked at 410°C for three hrs. The *W* loop, Ta filaments, and Ba coil were heated so hot that further heating did not increase the pressure. The tube was sealed off at a pressure of $2 \times 10^{-7}$ mm with both Ta filaments at a high temperature. Soon after the tube was sealed off the patterns for clean *W* and Ba on *W* were quite unsteady: there were rapid variations in intensity of small bright spots or flickering and there were more gradual changes in the pattern over large areas. After glowing the *W* loop, Ta filaments, and Ba coil many times and at successively higher temperatures, the flickering disappeared completely and the slow large-area changes became less pronounced or required a longer time to appear. Characteristic patterns could be reproduced at will for any particular treatment. In the early stages the clean *W* pattern changed noticeably in one minute; later, the time required for a definite change to occur increased to ten minutes, then one hour, then one day, and finally one week or even one month. The effect of the residual gas was to enlarge the 211 planes and darken the 111 zone. The effect of this residual gas could be removed at $T = 800$°K in a minute. We suspect that the residual gas is mostly CO.

During the course of many experiments, the *W* loop was raised from 2200 to 2800°K. Gradually the voltage required to obtain a field emission of 10 microamps from clean *W* increased from 6000 to 9000 volts. In accordance with Mueller's results[2] we ascribe this to an increase in the radius of curvature of the point from 2 to $3 \times 10^{-5}$ cm.

The lower portion of Fig. 1a shows an enlarged view of the *W* point and indicates that the tip of the point consists of a single crystal. Hence all possible crystal orientations should be represented on the surface. Figure 1b shows a still further enlargement of the tip of the point. We assume that near the tip of the point the surface is a paraboloid. If the origin is taken at the vertex, and $y$ is measured along the axis and $x$ perpendicular thereto, the equation for the paraboloid is

$$y = x^2/2r \tag{1}$$

where $r$ is the radius of curvature of the "point." The field near such a sur-

face can be calculated if the anode is a larger paraboloid whose equation is given by

$$y = x^2/2(2d + r) \tag{2}$$

provided the origin is at its vertex, $d$ is the distance between the two vertices, and the axes of the two paraboloids are the same. The field $F_h$ for points at which $y = h$, is given by*

$$F_h = K_h V = \frac{2V}{r(1 + 2h/r)^{\frac{1}{2}} ln(1 + 2d/r)} \tag{3}$$

where $h$ is distance along the axis of the small paraboloid. At the tip or vertex of the $W$ point, $h = 0$ and

$$F_0 = \frac{2V}{r ln(1 + 2d/r)} \equiv K_0 V \tag{4}$$

$$\text{Hence } F_h = F_0 \Big/ \left(1 + \frac{2h}{r}\right)^{\frac{1}{2}} \tag{5}$$

For an angle of 60° with the axis, $h/r = .47$ and

$$F_h = .72\, F_0$$

For clean $W$, this predicts that the emission density at an angle of 60° with the axis should be .008 of the emission density along the axis. For angles less than 10° the field and emission densities should differ only slightly from that for the axis values. Experiment shows that these predictions are qualitatively fulfilled.

Subsequent photographs will show that for clean $W$ most of the emission comes from regions which surround the 100 plane. For the 611 plane $\varphi =$ 4.4 volts.[5] The area of these highly emitting regions corresponds to about $\frac{1}{3}$ of the area of the screen which in turn corresponds to about $\pi r^2$ $cm^2$ on the $W$ point. Hence we have taken the highly emitting area to be $r^2$ $cm^2$. The highest emitting areas make an angle of about 25° with the axis of the $W$ point or with the 110 direction. From Eq. (3) we calculate that the field is about .924 that at the tip of the point.

In order to obtain a value of $r$, the radius of curvature of our $W$ point, we proceeded as follows: We observed the emission current $i$ as a function of the applied voltage $V$ and plotted $\log i - 2 \log V$ vs $1/V$. Straight lines were obtained whose slopes and intercepts for clean $W$ depended on the highest temperature and time at which the $W$ loop was glowed. We then plotted a similar family of theoretical lines for various assumed values of $r$. The ex-

* We are indebted to our colleague S. P. Morgan for Eqs. (1) to (4).

perimental curves agreed fairly well with the theoretical ones for both slope and intercept. The values of $r$, deduced from the location of the experimental lines, ranged from $2 \times 10^{-5}$ cm for low temperature treatment to $3 \times 10^{-5}$ cm after repeated glowing at 2800°K.

The theoretical family of curves was based on the Fowler-Nordheim equation modified for the electron image effect:[3]

$$j(\text{amps/cm}^2) = \frac{1.5 \times 10^{-6} K^2 V^2}{\varphi} \epsilon^{-(\varphi^{3/2}/KV)6.8 \times 10^7 f(x)} \tag{6}$$

in which the field $= K \times V$ volts/cm and $x = \dfrac{3.78 \times 10^{-4}\sqrt{KV}}{\varphi}$. Nordheim[6] gives a table of $f(x)$ vs $x$. From this we plotted $f(x)$ vs $KV$ for $\varphi = 4.4$ volts, and found that for values of the field $KV$ from 15 million to 95 million volts/cm, $f(x)$ was given by

$$f(x) = .968 - 5.54 \times 10^{-9} KV \tag{7}$$

This range of field covers nearly all values which are usually encountered in field emission. By substituting Eq. (7) in Eq. (6) and putting $\varphi = 4.4$ we obtain

$$j = 3.42 \times 10^{-7} K^2 V^2 \epsilon^{3.47} \epsilon^{-(6.06 \times 10^8/KV)} \tag{8}$$

For a $W$ point in which the major part of the current comes from an area of about $r^2$ $cm^2$ having a work function $\varphi$ of 4.4 volts, and making an angle of about 25° with the axis, the current $i$ in amperes is given by

$$\log i = -5.00 + .04 + 2 \log r + 2 \log K + 2 \log V - \frac{2.64 \times 10^8}{KV} \tag{9}$$

in which $K$ is a function of $r$, $h$, and $d$ given by Eq. (3). For the experimental tube the point to anode distance was 4.0 cm. Since the field at the point will vary only slightly with the exact shape of the anode we have put $d = 4.0$ cm. For $r = 3 \times 10^{-5}$ cm

$$K_0 = 5300 \text{ cm}^{-1}. \quad K_h = .924\ K_0 = 4900 \text{ cm}^{-1}$$

and

$$\log i = -7 + .41 + 2 \log V - 5.40 \times 10^4/V \tag{10}$$

Similar equations can be deduced for other values of $r$.

From Eq. (6) it follows that the current density and hence the screen brightness depend on the work function $\varphi$, and the field $KV$: the current density increases as $\varphi$ decreases, and increases as $K$ increases. Since the single crystal point exposes all possible planes and since it is known that different

planes of $W$ have different work functions, it is to be expected that different regions of the point will emit various current densities. Quantitative calculations show that the ratio of highest to lowest current densities should be at least 300. The current density might also be expected to vary if $KV$ varies due to small local elevations or depressions from the paraboloid. Such hills and valleys or ridges might result in 10-fold variations in current density. Both types of variations are illustrated in photographs which are to follow.

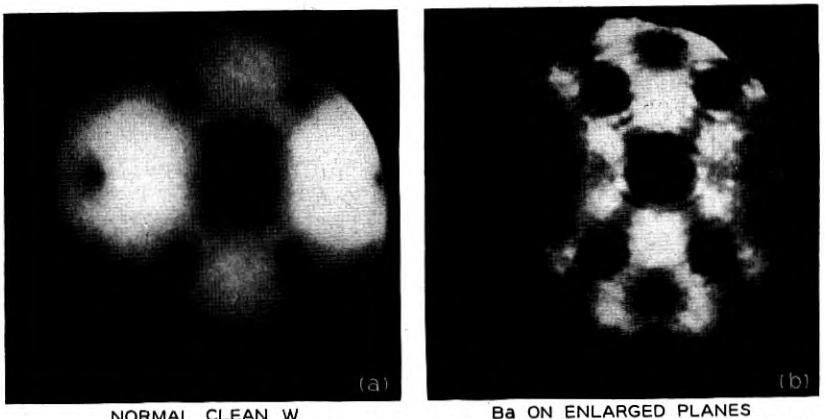
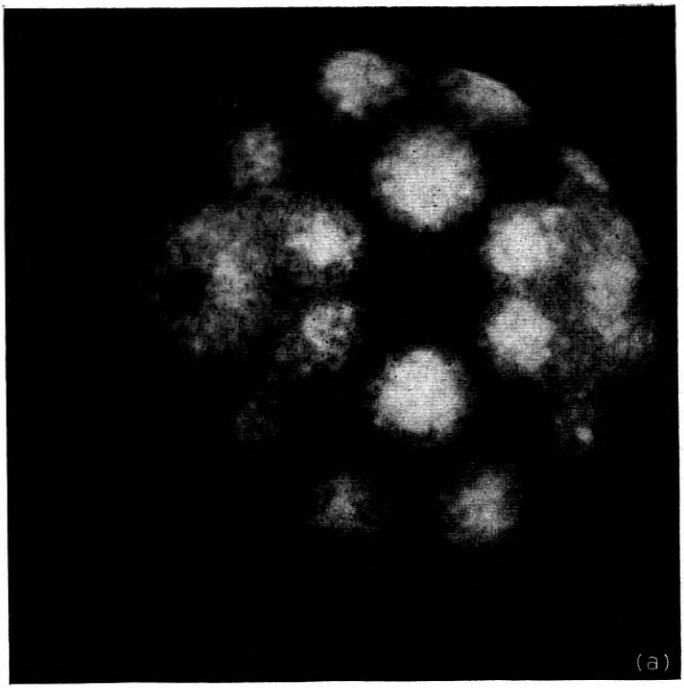



Fig. 2—Field emission patterns from normal clean tungsten and from *Ba* on $W$ with enlarged planes. Note elliptical structure around central dark region in (b). The treatment which conditioned the $W$ point is given above the black line; the constants used in taking the photograph are given below the lines.

The various physical constants pertaining to the experimental tube are summarized in the left part of Fig. 1b. The error in the value of $r$ is probably less than 20% and the error in the magnification is probably less than 30%.

## Field Emission from a Paraboloidal Surface of a $W$ Single Crystal

Reproductions of photographs of the screen are shown in Figs. 2, (a) and (b), and 3 (a). Figure 3 (b) shows the indices of the principal regions and zones appearing in these and subsequent reproductions. For the experimental tube and for a point with a radius of curvature of $3 \times 10^{-5}$ cm, the distance between two nearest 211 planes is about 2000 Å. Other dimensions can be scaled off on the reproductions. In these and other reproductions, the treatment given the surface is described in the upper part of the print-



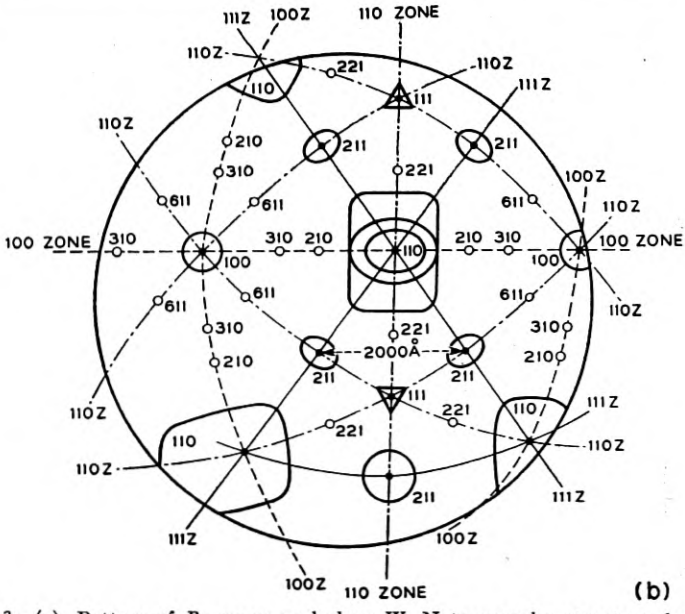




Fig. 3—(a). Pattern of *Ba* on normal clean *W*. Note granular structure due to *Ba* clusters.

(b). Location of principal planes and zones as they appear on the screen.

ing; the lower part describes the conditions under which the photograph was taken. For example, for Fig. 2 (a) the $W$ point was heated to 2400°K with zero applied voltage for 30 sec. The photograph was taken with the point at room temperature, assumed to be 300°K, with +9000 volts applied to the anode, while a field emission current of 40 microamps was drawn from the point, and for an exposure of $\frac{1}{5}$ sec.

From Fig. 2(a) it is easily seen that for "normal clean $W$" which we define as $W$ glowed at 2400°K, the 110, 211, and 100 regions emit poorly; the 111 region emits moderately; the greatest emission density comes from a rather broad band surrounding the 100 region. The center of this band makes an angle of about 20° with the 100 direction. The 611 plane lies in the central part of this band. In Jenkins' *Report*[3] it is shown that the 110, 211, and 100 dark regions are planes and that the extent of these planes can be increased by applying high positive fields while the point is at temperatures near 1200°K.

Figure 2(b) shows a photograph for a point which has been treated in this manner; Ba was then evaporated onto the $W$ and the $W$ loop was heated until the Ba migrated over the surface. Figure 3(a) shows the emission from migrated Ba on normal clean $W$. In Figs. 2(b) and 3(a) the amount of Ba is rather small, about .10 monolayer. For Ba on tungsten the emission comes from small "circular" regions with an average diameter of 100 Å. We interpret these to be small regions in which the Ba atoms cluster together, thus reducing the work function more than in neighboring regions, and hence we call such regions clusters.

A careful inspection of the negatives for clean $W$ show that, in regions other than the 110, 211, and 100 planes, the emission shows a granular structure with small regions of the order of 100 Å diameter surrounded by slightly darker regions. We believe this to be due to submicroscopic facets on the paraboloidal surfaces; these facets form small hills or plateaus and valleys. On the small hills the local field is slightly greater than in the valleys and hence the emission from clean $W$ is slightly greater than from the valleys. On the other hand we believe that the Ba atoms are held more firmly in the valleys or troughs where they can contact more $W$ atoms, thus accounting for the Ba clusters discussed in the preceding paragraph.

Figure 2(b) also shows a series of large elliptical rings with their center in the 110 plane and their major axis in the 100 zone. The evidence for these is especially pronounced in the 111 zones. The separation between them is about 170 Å. This suggests that the 110 plane and the region surrounding it consists of a series of plateaus of 110 planes elliptical in shape.

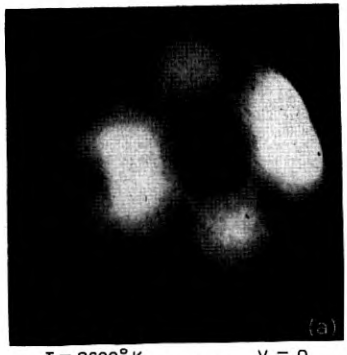
## Effect of Temperature in Determining the Size of 110, 211 & 100 Planes

Figure 4 shows the effect of glowing the point at temperatures from 2600 to 1200°K. In each case the temperature was held at a constant value until an approximate steady state was reached. When the photographs were taken, the applied voltages were adjusted slightly so as to maintain a constant average screen intensity. After the 1200°K glowing, the 2400°K was repeated. All the films came from the same pack and were developed together. Other tests showed that the result for a given temperature did not depend upon temperature treatments that preceded it.

From this series of tests we conclude: (1) A single crystal of tungsten which is approximately a paraboloid with a radius of curvature of $3 \times 10^{-5}$ cm and which has been heated for approximately an hour at 2400°K, assumes a surface configuration in which 110, 211, and 100 planes develop. The remainder of the surface consists of small facets or crystal planes of the order of 100 Å across. (2) The 110 planes are rectangular in shape with rounded corners; the 211 planes are slightly elliptical; the 100 planes are circular. (3) As the temperature of glowing decreases from 2600 to 1200°K the diameter of the 100 plane increases from 260 Å to 500 Å; the major axis of the 211 plane increases from 400 Å to 750 Å; the 110 plane changes only slightly in size. (4) At 1200°K small 310 and 111 planes develop and step-like structures develop in the 111 zone between the 110 and 211 planes. (5) The rate at which the surface changes from one steady state configuration to another decreases with temperature. Below 1050°K this rate becomes too slow for convenient observation. (6) At 1050°K in one hour the 211 planes enlarge and the *W* atoms migrate in the 111 direction on this plane; the excess *W* is deposited on the adjoining paraboloidal surface.

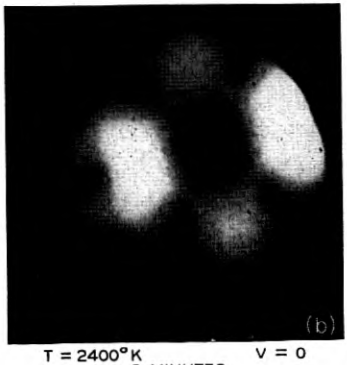
## Effect of High Fields and Temperature in Determining Surface Configuration

Figure 5 shows a series of photographs in which normally clean *W* was kept at 1400°K while approximately 8000 volts was applied to the anode for one minute to 39 minutes. The last photograph was taken after the *W* was heated to 1430°K for three minutes without an applied voltage. All photographs were taken with the *W* at room *T*. The voltage was adjusted until the total field emission was 30 microamps. As the treatment progresses: the 211 and 100 planes enlarge, while the 110 planes change their shape; the emission density from the 100 and surrounding regions decreases while that near the 111 and regions surrounding the 110–211 planes increases;

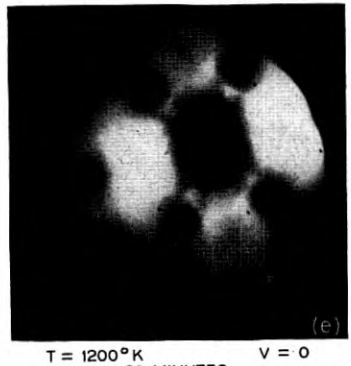



T = 2600°K V = 0
10 SECONDS
V = 9230 i = 17 μA
1/5 SECOND

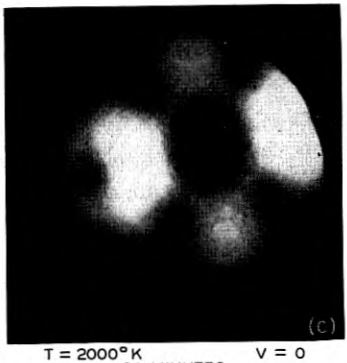

T = 1600°K V = 0
20 MINUTES
V = 9220 i = 18 μA
1/5 SECOND



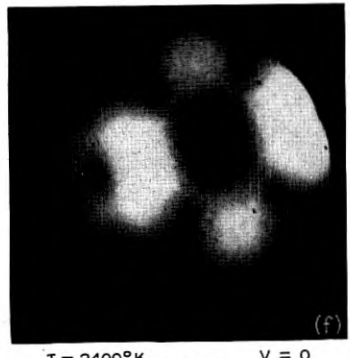

T = 2400°K V = 0
3 MINUTES
V = 9230 i = 17 μA
1/5 SECOND




T = 1200°K V = 0
20 MINUTES
V = 9170 i = 23.2 μA
1/5 SECOND

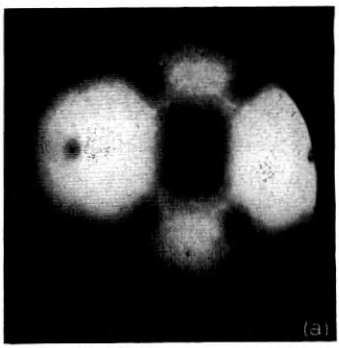



T = 2000°K V = 0
20 MINUTES
V = 9220 i = 18 μA
1/5 SECOND



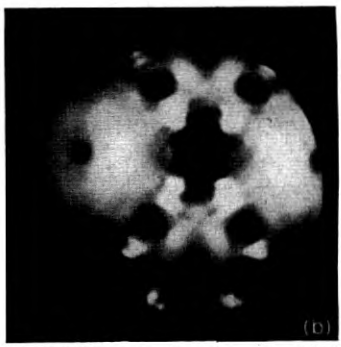

T = 2400°K V = 0
3 MINUTES
V = 9230 i = 16.8 μA
1/5 SECOND

Fig. 4—Effect of temperature on size and shape of principal planes for clean $W$.

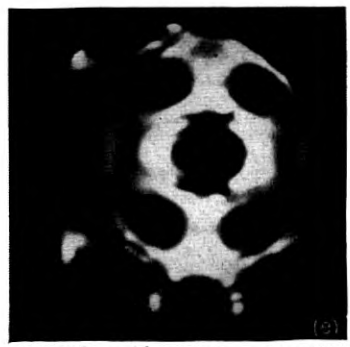


T = 2400°K V = 0
1 MINUTE
V = 8900 i = 30 μA
2/5 SECOND

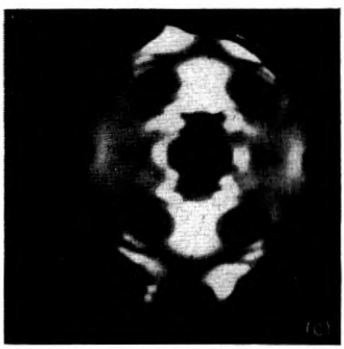
T = 1400°K i = 50 μA V = 8000
21 MINUTES
V = 7850 i = 30 μA
2/5 SECOND



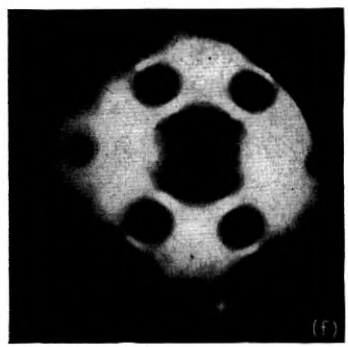
T = 1400°K i = 17 μA V = 8300
1 MINUTE
V = 8800 i = 30 μA
2/5 SECOND



T = 1400°K i = 25 μA V = 7900
39 MINUTES
V = 8050 i = 30 μA
2/5 SECOND



T = 1400°K i = 30 μA V = 8200
6 MINUTES
V = 8400 i = 30 μA
2/5 SECOND



T = 1430° K V = 0
3 MINUTES
V = 8500 i = 30 μA
2/5 SECOND

Fig. 5—Effect of high fields at 1400°K on size and shape of principal planes for clean *W*.

in the early stages the intensity increases in the 111 zone "steps" and in the region beyond the 211 plane going toward the 111 plane.

We have repeated similar experiments, about 20 times, varying the temperature and field. The rate at which the changes take place increases with temperature and with field. The rate increases more rapidly than the first power, perhaps as the square or cube of the field. In one experiment the direction of the field was reversed; this did not affect the kind of changes nor the rates. In another experiment, the temperature was reduced to 800°K while a field of about 40 million volts/cm was applied; in one hour the 211 planes enlarged perceptibly, the intensity increased just beyond this plane in the 111 direction, and the "steps" in the 111 zone appeared. For fields $>40 \times 10^6$ volts/cm and $T = 1500$ to 1600°K, most of the surface can be developed into planes; the highest emission comes from broad lines where the planes intersect; and the voltage necessary to obtain a given current is greatly reduced.

Many of these observations can be explained readily if we postulate that $W$ atoms can be polarized and that such atoms will tend to move from low to high fields. The effects of such polarization forces will of course be superimposed on the forces which tend to hold the $W$ atoms in certain crystalline positions. Consider normal clean tungsten after glowing at 2400°K, and concentrate attention on the 211 plane. In the previous section we concluded that above 1050°K, $W$ atoms are mobile on the surface and travel in the 111 direction until they reach the adjoining paraboloidal surface. A model of the 211 plane for $W$ shows that, in the 111 direction, the atoms touch each other but the rows of atoms are separated by 1.635 atom diameters; hence we would expect that atoms on this plane could move quite readily in the 111 direction. Now consider the effects of a high field. At the edge of the 211 plane the field will be larger than average, while at the center it will be less than average so that the field must increase toward the edge. Hence there should be a net force due to the field tending to take atoms off the edge of the plane, and the rate at which the planes develop should be greater with a field than without. Furthermore the extent to which the plane develops should be greater with a field. Since the polarization and the field gradient are probably proportional to the field, and the force is the product of the two, one would expect the field effect to increase with the square of the field. Because the force on the polarized atom due to the field is away from the surface for both positive and negative fields, the field effects should be independent of the direction of the field.

The polarization postulate also explains the observation that after 39 minutes of applied field most of the emission comes from the 111 region and regions surrounding the 211 and 110 planes; and that the emission from

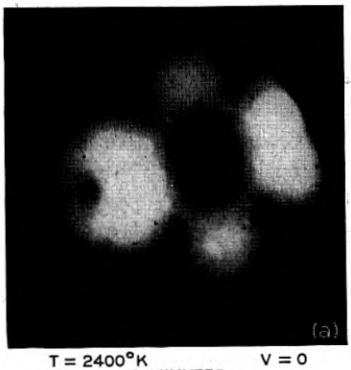
the regions surrounding the 100 plane, such as the 611 or 310 regions, is greatly reduced. Since the mobility of *W* atoms on 211 and 110 planes is greater than that on 100 planes, we conclude that the forces required to move a *W* atom on a 211 or 110 plane from a position of equilibrium to a neighboring position of equilibrium are less than those required to move an atom on a 100 plane. A study of models of these planes leads to the same conclusion. The field effect reduces the displacement work on all these planes, and hence we would expect that the 211 and 110 planes would change their shape faster than the 100 plane. The *W* atoms involved in such changes pile up in regions near their respective planes and thus increase the local field in such regions. Since the rate of migration increases rapidly with the field, these regions will grow at a still faster rate than before. Hence we would expect that such regions would pile up *W* atoms at the expense of other regions in which the migration rate started out more slowly. The experimental results, interpreted in this way, lead to the conclusion that high fields can result in movement of *W* atoms over distances of several thousand Angstroms.

We have made about five observations in which this effect continued even at room temperature. If by temperature and field treatment one obtains a pattern in which the emission from a few small spots materially exceeds that of other regions, and if the temperature is then reduced to 300°K while the voltage is kept on for hours, it is found that one or two of these spots will grow in size and intensity while other spots and regions get relatively less intense.

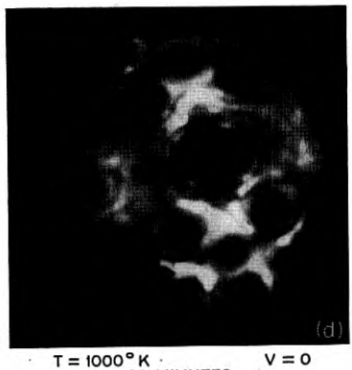
In such cases and in all cases of enhanced local field emission, the pattern can be brought back toward the normal condition by merely glowing the point at temperatures near 1000°K. The more the pattern differed from the normal one, the lower the temperature required to observe changes back toward the normal. One instance of the tendency to approach a normal pattern is that of the last photo in Fig. 5. A series of photos showing this tendency is given in Fig. 6.

### Effect of Temperature in Changing an Abnormal to a Normal Pattern

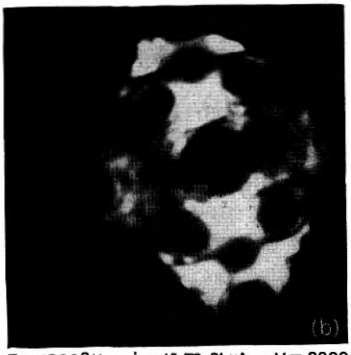
Figure 6 shows first a pattern of normal clean *W* for 2400°K; then follows a pattern produced by treatment at 1200°K with 8900 volts applied for 90 min. During this time the emission increased from 15 to 31 microamps; the next four patterns show the effect of glowing at successively higher temperatures for about an hour. A comparison of photos b and c shows that at 900°K the changes are slight: only a few of the brighter spots near the perimeter of the 110 and 211 planes have decreased in intensity.



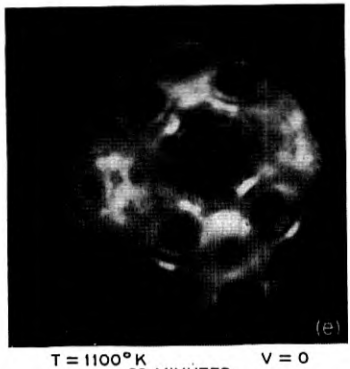

T = 2400°K V = 0
3 MINUTES
V = 9250 i = 14.5 μA
1/5 SECOND



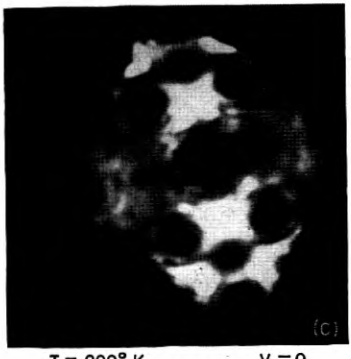

T = 1000°K V = 0
90 MINUTES
V = 8940 i = 21 μA
1/5 SECOND



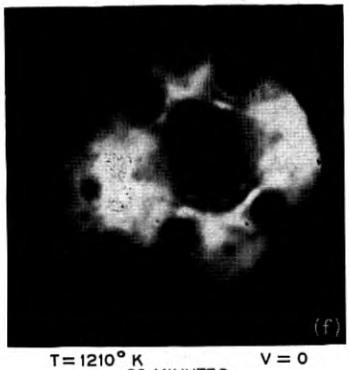

T = 1200°K i = 15 TO 31 μA V = 8900
90 MINUTES
V = 8840 i = 31 μA
1/5 SECOND




T = 1100°K V = 0
60 MINUTES
V = 8970 i = 17.5 μA
1/5 SECOND




T = 900° K V = 0
60 MINUTES
V = 8870 i = 28 μA
1/5 SECOND




T = 1210° K V = 0
60 MINUTES
V = 9000 i = 15 μA
1/5 SECOND

Fig. 6—Effect of temperature in changing an abnormal to a normal pattern.

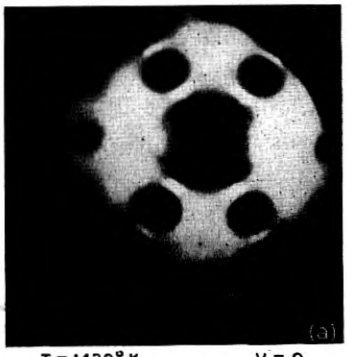
That the surface has changed detectably is shown by the fact that at room $T$ the emission decreased from 31 to 28 $\mu a$ even though $V$ increased from 8840 to 8870 volts. This means that the abnormal "bumps" have decreased slightly. These effects get progressively more pronounced as the glowing $T$ was increased to 1200°K. The 611 region is now the brightest; the small 310 and 111 planes are still poor emitters. A continuation of the test showed that the 111 plane became normal after one hour at 1300°K. The 310 plane became normal after one hour at 1500°K. After one hour at 1600°K, the pattern looked like normal clean $W$ except that the 100 and 211 planes were larger than in photo a of Fig. 6; it was similar to Fig. 4, photo d, after glowing at 1600°K.

## PART II: EMISSION AND ADSORPTION PROPERTIES OF Ba ON $W$

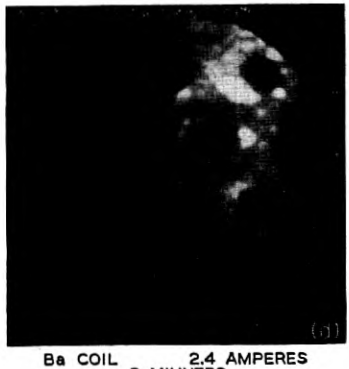
### Field Emission from Ba on $W$

Figure 7 shows a series of photographs in which successive units or "shots" of Ba were vaporized onto the $W$ point. The geometry of the tube was such that the greatest rate of deposition occurred on the upper right 611 region (Fig. 3b) and tapered off to zero on an "arc" which passes slightly to the left of the upper left 211, central 110, and lower right 211 planes. (Photo f of Fig. 7) In this series a "shot" of Ba was produced by heating the Ba coil with 2.4 amps for one minute. Later calculations will show that one "shot" deposited about 0.5 to 0.7 of a monolayer in the 611 region so that 7 shots deposited 3 to 5 layers in this region and deposited about 1 layer near the "arc" region.

The first photo shows clean tungsten treated so as to enlarge the 100 and 211 planes and to modify the shape of the 110 plane; the remainder of the surface is approximately on a paraboloid. In these latter regions the emission density is nearly uniform. In the 110 plane on the negative, there is a clear but faint ellipse. This ellipse is enhanced by the Ba in photos b, c, and d. We believe this ellipse to be due to the edge of a 110 plane which extends over only part of the larger underlying 110 plane. At this edge the local field is larger than in nearby regions and hence produces slightly greater emissions even on clean $W$. When Ba is deposited on this plane the edge serves as a nucleation center for Ba clusters even at 300°K. Prominent clusters also appear on the edges of the 211 planes in photos b, c, and d. Clusters also appear on the paraboloidal surfaces. The existence of these clusters shows that Ba atoms can move over a short distance—about 200 Å—even at room temperature.

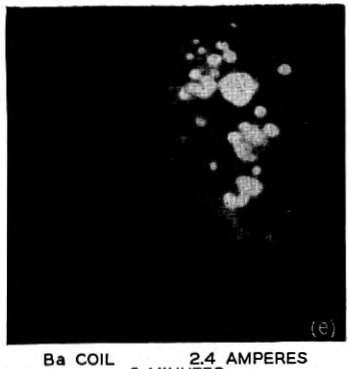


T = 1430° K, V = 0, 3 MINUTES
V = 8500, i = 30 μA, 2/5 SECOND

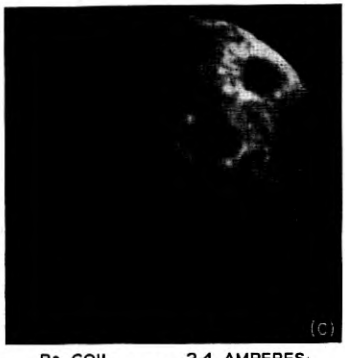


Ba COIL, 2.4 AMPERES, 3 MINUTES
V = 3150, i = 50 μA, 2/5 SECOND

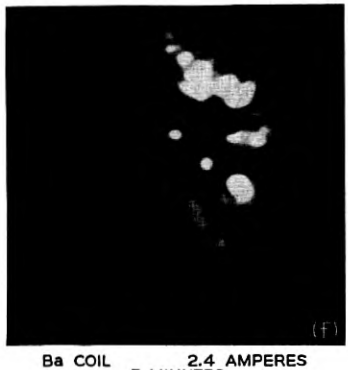
Ba COIL, 2.4 AMPERES, 1 MINUTE
V = 5000, i = 70 μA, 2/5 SECOND



Ba COIL, 2.4 AMPERES, 5 MINUTES
V = 3000, i = 50 μA, 2/5 SECOND



Ba COIL, 2.4 AMPERES, 2 MINUTES
V = 3370, i = 50 μA, 2/5 SECOND



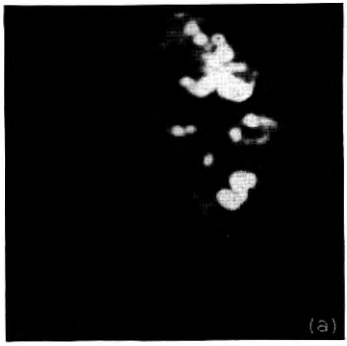
Ba COIL, 2.4 AMPERES, 7 MINUTES
V = 3100, i = 50 μA, 2/5 SECOND

Fig. 7—Patterns for successively increasing amounts of *Ba* on *W* with enlarged planes.

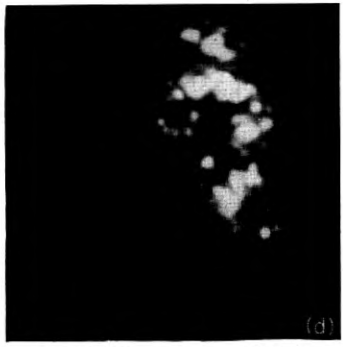
### Formation and Disappearance of Crystallites

As the number of Ba shots increases in Fig. 7, a large dark region develops and enlarges from the right 100 region. This is due to the well known fact that when the Ba concentration exceeds one monolayer the work function increases. If the Ba atoms remained where they were deposited, one would expect the patterns to consist of broad bright arcs whose centers were at the region of greatest deposition and whose radii would increase with amount deposited; for 7 shots of Ba one would expect a narrow bright arc of nearly maximum possible radius. Such an arc does indeed appear after 6 and 7 shots; but the regions with more than a monolayer are not dark as expected; instead there appear in these regions intensely bright large area emission centers. These just begin to show after 3 shots and become more prominent for 4 to 7 shots. These bright emission centers have properties different from those of the clusters previously described; they are larger, may appear on any plane, are in a continuous state of flicker even at 300°K, disappear at much lower glowing temperatures and can be observed at much lower applied voltages. In accord with Haefer,[7] we believe that they are due to Ba crystals which grow normal to the surface; hence the term crystallites seems appropriate. At the crystallites the local field should be much greater than the average field and hence they should be observable at low applied voltages. Different crystallites should have a range of sizes and hence a range of spot sizes. Crystallites should occur only for Ba concentrations greater than monolayers and hence should be nearly independent of the underlying tungsten. Because of the very high current densities through a crystallite, one would expect a considerable increase in local temperature, perhaps even to the melting point of Ba which would change the size and shape of the crystallite and hence the emission; this accounts for the flickering and agrees with the observation that the amount of flickering increases with the applied voltage and emission current. If the crystallites are solid Ba they should evaporate more easily than Ba clusters adsorbed on $W$. Furthermore the existence of similar crystallites for evaporated films has been deduced from electron diffraction experiments. Hence the evidence for crystallites is quite good.

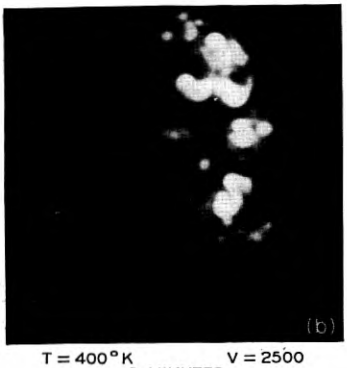
Figure 8 shows the evidence for the disappearance of crystallites in the temperature range 370 to 615°K. Note that, as $T$ increases from 370 to 510°K, the voltage required to get 50 microamps decreases from 3150 to 2500. Note also that up to 615°K no detectable amount of Ba migrates into the region beyond the outermost arc. The intensity of this arc decreases, presumably because $V$ is decreased.



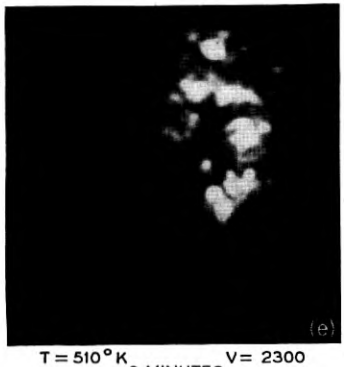

T = 370°K V = 2500
4 MINUTES
V = 2950 i = 50 μA
2/5 SECOND



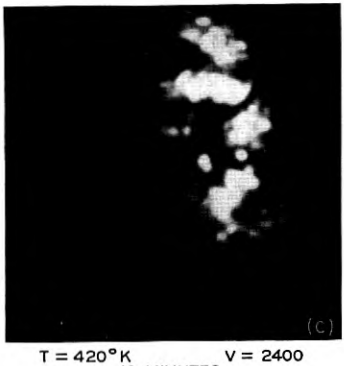

T = 450°K V = 0
5 MINUTES
V = 2650 i = 50 μA
2/5 SECOND



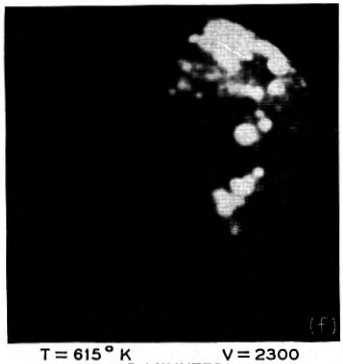

T = 400°K V = 2500
5 MINUTES
V = 2900 i = 50 μA
2/5 SECOND




T = 510°K V = 2300
6 MINUTES
V = 2500 i = 50 μA
2/5 SECOND




T = 420°K V = 2400
10 MINUTES
V = 2900 i = 70 μA
2/5 SECOND




T = 615°K V = 2300
5 MINUTES
V = 2570 i = 50 μA
2/5 SECOND

Fig. 8—The disappearance of crystallites from 370 to 615°K.

### Migration of Ba on $W$

Starting at about 800°K, Ba migrates over large distances—from one side of the paraboloid to the other or about 4000 Å. Figure 9 shows the steps in the migration process and the early stages of evaporation. The migration process can be followed continuously by observing the screen for moderate $V$ between 800 and 1000°K. Migration is essentially complete after five minutes at 1045°K. By then the crystallites have disappeared completely and with them abnormally high local fields. It is therefore possible to compute the field from the applied voltage. Then, as explained above, values of $\varphi$ and $\theta$ averaged for the whole surface can be computed. In this way we find that for photos c and d in Fig. 9, $\varphi = 2.00$ and $\theta$ is about 1.00.

### Evaporation of Ba on $W$

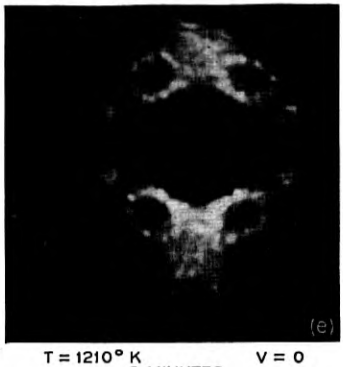
For $T \geq 1200°$ evaporation can be observed in five minutes. This is evidenced by the fact that after such glowing the value of $V$ required for a given current increases. The details of the evaporation are continued in Fig. 10. From the values of $V$ and $i$, values of $\varphi$ and $\theta$ have been calculated and are shown in Table I. From this table it appears that nearly all the Ba is evaporated in five minutes at 1600°K.

Further information on how the evaporation rate at a given $T$ varies with $\theta$ can be deduced from experiments for which no photos are shown. Suppose, in the above series of experiments, the point had been glowed for twenty minutes instead of five minutes, the calculated Ba concentration $\theta$ would have reached a somewhat lower value than .80, say .75. Still further glowing would have reduced $\theta$ only slightly. From this we conclude that at $T = 1200$ and $\theta = .75$ the rate of evaporation or $d\theta/dt$ is so small that additional glowing for twenty minutes reduces $\theta$ by small amounts. If now $T$ is raised to 1300°K for one minute, $\theta$ is substantially reduced, perhaps to .65. After five minutes at 1300°K, $\theta$ might be .55. After twenty minutes at 1300, $\theta$ might be .51. Long times at 1300°K might reduce $\theta$ to .50. Only by raising $T$ above 1300°K could $\theta$ be substantially reduced below .50. These observations suggest that the rate of evaporation of Ba on $W$ depends not only on $T$ but also on $\theta$: for a constant $T$ it is substantially 0 for all values of $\theta$ less than a critical value $\theta_c$. Above $\theta_c$, the evaporation rate increases rapidly with $\theta$, perhaps exponentially. Hence the probability of evaporation of a particular Ba atom depends on the proximity of neighboring atoms. This must mean that the forces between adsorbed Ba atoms are comparable to though smaller than the forces between Ba and $W$. A plot of the $\theta$ values in Table I vs $T$ would show that $\theta_c$ varies linearly with $T$ between 1130 and 1430 or between $\theta = 1.00$ and .18.


T = 800°K V = 2400
6.5 MINUTES
V = 2800 L = 50 μA
2/5 SECOND

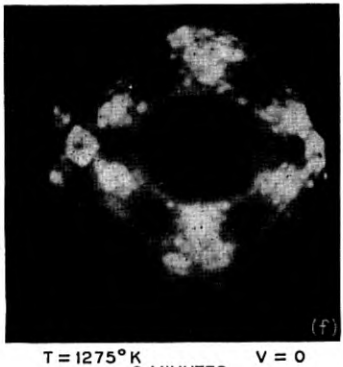

T = 1130°K V = VARIED
5 MINUTES
V = 2850 L = 50 μA
2/5 SECOND


T = 940°K V = 0
5 MINUTES
V = 2650 L = 50 μA
2/5 SECOND



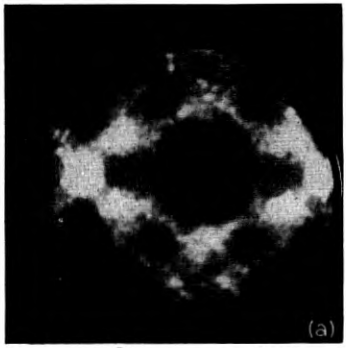

T = 1210° K V = 0
5 MINUTES
V = 3250 L = 50 μA
2/5 SECOND


T = 1045° K V = 0
5 MINUTES
V = 2800 L = 50 μA
2/5 SECOND

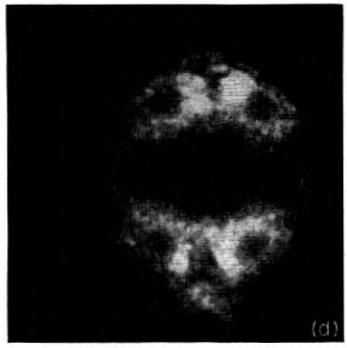



T = 1275° K V = 0
6 MINUTES
V = 4400 L = 50 μA
2/5 SECOND

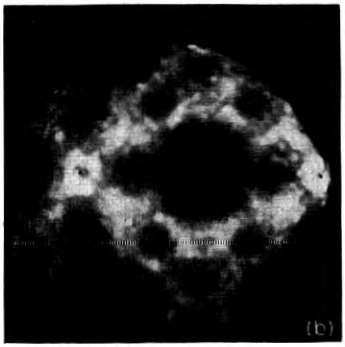
Fig. 9—Migration of *Ba* on *W* from 800 to 1045°K. Evaporation of *Ba* on *W* from 1130 to 1275°K.



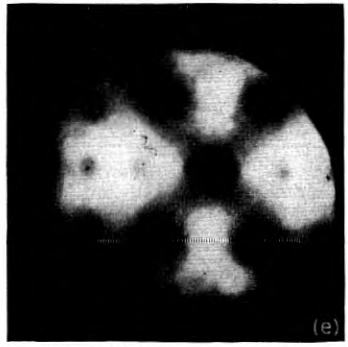

T = 1330°K V = 0
5 MINUTES
V = 5000 i = 50 μA
2/5 SECOND



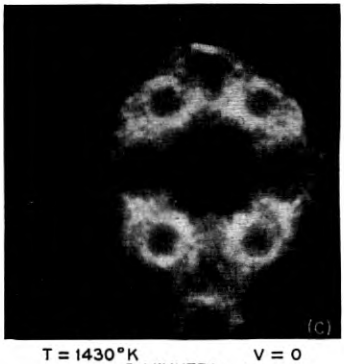

T = 1515°K V = 0
5 MINUTES
V = 8350 i = 70 μA
1/5 SECOND



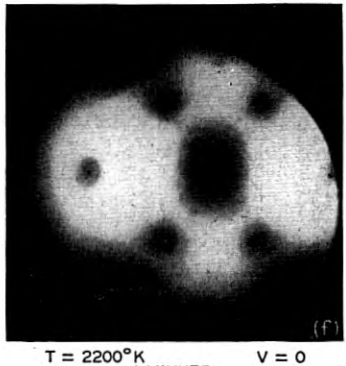

T = 1380°K V = 0
5 MINUTES
V = 6000 i = 50 μA
2/5 SECOND




T = 1670°K V = 0
3 MINUTES
V = 9250 i = 27 μA
2/5 SECOND




T = 1430°K V = 0
5 MINUTES
V = 7130 i = 70 μA
1/5 SECOND




T = 2200°K V = 0
1 MINUTE
V = 9000 i = 30 μA
2/5 SECOND

**Fig. 10—Evaporation of *Ba* on *W* from 1330 to 1670°K.**

### Temperature Effect on Clusters

The photos in Figs. 9 and 10 show that, for all values of $\theta$ from 1.0 to .10, the emission comes largely from clusters. It is interesting to observe, but difficult to portray in photographs, what happens if the temperature is raised above room $T$ but kept below that at which it had previously been heated to reduce $\theta$. As a specific instance we choose a case in which the treatment $T$ was 1380°K for five minutes—photo b in Fig. 10—for which $\theta = .28$. With an applied voltage at $T = 300$°K, the whole pattern and the clusters in particular are very steady. If $T$ is now raised to about 700°K, the clusters bordering on the 211 planes and those in the 111 zone appear to be agitated: the brightness of any one cluster fluctuates up and down and the center of the cluster moves over about half a cluster diameter. Clusters in other regions are perfectly steady. As $T$ is raised the 211 clusters agitate more violently and the clusters in nearby regions begin to agitate. At still higher $T$, the clusters in the 111 region begin to agitate but those surround-

Table I

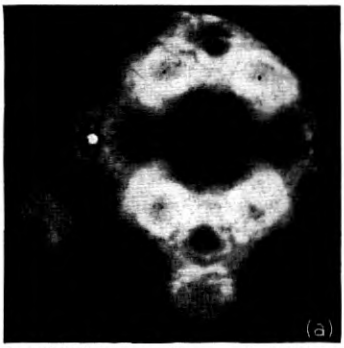
Dependence of $\varphi$ (Average Work Function) and $\theta$ (Layers of Ba) on Temperature and Time

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $T$ in °K..... | 1045 | 1130 | 1210 | 1275 | 1330 | 1380 | 1430 | 1515 | 1670 | 2200 |
| $t$ in min..... | 5 | 5 | 5 | 6 | 5 | 5 | 5 | 5 | 3 | 1 |
| $\varphi$ volts..... | 1.98 | 2.00 | 2.20 | 2.47 | 2.81 | 3.30 | 3.70 | 4.10 | 4.53 | 4.40 |
| $\theta$..... | ~1.0 | ~1.0 | .80 | .62 | .46 | .28 | .18 | .08 | ~.00 | .0$_0$ |

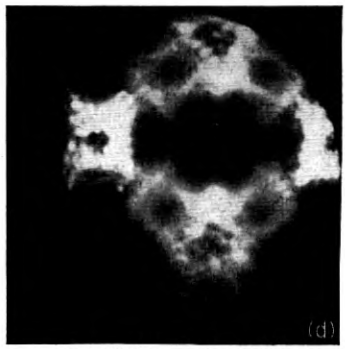
ing the 100 plane are still steady. For $T$ near 800°K all clusters show some agitation. At 1045°K, the clusters near the 110, 211, and 111 planes agitate so violently that individual clusters can no longer be distinguished but merge into one another producing bright bands which presumably reveal contours on the tungsten surface. However, the clusters in the regions surrounding the 100 plane agitate so slowly that in a photo of $\frac{1}{5}$ sec exposure they appear to be stationary. Photo a, Fig. 11 shows the result. Photo b shows the pattern immediately afterward at $T = 300$°K. These observations can be repeated as often as one pleases.

### Effect of Field on the Redistribution of Ba on $W$

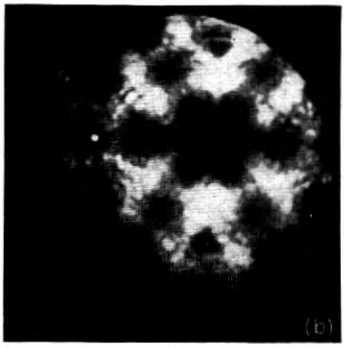
Photos c to f of Fig. 11 show that fields of 30 to 40 million volts/cm can redistribute some Ba from the 110, 211 and 111 regions to the regions surrounding 100. Photo c shows the pattern after glowing at 1430°K for five minutes with $V = 0$. Photo d shows the pattern at $T = 800$°K after 3 min. with $T = 800$°K and $V = 7380$ volts. When $T$ was reduced to 300°K, the pattern did not change appreciably. However, when $T$ was kept at 800°K for three minutes with $V = 0$, the pattern changed drastically as shown in photo e. Photo f shows that the redistribution is not the result of glowing



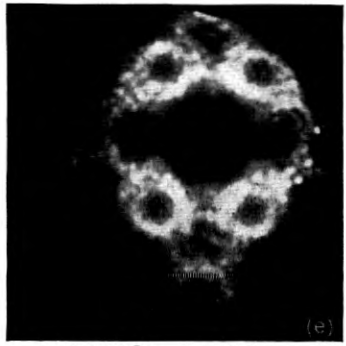

T = 1380°K V = 0 5 MINUTES
T = 1045°K V = 6900 5 MINUTES
T = 1045°K i = 70 μA V = 7100
1/5 SECOND



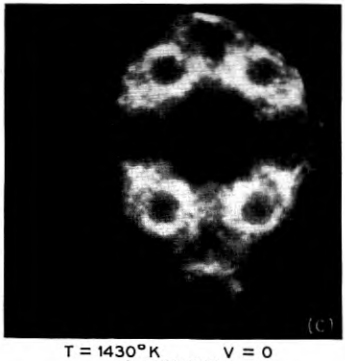

T = 800°K V = 7380
3 MINUTES
T = 800°K i = 70 μA V = 8000
1/5 SECOND



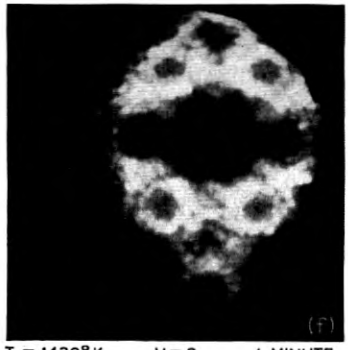

NO
TREATMENT
T = 300°K i = 70 μA V = 7750
1/5 SECOND




T = 800°K V = 0
3 MINUTES
T = 300°K i = 70 μA V = 7370
1/5 SECOND




T = 1430°K V = 0
5 MINUTES
T = 300°K i = 70 μA V = 7130
1/5 SECOND




T = 1430°K V = 0 1 MINUTE
T = 800°K V = 0 3 MINUTES
T = 300°K i = 70 μA V = 7320
1/5 SECOND

Fig. 11—Effect of temperature and field on the agitation of clusters and the redistribution of *Ba*. At 800 to 1000°K, *Ba* clusters are violently agitated near 211 and 111 planes but relatively stationary near 100 planes. High fields cause *Ba* to migrate from 111 and 211 regions to 100 regions.

at 800°K. To get the redistribution effect it is necessary to have both a high $T$ and a high $V$.

It is fascinating and instructive to watch this redistribution progress slowly. To do this we start with a pattern like that in photo e with $T$ = 300°K and $V$ = 7370 volts. The clusters are steady everywhere. $T$ is then raised to 800°K and the pattern observed for three to ten minutes. At first the clusters in the 110, 211, and 111 regions are in violent agitation while

TABLE II

SUMMARY OF ADSORPTION PROPERTIES FOR FIVE PLANES ON A SINGLE CRYSTAL

| Plane | 110 | 211 | 100 | 111 | 611 |
|---|---|---|---|---|---|
| **Clean W** | | | | | |
| Approximate $\varphi$ in volts | >4.9 | 4.8 | 4.6 | 4.40 | 4.2 |
| Size in Å for $T$ = 2400°K | 1700 × 1300 | 400 | 260 | <20 | <20 |
| Size in Å for $T$ = 1200°K | 1700 × 1100 | 750 | 500 | 75 | <20 |
| Size for large $T$ and $F$ | Changes shape | 1300 | 800 | 400 | <20 |
| °K for $W$ migration, $V$ = 0 | 1050 | 1050 | 1200 | | |
| °K for $W$ migration $V$ = 9000 | 800 | 800 | | | |
| **Ba on W; $\theta$ <1** | | | | | |
| Size of clusters in Å | 50 to 200; median 100 | | | | |
| °K for cluster formation | 300 | 370 | | 300 | 300 |
| °K for cluster disappearance | 370 | 420 | | | |
| °K for Ba migration | 370 | 420 | 800 | ~500 | ~600 |
| °K for evaporation: depends on $\theta$ | 1050 to 1600 | | | | |
| **Ba on W; $\theta$ >1** | | | | | |
| Size of crystallites in Å | 200 to 600; median 400 | | | | |
| °K for crystal formation | 300 | 300 | | 300 | 300 |
| °K for crystal disappearance | 600 | 600 | | 800 | 800 |

the few clusters in the 100 region are stationary. In ten seconds a few more clusters appear in the 100 region. When a new cluster appears it seems to do so suddenly. One gets the impression that one cluster pushed a neighboring one closer to the 100 plane or that one cluster grew at the expense of material from a neighboring one. Gradually as the concentration of clusters in the 100 region increases, the brightness of this region increases while the brightness in the 110, 211, and 111 regions decreases. A steady state is reached in three to five minutes.

REFERENCES

1. E. W. Müller, *Phys. Zs. 37*, 838, 1836.
2. E. W. Müller, *Zs. f. Physik 106*, 541, 1937; *126*, 642, 1949; *Naturwissenschaften*, July 1950.
3. R. O. Jenkins, *Reports on Progress in Physics 9*, 177, 1942–43.
4. F. Ashworth, *Advances in Electronics, 3*, Feb. 1951.
5. M. H. Nichols, *Phys. Rev., 57*, 1940 p. 297.
6. Nordheim, *Proc. Roy. Soc. 121*, 1928 p. 638.
7. R. Haefer, *Zeits. f. Physik 116*, 604, (1940).

# Heat Dissipation at the Electrodes of a Short Electric Arc

By L. H. GERMER

Platinum contacts are brought together 60 times a second, discharging on each closure a condenser of 0.01 mf capacity charged to 40 volts. The heat flowing along each electrode is calculated from a temperature difference measured by thermocouples, and from this is determined the energy dissipated at each contact. If there is no arc on closure, the energy is the same on the two contacts, and is small. If there is an arc between the contacts before they touch, about 58 per cent of its energy is dissipated upon the anode and about 42 per cent upon the cathode. The distribution is the same in an arc between clean "inactive" electrodes and in the entirely different kind of arc occurring between carbonized "active" electrodes. This information may be significant in developing an understanding of closure arcs which are the sole cause of the erosion of electrical contacts on closure.

THIS paper is an account of direct calorimetric measurements of the energy dissipated at positive and negative electrodes when they are brought together to discharge a condenser. The experiments are called for by the fact that the erosion at the closure of electrical contacts is due to arcing,[1] and understanding how the energy of a closure arc is distributed between the electrodes is likely to help in developing a comprehensive theory of this arc which in turn may aid in the control of contact erosion.

The experimental method is an adaptation to the present problem of a procedure[2] used earlier in which crossed wires are separated and brought together 60 times per second by means of a magnetic loudspeaker unit, each closure discharging a condenser which is recharged after the wires have been separated. For the present experiments the two wires are made of platinum and are rather heavy, and the flow of heat in each of them is measured by a pair of thermocouples. There is a known length of wire between the two thermocouples of each pair which are connected in series to oppose each other, so that a galvanometer in either circuit will give a deflection proportional to the difference in temperature across the wire.[3] The flow of heat along each wire is calculated from this temperature difference and the thermal conductivity and dimensions of the wire. After making some corrections this gives the amount of energy dissipated upon the electrode at each discharge of the condenser.

The two platinum test wires have diameters of 0.0635 cm and each is about 2.2 cm long from its end to the point where it is clamped in a very

[1] L. H. Germer, *Jl. App. Phys. 22*, 955 (1951).
[2] J. J. Lander and L. H. Germer, *Jl. App. Phys. 19*, 910 (1948), pp. 918–919.
[3] This is the experimental arrangement used by J. J. Lander in measuring heat flow in his determinations of Thomson coefficients. *Phys. Rev. 74*, 479 (1948), Fig. 3.

heavy copper block. The thermocouples are Chromel-Alumel wires of 0.012 cm diameter and one couple of each pair is welded to its platinum wire 2.0 cm from the point where the wire is clamped in its copper block; the other couple of the pair is electrically insulated by a glass coating and is buried in a deep hole in the block. The electrical contact is made between points of the platinum wires a little beyond the welded thermocouples, and opening and closing of the circuit is achieved by striking one of the platinum wires beyond the point of electrical contact with the insulated armature of a speaker unit vibrating at 60 cycles. Each heavy copper block with its platinum wire is mounted on a cantilever bar, the end of which can be moved by a screw to permit fine adjustment of the contacts. Adjustment can be made also by varying the voltage supplied to the speaker unit.

In order to minimize thermal disturbances this equipment is mounted on a heavy steel base, and the speaker unit, which dissipates about 0.01 watt during operation, is thermally insulated from the contacts by three concentric heavy aluminum covers each in very good thermal contact with the steel base. All of this equipment is covered by a silvered bell jar of 21 cm inside diameter. An aluminum covered Celotex housing surrounds the bell jar and the thermocouple galvanometer, except for a small glass window for reading the galvanometer. The galvanometer light is turned on for only about one second at the time of each reading. The experiments are made in a constant temperature room.

All of the significant tests consist in measurements of the heat flow along the platinum wires when they are brought together 60 times per second discharging at each closure a capacity of $10^{-8}$ f charged to a potential of 40 volts. At this potential an arc occurs between clean platinum electrodes if the circuit inductance is less than about $10^{-6}$ h, but there is no arc if the inductance is much higher than this.[1] If the electrodes are operated in the presence of any one of various organic vapors they become coated with carbonaceous material and arcing then occurs at every closure even when the inductance is quite high.[1] Measurements have been carried out under three different experimental conditions: (1) clean electrodes with a circuit inductance of $0.05 \times 10^{-6}$ h and an *arc at every closure*, (2) clean electrodes with a circuit inductance of $10 \times 10^{-6}$ h and *no arcing*, and (3) electrodes slightly carbonized by d-limonene vapor with a circuit inductance of $10 \times 10^{-6}$ h and an *arc at every closure*. The condition of arcing on every closure, or of complete absence of all arcing, was readily determined for each experiment by continuous oscilloscopic observation.[4] The potential of 40 volts was chosen as the highest at which there is never a second arc (in the reverse

[4] See reference 1, Fig. 1.

direction) which would impossibly complicate interpretation of the data. Experiments under conditions 3 were carried out with the limonene vapor pressure maintained at about the lowest value at which activation can be produced (0.06 mm Hg in most experiments). At this low pressure activation does not develop until the electrodes have been operating for some time, but when it develops the open circuit potential after an arc is −5 volts

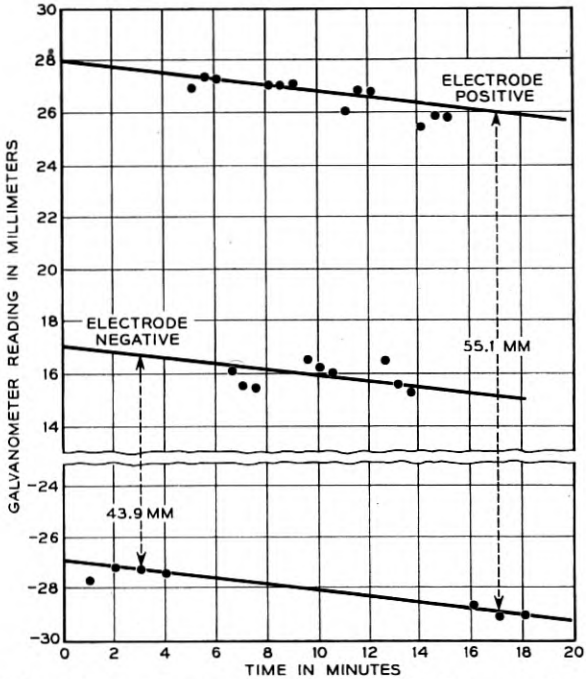



Fig. 1—Readings of the galvanometer in series with the thermocouples in the moving electrode, when the electrode was positive and when it was negative. Electrodes activated by vapor of *d*-limonene at a pressure of 0.06 mm Hg with 60 closures per second and an arc at every closure. Condenser of 0.01 mf charged to 40 volts, discharged on each closure through an inductance of $10 \times 10^{-6}h$. Galvanometer deflections in mm are transformed into ergs per closure by multiplying by 0.510. The experimental points obtained from this figure are marked by small arrows on Fig. 2.

which is also the value reached after arcing in the inactive condition. The observance of this open circuit voltage is proof that each arc is an arc in platinum vapor, and not carbon.

An example of data taken upon active (carbonized) electrodes with a circuit inductance of $10 \times 10^{-6}$ h (conditions 3), and an arc at every closure ending in an open circuit potential of −5 volts as verified by continuous observation of the oscilloscope recording the potential across the contacts,

is given in Fig. 1. The ordinates are readings of the galvanometer when it was in series with the thermocouples in the moving contact. The circuit for charging the condenser to 40 volts prior to each closure was turned on at 4 minutes with the potential of the moving contact positive. The potential was reversed at intervals of $1\frac{1}{2}$ minutes and finally turned off at 15 minutes. The deflections of 43.9 and 55.1 mm occasioned by the energy dissipated at the contact by the discharge of the condenser respectively when the contact was negative and when it was positive are translated into temperature differences of $\Delta T = 0.1213$ and $\Delta T = 0.1523$°C by multiplying by $\alpha r \beta$ where $\alpha = 3.36 \times 10^{-9}$ amp/mm of the galvanometer deflection, $r = 33.7$ ohms circuit resistance, and $\beta = 2.44 \times 10^{4}$°C/volt thermocouple sensitivity. Neglecting for the moment small corrections due to radiation and convection losses, these temperature differences are converted into heat flow along the wire of 22.4 and 28.1 ergs per closure by multiplying them by the factor $B = 184.5$ obtained from the dimensions of the wire, the thermal conductivity of platinum $k = 0.699$ watt/cm°C, and the factor 60 representing the number of closures per second.

The heat flow in one of the wires differs from the heat dissipated by the arcs upon that wire because of radiation and convection losses, and because the higher temperature of the positive electrode results in some conduction of heat to the negative electrode at their point of contact. It has been found expedient first to obtain data which are intended to be free from the last of these three sources of error and then to correct for radiation and convection losses as obtained by calculation.

The energy in the electrode wires corresponding to the average excess temperature of the wires above their surroundings (0.07°C) represents the total energy of about 500 arcs. Thus the large scale temperature distribution in one wire is inappreciably changed during the time the wires are in contact after an arc, and the transfer of heat from one to the other can be corrected for by making measurements of $\Delta T$ across each wire for different fractions $x$ of each cycle during which the wires are in contact and extrapolating the values so obtained to find $\Delta T_0$ for zero time of contact. Data of this sort for experimental conditions listed as (1) above are plotted at the lower left side of Fig. 2, and for conditions (3) at the lower right side of the figure; the $\Delta T_0$ values from these curves are written down on the first line of Table 1. On the upper half of the figure is plotted the total heat flowing along *both* wires as calculated by multiplying $\Delta T$ by the factor $B = 184.5$ (not correcting for radiation and convection losses).

All the solid circles on Fig. 2 (and Fig. 3 also) represent measurements upon the moving electrode, and the open circles measurements upon the stationary electrode. Differences between the solid circles and the open

TABLE I
HEAT DISSIPATION DATA

| | Inactive Electrodes | | | | Active Electrodes | |
|---|---|---|---|---|---|---|
| | (1) $L = 0.05 \times 10^{-6}h$ Arc at Every Closure | | (2) $L = 10 \times 10^{-6}h$ No Arcing | | (3) $L = 10 \times 10^{-6}h$ Arc at Every Closure | |
| | Anode | Cathode | Anode | Cathode | Anode | Cathode |
| 1. $\Delta T_0$ from Fig. 2 | .1513°C | .1217°C | — | — | .1587°C | .1188°C |
| 2. Correction to $\Delta T_0$ from Fig. 3 | +.0050 | −.0050 | — | — | +.0010 | −.0010 |
| 3. $\Delta T_0$ corrected for conduction at closure | .1563 | .1167 | .0445 | .0453 | .1597 | .1178 |
| 4. Ergs/closure from $\Delta T_0$ | 28.83 | 21.53 | 8.21 | 8.36 | 29.48 | 21.73 |
| 5. Radiation correction (ergs/closure) | 0.12 | 0.09 | 0.03 | 0.03 | 0.12 | 0.09 |
| 6. Convection correction (ergs/closure) | 1.63 | 1.23 | 0.32 | 0.33 | 1.71 | 1.19 |
| 7. Ergs/closure, final values | 30.58 | 22.85 | 8.56 | 8.72 | $w_+ = 31.31$ | $w_- = 23.01$ |
| 8. Total ergs/closure | | 53.43 | | $w_0 = 17.28$ | | 54.32 |

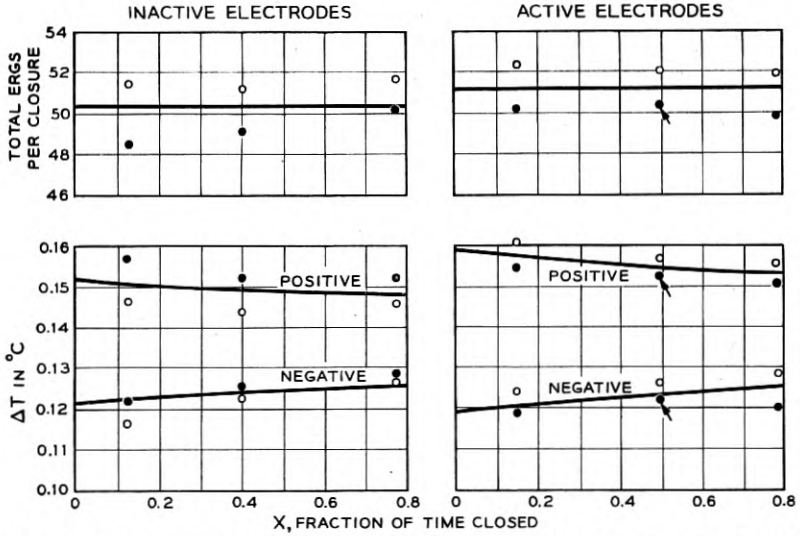



Fig. 2—Heat flow data at different values of $x$, the fraction of the time the electrodes are closed. Left hand curves, inactive electrodes, inductance $0.05 \times 10^{-6}h$, closing at 3.3 cm/sec. Right hand curves, active electrodes, inductance $10 \times 10^{-6}h$, closing at 2.5 cm/sec.

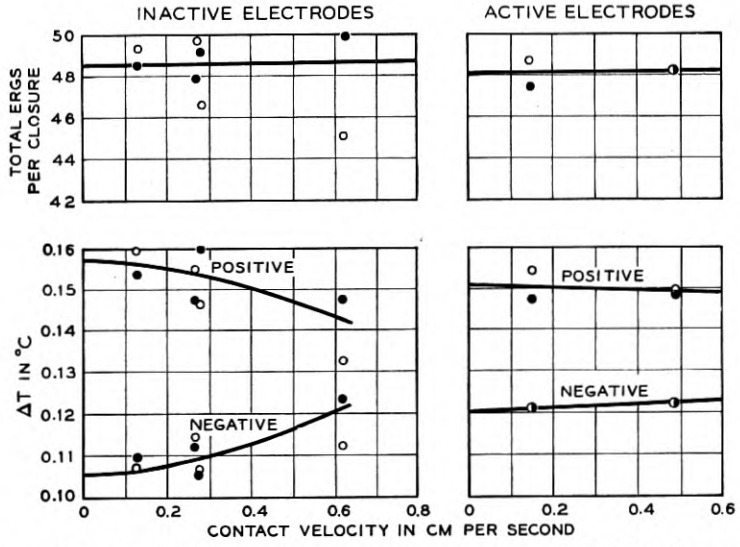



Fig. 3—Data for different velocities at closure, all for $x = 0.5$. Left hand curves, inactive electrodes, inductance $0.05 \times 10^{-6}h$. Right hand curves, active electrodes, inductance $10 \times 10^{-6}h$.

circles are to be attributed to differences between the electrodes, probably in the effective mean diameter. For observations in the active condition the solid circles of Fig. 2 are consistently higher than the open circles, but the opposite is true for measurements upon inactive surfaces. This reversal is not significant; it was brought about by an accident which necessitated rewelding the thermocouple to the moving electrode after the measurements upon the active surfaces had been completed and before the measurements upon the inactive surfaces. In earlier preliminary tests there was no difference of this sort; one must conclude that the welding operation altered the moving electrode. (In the case of the plots at the left of Fig. 3 below, some of the data were taken before the rewelding operation and some after.)

The continuous curves drawn on the lower half of Fig. 2 have the ordinates $W\epsilon_x/B$ for the positive electrode and $W(1 - \epsilon_x)/B$ for the negative, where $W$ is the total energy per closure represented by the horizontal lines at the top of the figure and $\epsilon_x$ is the fraction of the energy flowing down the positive electrode. $(0.5 < \epsilon_x < 1)$. The energy lost by the positive to the negative electrode by conduction per closure is clearly $(\epsilon_0 - \epsilon_x)W$. The temperature difference between the wires near the point of contact is $(W/B)(2\epsilon_x - 1)$ and, from analogy with the electrical formula for the spreading resistance of a circular contact of diameter $l$, the heat flow from one electrode to the other per second is found by multiplying this temperature difference by $lkx$ and the heat flow per closure by further dividing by 60. Equating the two expressions for heat flow one obtains $\epsilon_x = (60\,B\epsilon_0 + lkx)/(60\,B + 2lkx)$. The curves drawn on Fig. 2 have shapes determined by this expression with the two parameters $\epsilon_0$ and $l$ chosen to fit the experimental points. The resulting values of $\epsilon_0$ and $l$ are: $\epsilon_0 = 0.58$, $l = 11 \times 10^{-4}$ cm for the inactive electrodes and $\epsilon_0 = 0.57$, $l = 3.3 \times 10^{-4}$ cm for the active electrodes.

If the area of contact were truly circular and the contacting electrodes were crossed cylinders of perfect cross-section, these values of $l$ would be simply related by elastic theory to the forces $F$ holding the electrodes together when they are in contact. The formula[5] is $l = [6FD(1 - v^2)/E]^{\frac{1}{3}}$ where $D$ is the diameter of the wires, $E$ is Young's modulus for platinum and $v$ is Poisson's ratio. If we take[6] $E = 13 \times 10^8$ gm/cm$^2$, the above values of $l$ correspond respectively to forces of 5 and 0.1 grams weight. No significance can, of course, be attached to these values of force other than to observe that they are not wildly unreasonable.

The average lapse of time between the end of each arc and contact of

[5] A. E. H. Love, "The Mathematical Theory of Elasticity," Cambridge, fourth edition, 1927, page 197, equation 56.

[6] R. Holm, "Electric Contacts," Hugo Gebers, Stockholm, 1946, p. 389.

the electrodes at or near the place where the arc occurred can be calculated from the velocity of the moving electrode at contact and the average separation of the electrodes when the arc took place. Measurements of this separation have been made earlier.[7] Rough calculation of the local temperatures of the electrodes near the point of contact, making use of this average elapsed time, have shown that when contact is made these local temperatures are still far above the mean temperatures at the ends of the wires (perhaps higher by 10°C). The local temperatures reach the mean temperatures in a time which is very short in comparison with 1/60 second, and thus the conduction of heat from the positive to the negative electrode due to this local high temperature is not corrected for by the extrapolations of the curves of Fig. 2. This correction can be made by obtaining $\Delta T$ measurements for different electrode velocities and extrapolating to zero velocity. Such data are plotted in Fig. 3. It is obvious that the curves of the lower half of this figure must become horizontal as they approach zero velocity, but no other theoretical deduction has been made regarding their shapes. The differences between the $\Delta T$ values at the velocities of the data of Fig. 2 and at zero velocity are the required corrections, and these are written down on line 2 of Table 1. The correction seems to be small (0.001°C), or perhaps zero, for the active electrodes (right-hand side of Fig. 3) but amounting to about $\pm$ 0.005°C for the inactive electrodes (left-hand side). That the former should be smaller than the latter is in line with our knowledge that an arc between electrodes which are approaching each other will occur when they are farther apart if the electrodes are active than if they are inactive.[7]

The values of $\Delta T_0$ after applying the corrections of line 2 of Table I are written down on line 3. On line 3 are given also (columns 2) measured values of $\Delta T$ for inactive electrodes in a circuit containing an inductance of $10 \times 10^{-6}$ h which completely prevented any arcing. On line 4 are values of the energy dissipated upon the electrodes per closure calculated from the $\Delta T_0$ values of line 3. One must still consider corrections due to radiation and convection losses from the surfaces of the wires and radiation loss from the arc itself. These are taken up one at a time in the following paragraphs.

If the only correction were due to radiation from the surfaces of the warm wires, the heat put into the end of a wire per second $w$ would be related to $\Delta T_0$ by the equation

$$w = k\omega\Delta T_0/L + HA\Delta T_0/3, \tag{1}$$

where $k$ is thermal conductivity, $\omega$, $L$ and $A$ are respectively the cross-sectional area, length and surface area of the wire, and $H = 4T_0^3\sigma\epsilon$, the

[7] L. H. Germer, *Jl. App. Phys.* September 1951, Table I, line 3. (in press)

"outer conduction."[8] For $T_0 = 300°K$ and the emissivity $\epsilon = 0.05$ at room temperature,[9] the second term of this expression reduced to ergs per closure has the values listed as "radiation correction" on line 5 of Table I; these corrections are negligible.

The convection loss from a horizontal cylinder of diameter $D$ has been given[10] as $0.27A(\Delta T)^{5/4}/D^{1/4}$ in B.T.U. per hour with $\Delta T$ in °F, $D$ in feet and $A$ in square feet. For our system of units this becomes $4180A(\Delta T)^{5/4}/D^{1/4}$ ergs/sec. To make the differential equation for heat flow linear this can be written approximately $4180A(\Delta T_0/2)^{1/4}\Delta T/D^{1/4}$, and the heat put into the end of the wire per second taking account of convection loss would then be given by equation (1) with $H = 4180(\Delta T_0/2D)^{1/4}$. The second term of this expression reduced to ergs per closure has the values listed as "convection correction" on line 6 of the table.

That the heat lost from the arc itself is quite negligible is clear from estimates of the duration of the arc and of its superficial area. The arc time is $\pi(LC)^{1/2}$ which has the values 0.07 and $1.0 \times 10^{-6}$ sec respectively for the inactive and for the active surfaces. The average area of the arc which is effective in radiating is probably a great deal less than the area of the pit formed on one of the electrodes $\pi d^2/4$. In some experiments it was found[11] that $d^3 = 3.8 \times 10^{-11}$ cm³/erg. If this estimate of pit diameter is right for the present tests, and we take the arc temperature to be the boiling point of platinum[12] 4803°C and the duration of the arc $1 \times 10^{-6}$ sec, the radiation loss comes out to be 0.01 erg. This is a gross upper limit.

## Reduction of the Data

Not all of the energy in the charged condenser is dissipated in an arc on closure. *During* the arc some energy is dissipated by current flowing through circuit resistance, including spreading resistance in the electrodes at the site of the arc, and *after* the arc is over all of the remaining energy is so dissipated. We need to sort out the amounts of energy which are spent in these different ways in order to make a careful analysis of the data represented by the numbers on lines 7 and 8 of Table I.

The total energy is $e_0 = CV_0^2/2$ where $C = 10^{-8}f$ and $V_0 = 40$ volts in all of the experiments of this paper ($e_0 = 80$ ergs). The energy dissipated in an arc is $e_a = C(V_0 - V_1)v$, where $V_1 = -5$ volts is the potential across the

[8] H. S. Carslaw and J. C. Jaeger, "Conduction of Heat on Solids," Oxford, 1948, equation (6), p 119.

[9] This low value seems to be well established. See the paper by A. G. Worthing in a book "Temperature," Reinhold Pub. Co., 1941, Fig. 7 on p. 1175.

[10] W. H. McAdams, "Heat Transmission," McGraw-Hill, 1942, equations (13a) and (19), pp. 240–241.

[11] L. H. Germer and F. E. Haworth, *Jl. App. Phys. 20*, 1085 (1949), Fig. 5 on page 1088.

[12] Reference 2, Table II on page 914.

open electrodes when the arc is over and $v = 15$ (for platinum)[13] is the arc voltage assumed to be strictly constant during the life of each arc which is very closely true. The energy left in the circuit after an arc is over is $CV_1^2/2$. The total energy dissipated in the circuit is $C[V_0^2/2 - (V_0 - V_1)v - V_1^2/2]$ during the arc, plus $CV_1^2/2$ afterwards. When closure occurs without an arc (conditions 2) the total initial energy $CV_0^2/2$ is dissipated in the circuit. Some of the circuit energy appears in the electrode wires and is measured, as shown by the numbers of lines 7 and 8 of columns 2. It can probably be safely assumed that the fraction of the circuit energy which appears in the electrode wires is the same whether or not there is an arc. With this assumption, and knowledge of $V_1$ and v, we can use the data of columns 2 and 3 to calculate two parameters, $\eta$, the fraction of the circuit energy which appears in the electrode wires, and $\theta$, the fraction of the arc energy which is dissipated upon the positive electrode for the active condition. The data of columns 1 are not to be used with those of columns 2 and 3 because of a different electrical circuit and in consequence a different (and no doubt larger) value of $\eta$.

Quantities of interest are defined here:

| | | |
|---|---|---|
| total energy $e_0 = CV_0^2/2$ | | |
| arc energy $e_a = C(V_0 - V_1)v$ | | |
| energy which is measured (true value) | arc, $e_a + \eta(e_0 - e_a)$ | |
| | no arc, $\eta e_0$ | |
| factor by which all energy measurements are in error (i.e., experimental error) | | $\xi$ |
| fraction of arc energy at positive electrode | | $\theta$ |
| values obtained by measurement | arc, positive electrode, | $w_+$ |
| | arc, negative electrode, | $w_-$ |
| | no arc, total energy, | $w_0$ |

From the way these definitions have been given it is clear that

$$\begin{aligned} w_0 &= \xi\eta e_0 \\ w_+ + w_- &= \xi[e_a + \eta(e_0 - e_a)] \\ w_+ &= \xi[e_a\theta + (e_0 - e_a)\eta/2] \end{aligned}$$

These equations yield

$$\begin{aligned} \xi &= [(w_+ + w_-)e_0 - (e_0 - e_a)w_0]/e_0e_a \\ \eta &= w_0e_a/[(w_+ + w_-)e_0 - (e_0 - e_a)w_0] \\ \theta &= [2w_+ - (e_0 - e_a)\xi\eta]/2\xi e_a. \end{aligned}$$

The known numerical values of $C$, $V_0$, $V_1$ and $v$ give, in ergs, $e_0 = 80$, $e_a = 67.5$, circuit energy dissipated during an arc $= 11.25$, circuit energy dis-

[13] Reference 1, Table II, page 957.

sipated after an arc = 1.25. We identify $w_+$ and $w_-$ with the numbers so designated in Table I. When the value of $w_0$ is taken from the table we implicitly assume that the electrical spreading resistance at the contact is so much larger than the rest of the resistance of the electrode wires that substantially all of the heat $w_0$ is generated in the spreading resistance and not along the wires. With this assumption we obtain,

$$\xi = 0.765$$
$$\eta = 0.282$$
$$\theta = 0.580.$$

The final result of the experiment is represented by the number $\theta = .58$ which means that *a metal vapor arc between activated platinum electrodes dissipates 58 per cent of its energy upon the positive electrode and 42 per cent upon the negative electrode.* This result differs only slightly from $w_+/(w_+ + w_-) = 0.577$, the difference being the correction due to resistive heat developed equally in the two electrodes. For the inactive electrodes we obtain $w_+/(w_+ + w_-) = 0.572$ which is in close agreement, and it too must differ only slightly from the value which would be obtained if data were available for making the correction due to resistive heat.

## Reliability of Results

The fact that the experiments account for only $\eta = 0.765$ of the total energy need not be disturbing. It seems most likely that an inaccurate value for the thermal conductivity of the platinum wires and imperfect geometry of the wires account for this. The low resistance of the thermocouple circuits does not affect the indications of temperature difference which they give.

The large corrections of line 2 in columns 1 are highly uncertain as one sees readily by inspection of the corresponding curves of Fig. 3. If these corrections were taken to be zero one would obtain $\theta = 0.556$. It certainly seems unlikely that $\theta$ for inactive platinum electrodes can be less than this value.

## Interpretation

There are previous observations upon closure arcs between inactive electrodes which must be correlated with the results of these measurements There are no corresponding earlier data upon active electrodes.

A single closure arc between inactive platinum electrodes produces a pit on the positive electrode having a volume which is comparable with that to be expected if all of the energy of the arc is dissipated upon that electrode and is used there in melting and vaporizing metal with all of the

molten metal blown out from the arc crater by the pressure of metal vapor.[14] The metal from the crater is deposited in a rim about it and upon the negative electrode. There is considerable roughening of the negative electrode but none of its metal has been found upon the anode. This roughening indicates, no doubt, that a small fraction of the energy is dissipated directly upon the cathode. Estimates of the amount of metal transferred from positive to negative, made after many thousands of closure arcs, have shown that about 1 per cent of the metal from an anode crater reaches the cathode. Microscopic examination of the surfaces after a single arc reveals that the transferred metal upon the cathode seems to be much greater in amount than the 1 per cent found after many arcs. There is thus a distinct disagreement between the results of transfer measurements after many arcs and what one sees upon the surface of the cathode after a single arc.

Measurements of the present paper could be accounted for by assuming that most of the energy of a closure arc is dissipated upon the anode in melting and boiling metal, and that the energy is then located in this displaced metal with 58 per cent of it finally freezing on the anode and 42 per cent on the cathode. This tentative conclusion agrees with microscopic observations upon the electrodes after a single closure arc but is in sharp disagreement with the results of measurements of transfer of metal resulting from many arcs.

At the time this paper is being written it is felt that more penetrating experiments are called for, and in particular transfer measurements upon both active and inactive surfaces under experimental conditions which are better controlled than any which have been made previously.

[14] L. H. Germer and F. E. Haworth, *Phys. Rev. 73*, 1121 (1948) and Reference 11, Figs. 5 and 6.

# Detwinning Ferroelectric Crystals

By ELIZABETH A. WOOD

Unstrained single crystals of barium titanate can be detwinned under the influence of an electric field at elevated temperature, but strained crystals cannot. It seems probable that this is also true of crystals in a polycrystalline body such as a ceramic.

EACH of the ferroelectric[1] crystals so far discovered has a structure which closely approaches a more symmetrical structure into which it transforms at the Curie temperature. In all of them, the deviation from the more symmetrical structure is so slight (Table I) that the application of mechanical stress or electric field can produce a shift from one orientation of the lower symmetry structure to another. Since, in crystals grown from the melt, such as barium titanate, inhomogeneous mechanical stresses resulting from inhomogeneous cooling or differential thermal contraction of the surrounding flux material are present in the crystals as they pass through the Curie temperature, these crystals commonly comprise regions of two or more orientations of the lower-symmetry structure, symmetrically related. They are, in other words, twinned.

In this condition the electrically polar direction differs in orientation from one individual of the twin to another[2]. Since it is frequently desirable to have the polar direction oriented uniformly throughout the crystal, it is of interest to determine under what conditions this state can be achieved. It is not possible in all crystals.

The discussion in this paper will be confined to barium titanate because more experimental data are available for this crystal, but it is probable that similar considerations are applicable to the other ferroelectric crystals.

The process of causing the polar axis in a ferroelectric crystal to have the same orientation throughout the crystal has been called "poling." It is the process of detwinning the crystal. As C. J. Davisson and others pointed out in connection with the problem of detwinning quartz crystals during World War II, if the crystal is subjected to a stress which will be lessened if the "misoriented" regions change to the desired orientation and if the activation energy of the change is not too great, the crystal will be detwinned.

[1] Ferroelectric crystals are those crystals which exhibit, with respect to an electric field, most of the phenomena exhibited by ferromagnetic crystals with respect to a magnetic field, such as spontaneous polarization, domain structure, hysteresis of response to an alternating field and a Curie temperature above which these unusual characteristics are not present.

[2] By the ferromagnetic analogy each twin individual is called a ferroelectric "domain."

TABLE I

| Substance | Stable Structure at lower temperature | Closely allied more symmetrical structure and the Curie Temperature above which it is the stable structure | Difference |
|---|---|---|---|
| Barium titanate | Tetragonal $c = 4.04$, $a = 4.00$ | Cubic, $a_0 = 4.00$ $T_c$ = ca. 120°C | 1% of the length of the $c$ axis |
| Potassium niobate | Orthorhombic $a = 5.70$, $b = 5.74$, $c = 3.98$ equivalent to special case of monoclinic in which $a = c = 4.04$, $b = 3.98$, $\beta = 90°, 20' \pm 5'$ | Cubic, $a_0 = 4.04$ $T_c$ = ca. 435°C | 1.5% of the length of the $c$ axis + a shear angle of about 20′ |
| Rochelle Salt | Monoclinic | Orthorhombic $T_c$ = ca. 24°C | A shear angle of about 3′ |
| Potassium dihydrogen phosphate | Orthorhombic | Tetragonal $T_c$ = ca. −152°C | Small shear |

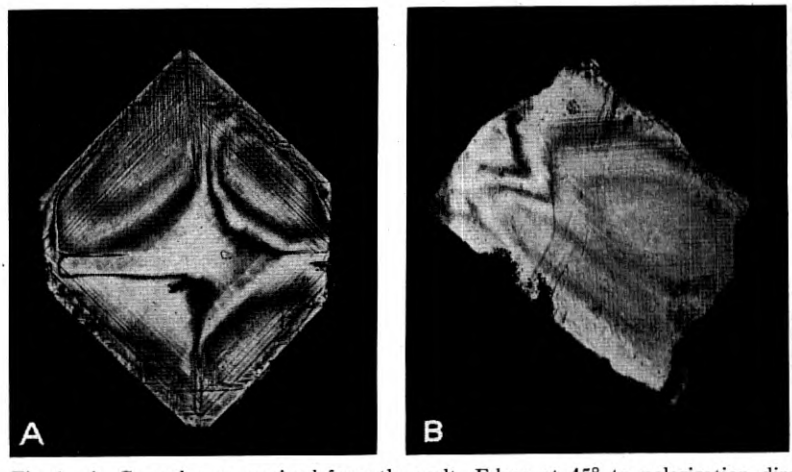



Fig. 1—A. Crystal $a$ as received from the melt. Edges at 45° to polarization directions of crossed nicols. Dimensions of surface: .2 x .2 mm.
B. Crystal $b$ as received from the melt. Edges at 45° to polarization directions of crossed nicols. Dimensions of surface: .15 x .2 mm.

With barium titanate as with quartz, the activation energy of this change can be reduced to zero by heating the crystal through a polymorphic transition above which its symmetry is such that the twinning can no longer exist. It is then cooled through the transition under the influence of the applied stress which favors one of the possible twin-orientations.

## Single-crystal Experiments

Parts A and B of Fig. 1 are photomicrographs of barium titanate crystals, both grown from the same melt by B. T. Matthias. The composition of the melt was 26 grams $BaCO_3$, 6.5 grams $TiO_2$, 50 grams $BaCl_2$, and the method followed was that described by Matthias in 1948[3]. Each of the crystals shows several domains and some inhomogeneous strain as indicated by birefringence evident between crossed nicols when the crystal is at the extinction position, Fig. 2, A and B, i.e. when its edges are parallel to the polarization directions in the polarizer and analyzer. An unstrained crystal in this position appears black between crossed nicols.

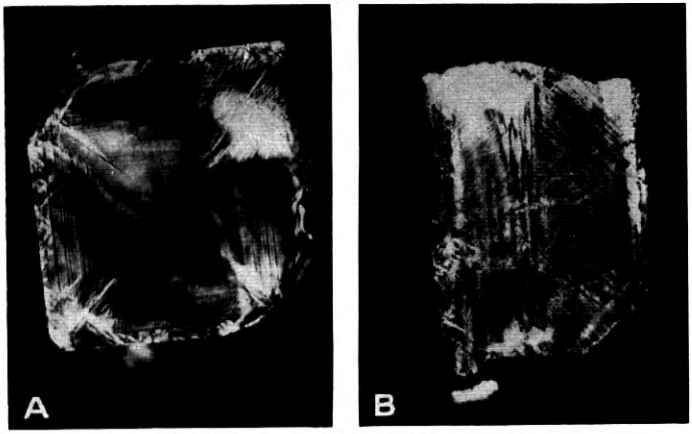



Fig 2—A. Same as Fig. 1A, but at extinction position.
B. Same as Fig. 1B, but at extinction position.

## Optical Evidence of the Effect of a High Field

However, crystal *a* can be made to assume essentially a single orientation throughout by the application of a high field at elevated temperature as shown in Fig. 3A, but the same treatment applied to crystal *b* results only in the formation of a large number of domains, as shown in Fig. 3B. A field of 16000 volts per cm. was applied across each crystal at 125°C. and continued until the crystal had cooled to less than 50°C. Parts A and B of Fig. 4 are the extinction-position photographs corresponding to Parts A and B of Fig. 3.

## X-ray Evidence of Inhomogeneous Strain

The reason for the difference in behavior of the two crystals is suggested by their back-reflection Laue photographs. Parts A and B of Fig. 5 are Laue

[3] Matthias, B. T., *Phys. Rev.* 73, 808–9, 1948.

photographs taken before the attempt was made to pole the crystals. Whereas a Laue photograph of a perfect single crystal would show a pattern of single spots, crystal *a* shows a pattern of groups of spots joined by

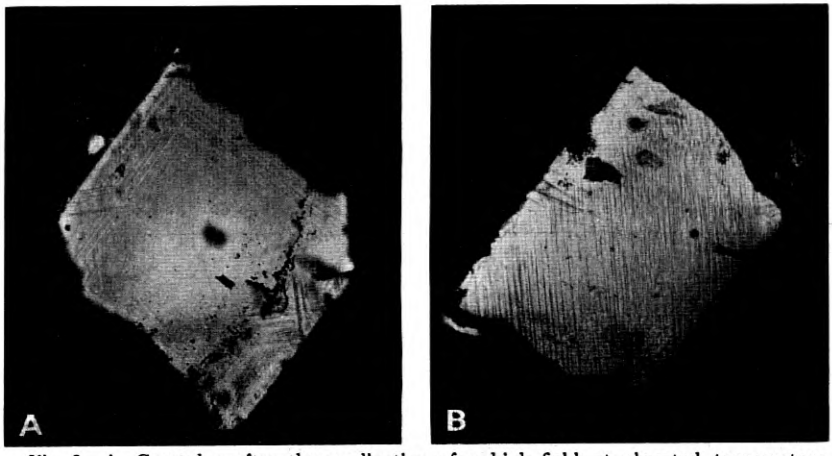



Fig. 3—A. Crystal *a* after the application of a high field at elevated temperature. Edges at 45° to polarization directions of crossed nicols
B. Crystal *b* after application of a high field at elevated temperature. Edges at 45° to polarization directions of crossed nicols.

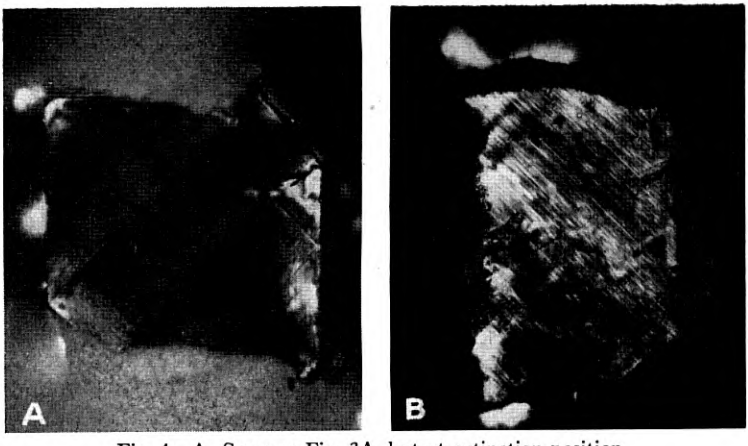



Fig. 4—A. Same as Fig. 3A, but at extinction position.
B. Same as Fig. 3B, but at extinction position.

fainter streaks and crystal *b* shows a pattern of short streaks. In both cases, the streaks indicate crystal material of continuously varying orientation, but in the case of crystal *a* it is transitional in orientation between two or

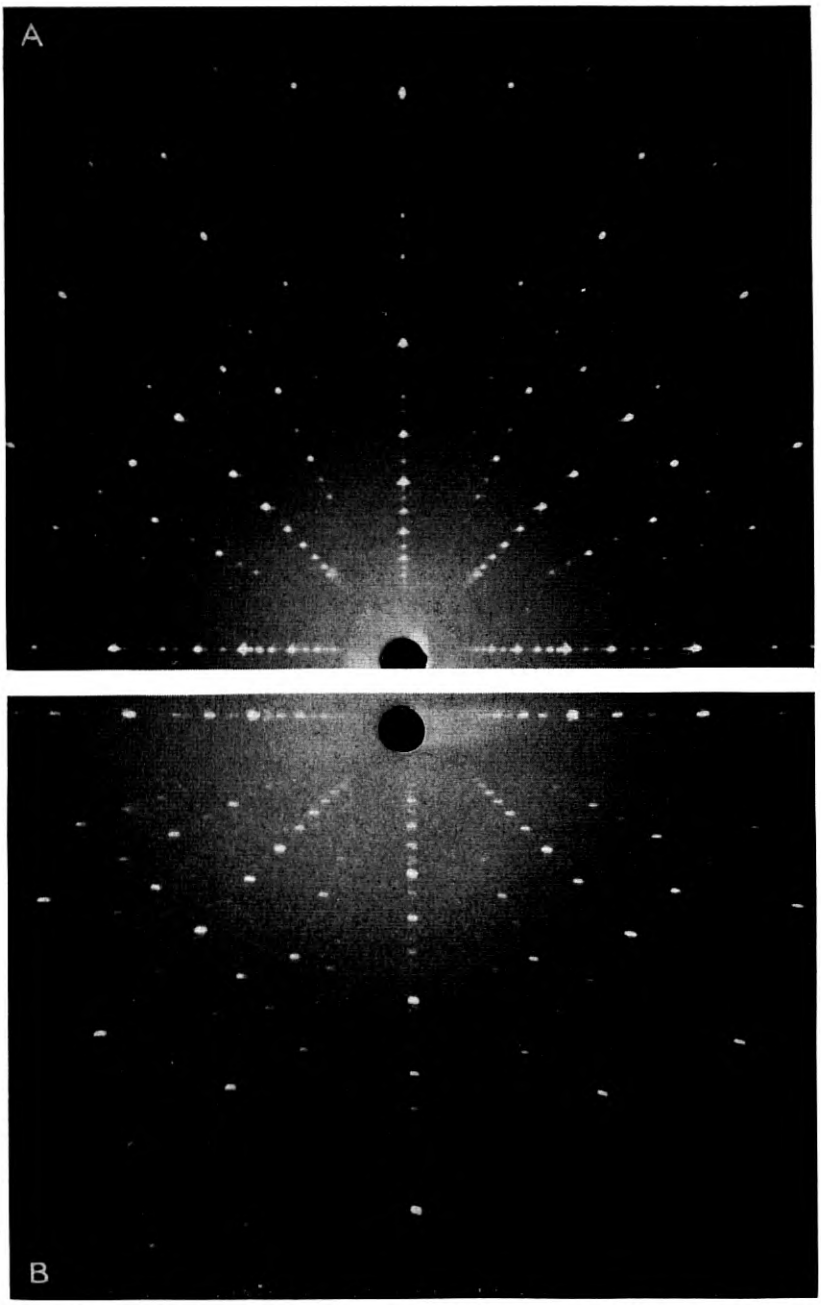



Fig. 5—Laue photographs of the two crystals before treatment. The symmetrical half of each photograph has been removed to facilitate close comparison. A: from crystal *a*; B: from crystal *b*.

more twin-related orientations and is probably twin-boundary material, whereas crystal *b* is a bent crystal.

The evidence for this interpretation lies in the facts that (1) the streaks in Fig. 5A converge at the reflections from the various (101) planes which are twin planes, whereas those in Fig. 5B do not; (2) the streaks in Fig. 5B show only that variation in length which is due to the use of a flat film whereas those in Fig. 5A show greater variation; and, finally, (3) the streaks in Fig. 5B are of nearly uniform intensity throughout, whereas those in Fig. 5A are faint streaks between strong end points. These three points are discussed in the following section.

## Distinction between Twin-Boundary and Other Inhomogeneous Strains in Crystals

The spots on a back-reflection Laue photograph may be considered as the intersections of the film with normals to atomic planes in the crystal, modified by a non-linear scale factor. The position of any spot is independent of the wave-length of the x-rays producing it and dependent only on the orientation of the reflecting plane.

When the x-ray beam falls on a twin boundary two families of twin-related spots appear on the film. In barium titanate twin-related spots from equivalent planes are close to each other. If the two spots of such a pair are joined by a line these lines will all converge toward the spot from the (101) plane which is the twin plane, the plane across which reflection of the structure would produce the twin configuration. (See Fig. 6, a back-reflection Laue photograph of a barium titanate crystal with only 2 twin-related orientations.) That this must be so will be clear from Fig. 7. With the exception of the twinning plane the planes in this figure represent zonal planes, planes containing two or more atomic-plane normals. The zonal planes on the two sides of the twin plane represent the zonal plane orientations in the two parts of the twin. The only zonal plane directions common to both parts of the twin are those normal to the twin plane since these are the only directions not changed by reflection across the twinning plane. The one direction common to all these unchanged zonal planes is the normal to the twin plane. Thus the zonal arcs on the plane photograph which are common to spots from both parts of the twin intersect in the reflection from the twin plane.

Referring now to Parts A and B of Fig. 5, we see that the streaks in Fig. 5A lie along the zonal arcs common to both parts of any given twin pair and are intermediate between the spots of the twin pair. They are therefore reflections from material transitional in orientation between the two twin orientations. The streaks in Fig. 5B, however, do not converge toward a

(101) plane-normal, but rather, when the non-linear scale factor has been taken into account, are all normal to the (100) axis which lies parallel to the film in a top-to-bottom direction. They therefore come from crystal planes bent around this axis of a twinned barium titanate crystal.

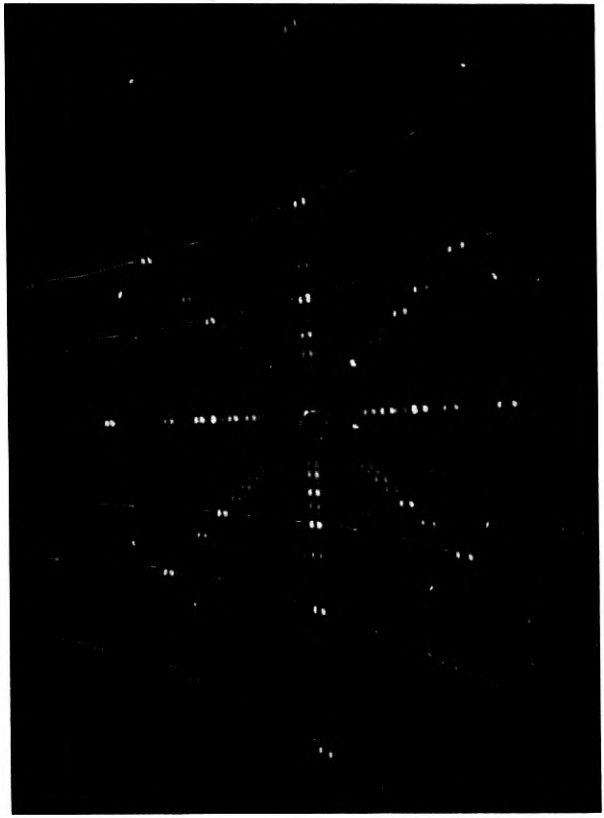


Fig. 6—Laue photograph of a barium titanate crystal with only two twin-related orientations.

A section through the reciprocal lattice of a twinned barium titanate crystal is shown in Fig. 8A: that of a bent crystal, in 8B. In the reciprocal lattice of a tetragonal crystal the direction of each point from the origin is the same as the direction of the normal to the set of planes it represents and the distance of each point from the origin is proportional to the reciprocal of their interplanar spacing. Since the back reflection Laue photograph shows only orientations of the atomic planes it may be thought of as a

shadowgraph of the reciprocal lattice, illuminated by a point source of light at the origin, as indicated by the dashed lines in Figs. 8A and 8B.

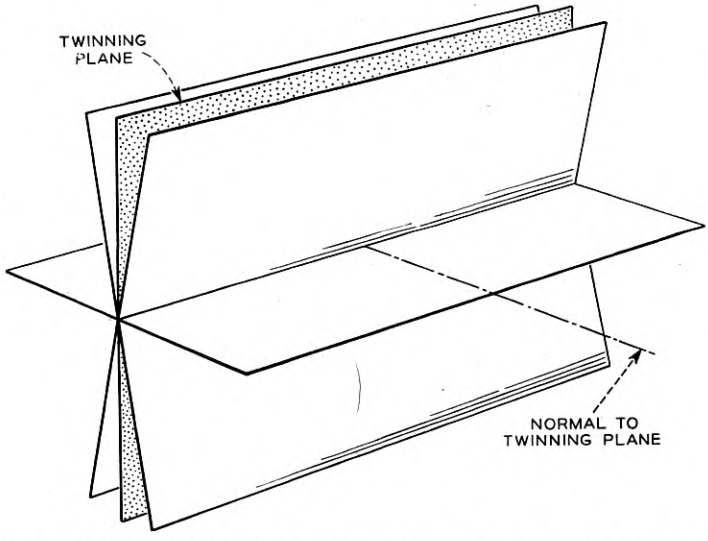



Fig. 7—Diagram of zonal relations between the two parts of a twinned crystal.

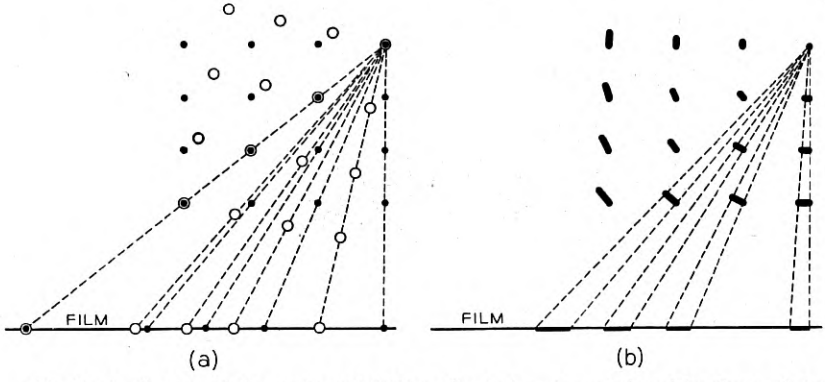



Fig. 8—A. Section of the reciprocal lattice of a twinned crystal and its Laue photograph.
B. Section of the reciprocal lattice of a strained crystal and its Laue photograph.

From these figures the second point of difference between Laue photographs 5A and 5B becomes clear, namely, that in the case of the bent crystal viewed normal to the bending axis the streaks appear rather uniform in length, whereas the streaks from the inter-twin oriented material

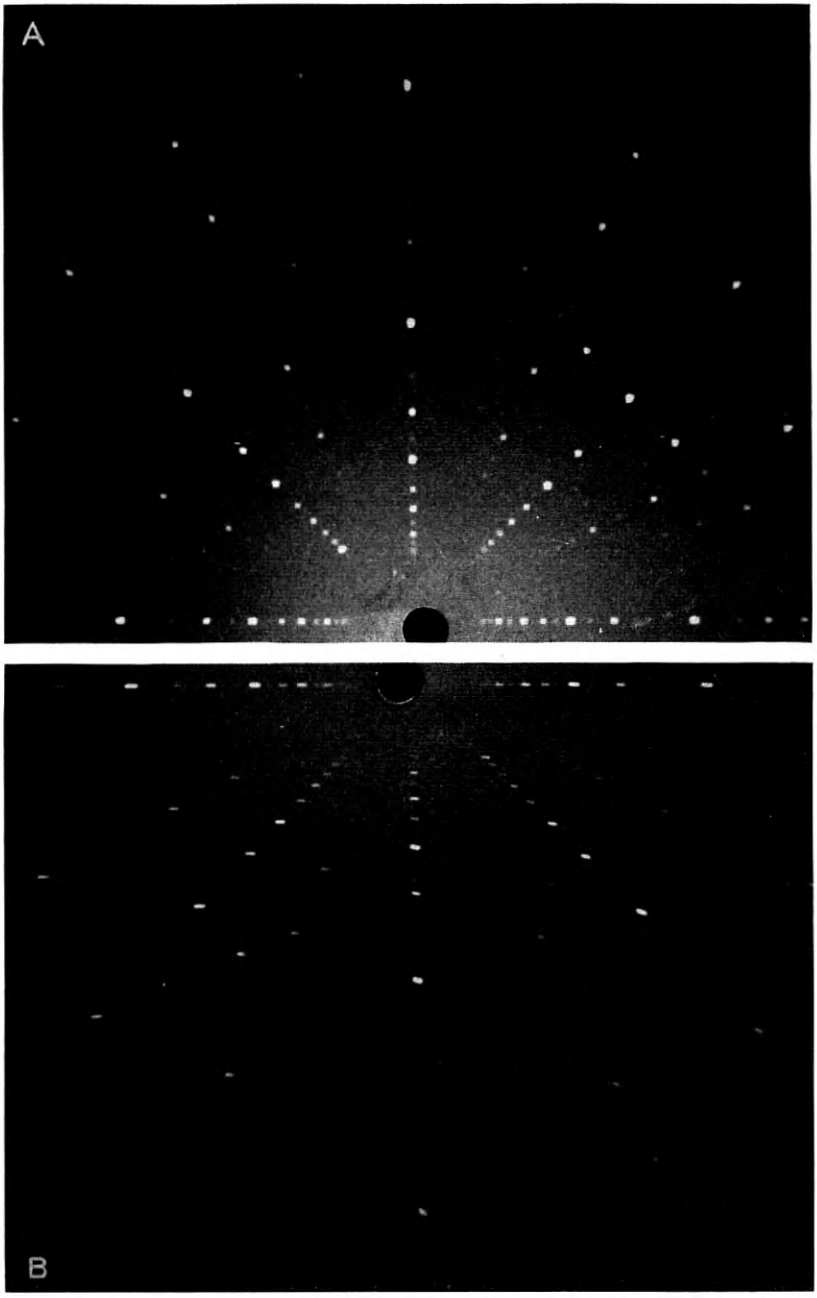



Fig. 9—Laue photographs of the two crystals after treatment, paired as in Fig. 5. A: from crystal *a*; B: from crystal *b*.

diminish in length as they approach the twin-plane normal. This is perhaps more obvious in Fig. 6 where only one pair of twins is shown.

Finally, the streaks from bent crystals are more uniform in intensity than those with intertwin-oriented material, but if the bending were non-uniform or the intertwin material more abundant this might not be so.

### X-ray Evidence of Effect of High Field

The Laue photograph of crystal *a*, taken after the poling process had caused it to become a single untwinned crystal, Figs. 3A and 4A, is shown in Fig. 9A. It shows a pattern of single spots. The absence of streaks agrees with the paucity of birefringent regions in Fig. 4A in indicating very little inhomogeneous strain in the crystal.

The Laue photograph of crystal *b*, taken after the poling attempt had produced only a regular multiple-twin pattern in it, Figs. 3B and 4B, is shown in Fig. 9B. The inhomogeneous strain has not disappeared, as indicated also by the birefringent regions in the photograph of Fig. 4B.

### Summary of Results of Single-Crystal Experiments

From the experiments described above it is concluded that barium-titanate crystals with only twin-boundary strain can, under the influence of a high field at elevated temperatures, be caused to have a single crystallographic orientation whereas barium titanate crystals otherwise strained cannot.

### Application to Barium Titanate Ceramics

Single crystals of barium titanate large enough for practical applications have not yet been grown. Therefore all practical applications using barium titanate have so far used it in the ceramic form.

Ceramics have been made for two different applications: condensers and electromechanical transducers. For the first, the maximum electrical polarizability for a given applied electric field is desired, since this results in a high dielectric constant. For the second application, however, it is desirable to have a ceramic which will deform mechanically in an electric field according to its own polarity. With this end in view, ceramics intended for electromechanical transducers have been poled by being subjected to a high field (roughly 15000 v/cm.) as they were cooled through the Curie temperature to room temperature.

In a series of unpublished experiments W. P. Mason and R. F. Wick of the Bell Laboratories have found that certain barium titanate ceramics, when poled in this way, retain their polarization in spite of high reverse fields ($\frac{1}{6}$ to $\frac{1}{3}$ the poling fields), i.e. require a high coercive force to change

the direction of their polarization. Such ceramics can be used in piezoelectric devices with high alternating fields without "depolarization" and can therefore achieve electro-mechanical coupling at higher power levels than ceramics that do not retain their polarization under the influence of a reverse field. Only a small proportion of ceramic specimens could be poled in this way and the factor common to these has not yet been ascertained.

In the light of the single-crystal experiments reported in this paper, it seems apparent that ceramics composed of inhomogeneously strained crystals (excluding twin-boundary strain) could not be poled. Three ceramic specimens whose poling history was known were available. Of these only one could be poled. X-ray diffraction photographs of the two unpolable ceramics showed streaked reflections from the individual grains, indicating strain. The grain-size of the polable specimen was much smaller, so small that reflections from individual grains could not be identified.

It is anticipated that ceramics for different uses should be differently fired and perhaps even differently composed as well as subjected to different electrical treatment subsequent to their formation.

# Longitudinal Modes of Elastic Waves in Isotropic Cylinders and Slabs

By A. N. HOLDEN

The general properties of the longitudinal modes in cylinders and slabs are developed with the aid of the close formal analogy between the dispersion equations for the two cases.

## 1. Introduction

THE classical exact treatments of the modes of propagation of elastic waves in isotropic media having stress-free surfaces but extending indefinitely in at least one dimension are those of Rayleigh[1] for semi-infinite media bounded by one plane, of Lamb[2] for slabs bounded by two parallel planes, and of Pochhammer[3] for solid cylinders. Rayleigh showed that a wave could be propagated without attenuation parallel to the surface, in which the displacement amplitude of the medium decreased exponentially with distance from the surface, at a velocity independent of frequency and somewhat lower than that of either the plane longitudinal or plane transverse waves in the infinite medium. Such "Rayleigh surface waves" have received application in earthquake theory.

For slabs or cylinders the treatments lead to a transcendental secular equation, establishing a relation (the "geometrical dispersion") between the frequency and the phase velocity, which for some time received only asymptotic application in justifying simpler approximate treatments. The past decade, however, has seen a revival of interest in the exact results[4, 5] stimulated by experimental application of ultrasonic techniques to rods[6, 9] and slabs,[7] by the use of rods and the like as acoustic transmission media, and perhaps by curiosity as to what qualitative correspondence may exist between such waves and the more intensively studied electromagnetic waves in wave guides. That this correspondence might not be close could be anticipated by observing that an attempt to build up modes by the superposition of plane waves in the medium reflected from boundaries would encounter an essential difference between the two cases: the elastic medium supports plane waves of two types (longitudinal and transverse) with different velocities, and reflection from a boundary transforms a wave of either type into a mixture of both.

On grounds both formal and physical it may be expected that solutions to the equations of small motion of the medium with a stress-free cylindrical boundary can be found with any integral number of diametral nodes of the

component of displacement along the rod, as well as for the "torsional" modes in which there is no displacement along the rod whatever. The classical results are for no such nodes, the "longitudinal" (or "elongational") modes, and for one such node, the "flexural" modes. The secular equation for modes with any number of such nodes has been exhibited by Hudson.[5] For any one of these types of mode, it may be expected that the secular equation will define a many-branched relation* between frequency and phase velocity, and that a different number of interior cylindrical nodal surfaces for the displacement components might be associated with each branch. Apart from the relatively simple torsional modes, the only branches whose properties have been intensively studied are the lowest branch of the longitudinal[4] and the lowest of the flexural[5] modes, because they (and the lowest torsional branch) are the only ones extending to zero frequency, the others exhibiting "cut-off" frequencies at which their phase velocities become infinite and below which they are rapidly attenuated as they progress through the medium.

Three qualitative results of these studies are of especial interest. In the first place, with increasing frequency the phase velocity in the lowest longitudinal and flexural branches approaches the velocity of the Rayleigh surface wave, and the disturbance becomes increasingly confined to the surface of the cylinder. In the second place, the dispersion is not monotonic as it is in the electromagnetic case: the phase velocity exhibits a minimum in the lowest longitudinal branch[4] and a maximum in the lowest flexural branch[5] with varying frequency. Finally, in the lowest longitudinal branch at least, the cylindrical nodes of the displacement components vary not only in radius but even in number with the frequency.[4]

The last result suggests that it would be difficult in practice to drive a cylindrical rod in that pure mode represented by its lowest longitudinal branch over any extended frequency range, since it is difficult to visualize a driving mechanism having suitable nodal properties. Longitudinal drivers which can be readily constructed may be expected to deliver energy to all longitudinal branches, in proportions varying with frequency. How satisfactory such a transmission device could be would depend importantly on how much the phase velocities at any one frequency differed from branch to branch.

This paper sketches the behavior of the higher longitudinal branches. That behavior could, of course, be determined exactly; Hudson[5] has shown how the calculation of the roots of the secular equations can be facilitated,

* This is true in particular of the flexural type of mode, and in his otherwise excellent treatment of flexure Hudson's statement to the contrary must be disregarded. Recent writings in this field have tended to distinguish as "branches" what in allied problems are commonly called "modes".

and Hueter[9] has used graphical methods. The alternative adopted here is a semi-quantitative treatment, assisted by extensive reference to the behavior of longitudinal waves in slabs,* for which the secular equation is simpler and closely analogous. The analogy in the case of flexural modes is considerably less close and will not be discussed.

The general consequences of the inquiry are that the higher longitudinal branches have phase velocities which are not necessarily monotonic functions of frequency. With increasing frequency, however, those velocities all approach that of the plane transverse wave,** not that of the Rayleigh surface wave (nor that of the plane longitudinal wave, as some investigators had guessed), a fact reflected perhaps in the experimental observation that driving a rod transversely usually provides purer transmission than driving it longitudinally.† Variation of nodal cylinders in location and number with frequency persists in the higher branches.

## 2. The Slab

The slab extends to infinity in the $y, z$ plane and has a thickness $2a$ in the $x$-direction. The displacements of its parts in the $x$, $y$, $z$ directions are $u$, $v$, $w$. Its material has density $\rho$ and Lamé elastic constants $\lambda$ and $\mu$, so that its longitudinal wave velocity is $\sqrt{(2\mu+\lambda)/\rho}$ and its transverse wave velocity is $\sqrt{\mu/\rho}$. That $\mu$ should be positive is a stability requirement of energetics; $\lambda$ will also be taken as positive since no material with negative $\lambda$ is known.

The equations of small motion are, in vector form,

$$(2\mu + \lambda)\ \text{grad div}\ (u, v, w) - \mu\ \text{curl curl}\ (u, v, w) = \rho \frac{\partial^2}{\partial t^2}(u, v, w).$$

Solutions representing longitudinal waves propagated in the $z$ direction can be of the form

$$u = Ue^{i(\omega t+\gamma z)}, \qquad v = 0, \qquad w = We^{i(\omega t+\gamma z)},$$

where $U$ is an odd function, and $W$ an even function, of $x$ alone, $\omega$ is the frequency in radians per second, and $\gamma = \omega/c$ where $c$ is the phase velocity. Solutions independent of $y$ are chosen here because they provide the simplest analogues to the case of the cylinder. Substitution shows that $U = Ae^{ikx}$,

* I am indebted to Dr. W. Shockley for the suggestion that this behavior might display a close enough analogy to that of the cylinder to provide insight; the work of Morse bears out the analogy.

** The fact is noted by Bancroft.

† Private communication from H. J. McSkimin.

$W = Be^{ikz}$, is a solution (where $A$ and $B$ are constants measuring the amplitude), if either

$$\text{(i)}\quad k_1^2 = \frac{\rho\omega^2}{2\mu + \lambda} - \gamma^2 \quad \text{and} \quad \gamma A_1 = k_1 B_1,$$

or

$$\text{(ii)}\quad k_2^2 = \frac{\rho\omega^2}{\mu} - \gamma^2 \quad \text{and} \quad k_2 A_2 = -\gamma B_2.$$

When solutions of both types are so superposed as to make $U$ odd and $W$ even

$$U = iA_1 \sin k_1 x + iA_2 \sin k_2 x, \tag{1}$$

$$W = A_1 \frac{\gamma}{k_1} \cos k_1 x - A_2 \frac{k_2}{\gamma} \cos k_2 x. \tag{2}$$

The normal and tangential stresses on planes perpendicular to $x$ are

$$X_x = (2\mu + \lambda)\frac{\partial u}{\partial x} + \lambda\left(\frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right), \qquad X_y = \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right),$$

$$X_z = \mu\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right)$$

and the requirement that they vanish at $x = \pm a$ leads to the boundary conditions

$$A_1(\lambda\gamma^2 + (2\mu + \lambda)k_1^2)\cos k_1 a + 2A_2\mu k_1 k_2 \cos k_2 a = 0,$$

$$2A_1\gamma^2 \sin k_1 a + A_2(\gamma^2 - k_2^2)\sin k_2 a = 0,$$

the vanishing of whose eliminant with regard to $A_1$ and $A_2$ is the secular equation. Although in principle that equation establishes a relation between $\gamma$ and $\omega$, it is more conveniently examined when expressed in terms of $\alpha \equiv k_1 a$ and $\beta \equiv k_2 a$, which are quadratically related to $\omega$ and $\gamma$ by (i) and (ii). In those terms it becomes

$$(\lambda\beta^2 + (2\mu + \lambda)\alpha^2)^2 \cos\alpha \sin\beta + 4(\mu + \lambda)\alpha\beta(\mu\beta^2 - (2\mu + \lambda)\alpha^2)\sin\alpha\cos\beta = 0. \tag{3}$$

The physically interesting quantities can be expressed in terms of $\alpha$ and $\beta$, with the aid of (i) and (ii). Thus

$$\rho\omega^2 = \frac{\mu(2\mu + \lambda)}{a^2(\mu + \lambda)}(\beta^2 - \alpha^2), \quad \text{and} \tag{4}$$

$$\gamma^2 = \frac{\mu\beta^2 - (2\mu + \lambda)\alpha^2}{a^2(\mu + \lambda)}. \tag{5}$$

Hence, denoting $l \equiv \beta/\alpha$, the phase velocity $c \equiv \omega/\gamma$ is given by

$$\rho c^2 = \frac{\mu(2\mu + \lambda)(l^2 - 1)}{\mu l^2 - (2\mu + \lambda)} \equiv E, \tag{6}$$

where $E$ is an "effective stiffness", a function of the elastic constants and $l$.

Since $\beta = 0$ is a trivial root of equation (3), it can be divided by $\beta$, and the expression on the left then becomes even in both $\alpha$ and $\beta$ and (3) can be regarded as an equation in $\alpha^2$ and $\beta^2$. From (4) and (5) it is evident that, if $\omega$ and $\gamma$ are both to be real, $\alpha^2$ and $\beta^2$ must be real and must obey the inequalities

$$\beta^2 > \alpha^2, \qquad \mu\beta^2 > (2\mu + \lambda)\alpha^2, \tag{7}$$

and thus the root $\beta = \alpha$ can be neglected. The general character of the desired roots can consequently be exhibited on a plot of $\beta^2$ against $\alpha^2$. Evidently on that plot lines of slope unity are lines of constant frequency (equation 4), and lines radiating from the origin are lines of constant velocity (equation 6). As will appear later, however, it is more convenient to use a linear rather than a quadratic plot, real $\alpha$ being measured to the right, imaginary $\alpha$ to the left, of the vertical axis, and real $\beta$ upward, imaginary $\beta$ downward, from the horizontal axis. Here radial lines are still lines of constant velocity, but lines of constant frequency are no longer simple.

In Fig. 2 such a plot has been sketched for the first few modes of a material obeying the Cauchy condition $\lambda = \mu$; the properties shown are restricted to those derived in the following paragraphs, and are lettered in Fig. 1 to correspond with those paragraphs.

(a) By virtue of (7), the significant portions of the roots lie above and to the left of the lines $\beta^2 = \alpha^2$, $\mu\beta^2 = (2\mu + \lambda)\alpha^2$. Setting $\mu\beta^2 = (2\mu + \lambda)\alpha^2$ in (3) reveals the cut-offs at $\sin\beta = 0$ and at $\cos\alpha = 0$: in other words at $\beta^2 = n^2\pi^2$, $\alpha^2 = \frac{\mu}{2\mu + \lambda} n^2\pi^2$, and also at $\beta^2 = \frac{2\mu + \lambda}{\mu}\left(n + \frac{1}{2}\right)^2 \pi^2$, $\alpha^2 = \left(n + \frac{1}{2}\right)^2 \pi^2$, where $n$ is any integer.

(b) Setting $\alpha = 0$ in (3), it can be seen that the roots intersect the line $\alpha^2 = 0$ at the points $\sin\beta = 0$. By calculating the derivative of $\beta^2$ with respect to $\alpha^2$, those points (at which $\alpha$ changes from pure real to pure imaginary) can be shown not to be multiple points, and the branches to have $\frac{d\beta}{d\alpha} = 0$ and $\frac{d(\beta^2)}{d(\alpha^2)} = -\frac{8\mu(\mu + \lambda)}{\lambda^2}$, independent of branch number and negative for

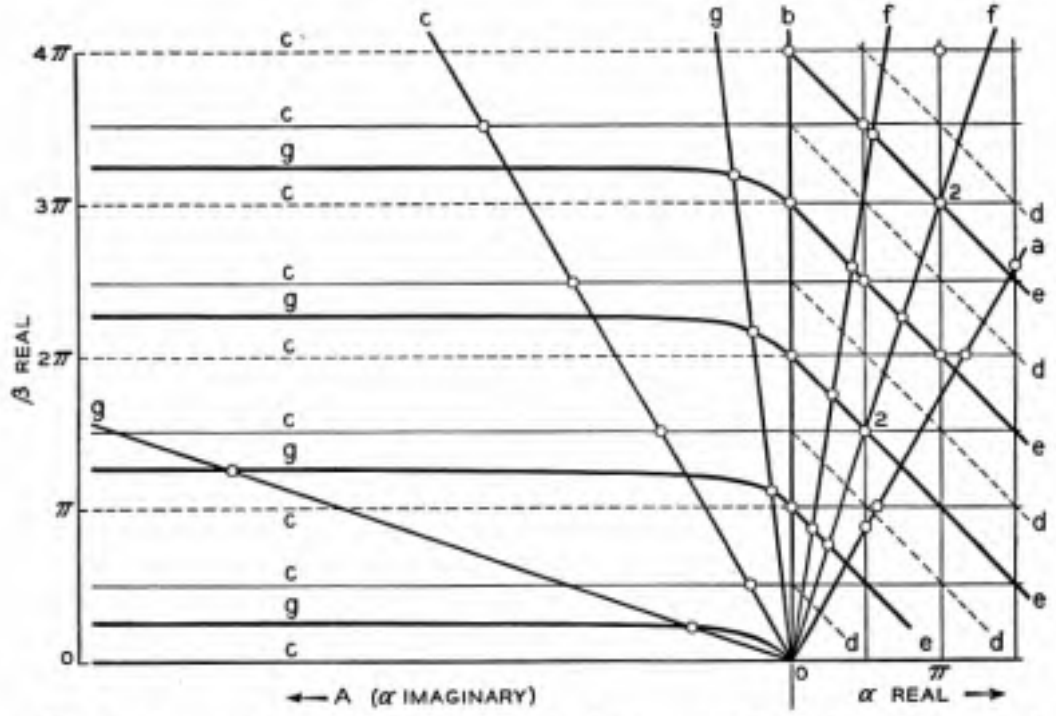



Fig. 1—Lines and intersections, discussed in the correspondingly lettered paragraphs of the text, which determine the properties of the first five branches of the longitudinal modes for a material obeying the Cauchy condition. Two coincident pairs of points are marked (2).

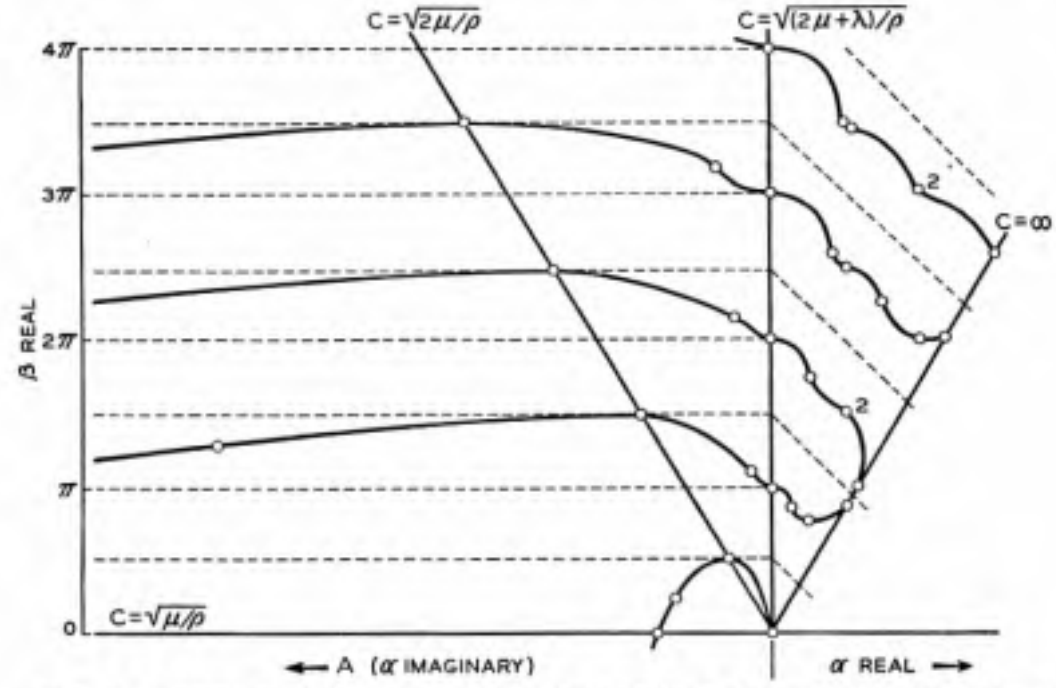



Fig. 2—A rough sketch of the branches determined by the properties illustrated in Fig. 1.

all materials. At those points the phase velocity is that of a plane longitudinal wave.

(c) Setting $\sin \beta = 0$ reduces (3) to $\alpha\beta(\mu\beta^2 - (2\mu + \lambda)\alpha^2) \sin \alpha = 0$, and hence in the region of positive $\beta^2$ and negative $\alpha^2$ the roots do not intersect the horizontal lines $\beta^2 = n^2\pi^2$ $(n \neq 0)$. As can be seen from (6) this confinement implies that the velocity is asymptotic to $\sqrt{\mu/\rho}$, that of a plane transverse wave, with increasing frequency. Notice that, in the region of positive $\beta^2$ and negative $\alpha^2$, $\beta$ takes the values $\cos \beta = 0$ only where $\lambda\beta^2 + (2\mu + \lambda)\alpha^2 = 0$ and that the roots have zero slope there. At those points the waves have a phase velocity $\sqrt{2}$ times that of a plane transverse wave.

(d) In the region of positive $\beta^2$ and positive $\alpha^2$ the roots exhibit a somewhat more complicated behavior, but confining lines can again be found: the diagonal lines $\beta = (n + \frac{1}{2})\pi - \alpha$. It is the nature of such critical lines as these which can be better exhibited on the linear than on the quadratic plot. Alternatively those lines can be written $\cos \alpha \cos \beta - \sin \alpha \sin \beta = 0$ (and thus will be shown to have analogues in the case of the cylinder), and substitution of this expression into (3) shows that if the roots intersect these lines they must do so for values of $\alpha$ and $\beta$ satisfying the relation

$$4(\mu + \lambda)\alpha\beta(\mu\beta^2 - (2\mu + \lambda)\alpha^2) = -(\lambda\beta^2 + (2\mu + \lambda)\alpha^2)^2 \cot^2 \alpha.$$

But the inequalities (7) make such values impossible.

(e) This suggests that in that region the roots may oscillate in a somewhat irregular manner about the diagonal lines $\beta = n\pi - \alpha$. Indeed it is immediately evident that they pass through the points $\cos \beta = \cos \alpha = 0$ and $\sin \beta = \sin \alpha = 0$.

(f) Expressing those lines as $\sin \alpha \cos \beta + \cos \alpha \sin \beta = 0$, and substituting into (3), shows that additional intersections may be afforded by any roots of the quartic equation

$$(\lambda\beta^2 + (2\mu + \lambda)\alpha^2)^2 - 4(\mu + \lambda)\alpha\beta(\mu\beta^2 - (2\mu + \lambda)\alpha^2) = 0$$

which obey the inequalities (7). Discarding the root $\alpha + \beta = 0$, and dividing by $\alpha^3$, yields the cubic equation

$$\lambda^2 l^3 - (2\mu + \lambda)^2 l^2 + (2\mu + \lambda)(2\mu + 3\lambda)l + (2\mu + \lambda)^2 = 0, \qquad (8)$$

whose roots are the negatives of the roots of the cubic equation for the Rayleigh surface wave velocity. It is well known that the Rayleigh cubic always yields one and only one significant positive root, and hence equation (8) can afford at most two additional significant intersections of any root of the secular equation with the line about which it oscillates. Although it

is not feasible to exhibit the roots of the Rayleigh cubic explicitly for arbitrary $\mu$ and $\lambda$, it is of some interest to exhibit its discriminant

$$D = \frac{4}{27}\lambda^4(2\mu + \lambda)^3(\mu + \lambda)^2(11\lambda^3 + 4\lambda^2\mu - 9\lambda\mu^2 - 10\mu^3),$$

and to note that for real positive values of $\lambda$ and $\mu$ it changes sign only once, at approximately $\lambda/\mu = 10/9$. Hence for $\lambda/\mu > 10/9$ two roots of the Rayleigh cubic are complex, while for $\lambda/\mu < 10/9$ two roots are real and negative. For a material obeying the Cauchy condition $\lambda/\mu = 1$, the roots of the Rayleigh cubic are $-3$, $-3 \pm 2\sqrt{3}$; thus $l = 3, 3 + 2\sqrt{3}$, both of which obey the inequalities (7), are relevant to intersections of each branch of the roots of the secular equation with the line about which it oscillates.

(g) The results of (e) and (f) suggest the value of a similar investigation in the region of imaginary $\alpha$. Here (denoting $\alpha \equiv iA$ and $L \equiv \beta/A$ where $A$ is taken positive and real) intersections occur between the branches and the lines $\sinh A \cos\beta - \cosh A \sin\beta = 0$ when $(\lambda L^2 - (2\mu + \lambda))^2 = 4(\mu + \lambda)L(\mu L^2 + (2\mu + \lambda))$. Clearly this quartic in $L$ has two and only two positive real roots, one greater and one less than $\sqrt{(2\mu + \lambda)/\lambda}$. In the case $\lambda = \mu$, those roots are approximately 9 and 1/3. This information, taken with that of (c), establishes that the branches are confined in the region of imaginary $\alpha$ to bands determined by $n\pi < \beta < (n + \frac{1}{2})\pi$, having one tangency to the lines $\beta = (n + \frac{1}{2})\pi$; and that at values of $A$ greater than correspond to the smaller root of $l$, the branches lie in the bands $n\pi < \beta < (n + \frac{1}{4})\pi$.

It is convenient to obtain assurance that in general the branches do not intersect at any point by noting that the confining lines of paragraphs (c) and (d) define bands within each of which in general one and only one cut-off point falls. Pivoting a ruler about the origin of Fig. 1, and recalling the cut-off conditions, avails. Degenerate cases arise when the elastic constants satisfy a condition $2\mu + \lambda = n^2\mu$ where $n$ is an integer; in those cases some cut-off points coincide in pairs on some confining lines. Calculation of derivatives at those points shows that the cases are not otherwise exceptional: the pair of roots forms a continuous curve which is tangent to the cut-off line at the double cut-off point.

From (6) it follows that the phase velocity will have a maximum or a minimum with frequency if $\frac{d(\beta^2)}{d(\alpha^2)} = \frac{\beta^2}{\alpha^2}$. That condition requires

$$\tan^2\alpha = -\frac{(\lambda\beta^2 + (2\mu + \lambda)\alpha^2)^2(\lambda^2\beta^2 + (2\mu + \lambda)(2\mu + 3\lambda)\alpha^2)}{4(\mu + \lambda)(2\mu + \lambda)^2\alpha^2(\beta^2 - \alpha^2)(\mu\beta^2 - (2\mu + \lambda)\alpha^2)}.$$

In the region of positive $\beta^2$ and $\alpha^2$ and of the inequalities (7), this is impossible, but when $\alpha^2$ is negative the condition may be satisfied. If it is satisfied in the higher branches, however, it must be satisfied an even number of times in any branch, so that the branch exhibits as many maxima as

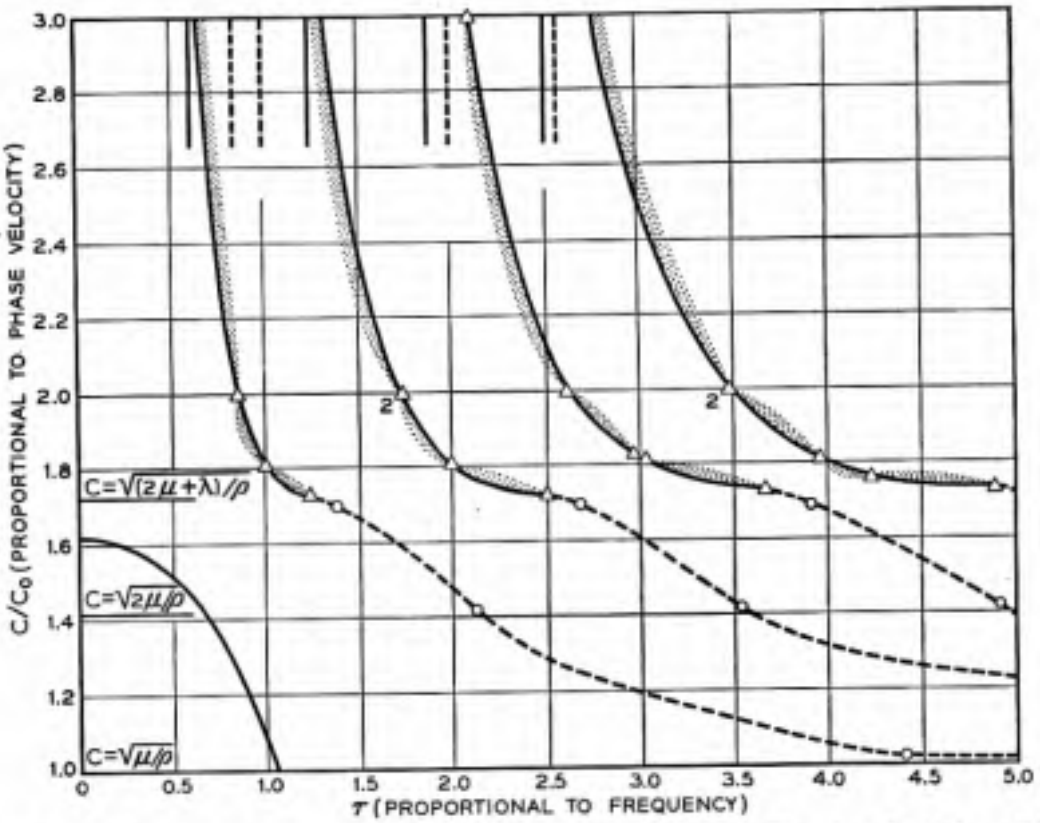



Fig. 3—The beginnings of the dispersion curves inferred from Figs. 1 and 2. The solid shaded lines are the curves about which the branches oscillate, intersecting them at the triangles and lying on the shaded side of them elsewhere, and the dashed extensions are the branches themselves. For increasing $\tau$ these branches all become asymptotic to the base line. The dashed lines at the top are the true cut-off frequencies; the solid "cut-off" lines are the asymptotes of the shaded curves. The beginning of the lowest branch is shown at the lower left; it becomes asymptotic to a line below this plot,

$$\frac{c}{c_0} = \sqrt{2 - \frac{2}{\sqrt{3}}},$$

after passing through a shallow minimum.

minima, for clearly the phase velocity is a decreasing function of frequency near the cut-off, and the velocity can also be shown to approach its asymptotic value at high frequencies from above in the higher branches.

In finally displaying the dispersion curves (Fig. 3) it is convenient to use as reduced variables $\tau$, the number of plane transverse wave lengths in

one slab thickness (which is proportional to the frequency), and $\frac{c}{c_0}$, the ratio of the velocity to that of a plane transverse wave. Evidently

$$\tau^2 \equiv \left(\frac{\omega a}{\pi c_0}\right)^2 = \frac{1}{\pi^2} \cdot \frac{2\mu + \lambda}{\mu + \lambda} (\beta^2 - \alpha^2); \qquad \left(\frac{c}{c_0}\right)^2 = \frac{(2\mu + \lambda)(l^2 - 1)}{\mu l^2 - (2\mu + \lambda)},$$

where $c_0 \equiv \sqrt{\mu/\rho}$.

For completeness, the lowest branch will be briefly sketched: that which originates at $\alpha = \beta = 0$. A calculation of $\frac{d(\beta^2)}{d(\alpha^2)}$ at that point yields only one non-trivial root, $-(2\mu + \lambda)(2\mu + 3\lambda)/\lambda^2$, and thus the phase velocity at low frequencies is found to correspond, as would be expected, with that given by the stiffness (a semi-Young's modulus, so to speak) of a material displacement-free in the $x$-direction but not in the $y$-direction, $E = 4\mu(\mu + \lambda)/(2\mu + \lambda)$. Since lines radiating from the origin of the $(\beta^2, \alpha^2)$ plot are lines of constant velocity, the dispersion curve for this branch starts with zero slope. The root curves over, intersecting the line $\lambda\beta^2 + (2\mu + \lambda)\alpha^2 = 0$ at $\beta = \frac{\pi}{2}$, and intersects the line $\beta^2 = 0$ again at $A^2$ $(\alpha = iA)$ where $(2\mu + \lambda)\, A \cosh A = 4(\mu + \lambda) \sinh A$. For large negative $\alpha^2$ and $\beta^2$, equation (3) approaches

$$\lambda^2 l^4 + 4\mu(\mu + \lambda) l^3 + 2\lambda(2\mu + \lambda) l^2 - 4(2\mu + \lambda)(\mu + \lambda) l + (2\mu + \lambda)^2 = 0,$$

which after discarding the trivial root $l = 1$ leaves the Rayleigh cubic. In the case of this one branch, the phase velocity approaches its asymptotic value at high frequencies from below, and hence the dispersion curve must have an odd number of maxima and minima, and in particular at least one minimum, as was discovered by numerical calculation for the corresponding branch in the case of the cylinder by Bancroft.[4]

The complicated behavior of the displacements in the higher branches is sufficiently illustrated by a brief consideration of their nodes: the values of $x$ at which the displacement is entirely along or entirely across the slab. From (1) and the boundary conditions, the $x$-dependence $U$ of the displacement component perpendicular to the slab will be given by

$$K_1 U = (\lambda\beta^2 + (2\mu + \lambda)\alpha^2) \sin\beta \sin\frac{\alpha x}{a} + 2(\mu\beta^2 - (2\mu + \lambda)\alpha^2)\sin\alpha \sin\frac{\beta x}{a} \tag{1a}$$

or by

$$K_2 U = 2(\mu + \lambda)\alpha\beta \cos\beta \sin\frac{\alpha x}{a} - (\lambda\beta^2 + (2\mu + \lambda)\alpha^2) \cos\alpha \sin\frac{\beta x}{a} \tag{1b}$$

where $K_1 = (\lambda\beta^2 + (2\mu + \lambda)\alpha^2)K \sin\beta$, $K_2 = 2(\mu + \lambda)\alpha\beta K \cos\beta$, at all points $(\alpha, \beta)$ satisfying (3). Similarly from (2) and the boundary conditions, the $x$-dependence $W$ of the displacement component along the slab will be given by

$$K_3 W = (\lambda\beta^2 + (2\mu + \lambda)\alpha^2) \sin\beta \cos\frac{\alpha x}{a} - 2(\mu + \lambda)\alpha\beta \sin\alpha \cos\frac{\beta x}{a} \tag{2a}$$

or by

$$K_4 W = 2(\mu\beta^2 - (2\mu + \lambda)\alpha^2) \cos\beta \cos\frac{\alpha x}{a} + (\lambda\beta^2 + (2\mu + \lambda)\alpha^2) \cos\alpha \cos\frac{\beta x}{a}, \tag{2b}$$

$$K_3 = i\alpha\gamma a(\mu + \lambda)\frac{\lambda\beta^2 + (2\mu + \lambda)\alpha^2}{\mu\beta^2 - (2\mu + \lambda)\alpha^2} K \sin\beta, \quad K_4 = 2i\alpha\gamma a(\mu + \lambda)K \cos\beta.$$

Examine now, for example, a material for which $\lambda = \mu$, in the branch whose cut-off is at $\beta = 2\pi$, $\alpha = 2\pi/\sqrt{3}$. It can be verified at once that the nodes of the two components at some of the values of $(\alpha, \beta)$ discussed earlier are described by the following table $\left(\text{in which } t \equiv \cos\frac{\pi x}{a}\right)$:

| $\alpha$ | $\beta$ | Values of $x/a$ for nodes of $U$ | Values of $x/a$ for nodes of $W$ |
|---|---|---|---|
| $2\pi/\sqrt{3}$ | $2\pi$ | 0, ± two values given by $2/\sqrt{3} \sin 2\pi x/a\sqrt{3} = \cos 2\pi/\sqrt{3} \sin 2\pi x/a$ | $\pm\frac{1}{4}, \pm\frac{3}{4}$ |
| $\pi$ | $2\pi$ | 0, $\pm 1$, ± one value given by $14t + 8 = 0$ | ± two values given by $14t^2 - 2t - 7 = 0$ |
| $\pi/2$ | $5\pi/2$ | 0, ± two values given by $22t^2 + 11t - 2 = 0$ | $\pm 1$, ± one value given by $5t^2 - 15t - 11 = 0$ |
| 0 | $3\pi$ | 0, $\pm\frac{1}{3}, \pm\frac{2}{3}, \pm 1$ | no nodes |
| $7\pi i/2\sqrt{3}$ | $7\pi/2$ | 0, $\pm\frac{2}{7}, \pm\frac{4}{7}, \pm\frac{6}{7}$ | $\pm\frac{1}{7}, \pm\frac{3}{7}, \pm\frac{5}{7}, \pm 1$ |
| $i\infty$ | $3\pi$ | 0, $\pm\frac{1}{2}, \pm\frac{2}{3}, \pm 1$ | $\pm\frac{1}{6}, \pm\frac{1}{3}, \pm\frac{2}{3}$ |

It is to be noted in general that the nodal variations become less extreme at high frequencies, since for all branches except the lowest $U$ and $W$ tend

to become proportional to $\sin\frac{\beta x}{a}$ and $\cos\frac{\beta x}{a}$ respectively, and $\beta$ approaches the value $n\pi$ where $n$ is the branch number in order of increasing cut-off frequency, with $n = 0$ ascribed to the lowest branch. Thus the feeling, derived from more familiar cases of wave motion, that the order in which the branches arrange themselves should be correlated with the number of nodes they display retains an asymptotic validity here, in respect of each displacement component.

Nodes of absolute displacement will occur only at special frequencies. With the notation $\alpha' \equiv \frac{\alpha x}{a}, \beta' \equiv \frac{\beta x}{a}$, the conditions for their occurrence can be written

$$(\mu\beta'^2 - (2\mu + \lambda)\alpha'^2)\sin\beta'\cos\alpha' + (\mu + \lambda)\alpha'\beta'\cos\beta'\sin\alpha' = 0,$$

$$\frac{\sin 2\alpha'}{\sin 2\alpha} = \frac{\sin 2\beta'}{\sin 2\beta}, \qquad \frac{\beta'}{\alpha'} = \frac{\beta}{\alpha}, \qquad \beta' \leq \beta, \qquad \alpha' \leq \alpha,$$

taken together with (3).

## 3. The Cylinder

Procedures analogous to those of the preceding section, and presented by Love[8], lead to Pochhammer's secular equation, which in the present notation* is

$$\begin{aligned}(\lambda\beta^2 + (2\mu + \lambda)\alpha^2)^2 J_0(\alpha)J_1(\beta) \\ + 4(\mu + \lambda)\alpha\beta(\mu\beta^2 - (2\mu + \lambda)\alpha^2)J_1(\alpha)J_0(\beta) \\ + 2(\mu + \lambda)(2\mu + \lambda)\alpha(\alpha^2 - \beta^2)J_1(\alpha)J_1(\beta) = 0,\end{aligned} \tag{9}$$

where (4), (5) and (6) still hold, with $a$ signifying the radius of the rod. The analogy between (3) and the first two terms of (9) is striking. Again the roots $\beta = 0$ and $\beta = \alpha$ can be neglected, and the equation when divided by $\beta$ becomes even in $\alpha$ and $\beta$, and a plot of $\beta^2$ against $\alpha^2$ becomes appropriate, with the restrictions (7) as to regions of significance. The following paragraphs are lettered to correspond with their analogues of the preceding section.

(a) Setting $\mu\beta^2 = (2\mu + \lambda)\alpha^2$ in (9) reveals the cut-offs at $J_1(\beta) = 0$ and at $(2\mu + \lambda)\alpha J_0(\alpha) = 2\mu J_1(\alpha)$.

(b) Setting $\alpha = 0$ in (9), it can be seen that the roots intersect the line

* In comparing this treatment with that of Hudson[5], interpret his symbols

$$x \to \beta,\ y \to \alpha,\ \tau_0 \to \frac{\omega a}{c_0},\ \alpha \to \frac{2\mu}{2\mu+\lambda}.$$

$\alpha^2 = 0$ at the points $J_1(\beta) = 0$, traversing them with $\frac{d(\beta^2)}{d(\alpha^2)} = \frac{-4\mu(\mu+\lambda)}{\lambda^2}$, or half that of their analogues. Again it is only at those points that the phase velocity has the value $\sqrt{(2\mu+\lambda)/\rho}$.

(c) Setting $J_1(\beta) = 0$ reduces (9) to $\alpha\beta(\mu\beta^2 - (2\mu+\lambda)\alpha^2)J_1(\alpha) = 0$, and hence the roots are confined between the horizontal lines $J_1(\beta) = 0$ in the region of positive $\beta^2$ and negative $\alpha^2$, and the velocity is asymptotic to $\sqrt{\mu/\rho}$ with increasing frequency. They meet the line $\lambda\beta^2 + (2\mu+\lambda)\alpha^2 = 0$ at the points $\beta J_0(\beta) - J_1(\beta) = 0$, or in other words at the maxima and minima of $J_1(\beta)$,* where again the phase velocity is $\sqrt{2\mu/\rho}$.

(d) In the case of the cylinder confining lines** are

$$\left[J_1'(\alpha) + \frac{\lambda(\beta^2-\alpha^2)}{\alpha(\lambda\beta^2+(2\mu+\lambda)\alpha^2)}J_1(\alpha)\right]J_1'(\beta) - J_1(\alpha)J_1(\beta) = 0,$$

since substitution shows that intersection of (9) with these lines would require

$$\frac{4(\mu+\lambda)\alpha\beta(\mu\beta^2-(2\mu+\lambda)\alpha^2)}{(\lambda\beta^2+(2\mu+\lambda)\alpha^2)^2}J_1^2(\alpha) = -\left[J_1'(\alpha) + \frac{\lambda(\beta^2-\alpha^2)}{\alpha(\lambda\beta^2+(2\mu+\lambda)\alpha^2)}J_1(\alpha)\right]^2,$$

which cannot be satisfied by permitted values of $\alpha^2$ and $\beta^2$.

(e) This suggests that in that region the roots may oscillate about the lines

$$J_1(\alpha)J_1'(\beta) + J_1(\beta)\left[J_1'(\alpha) + \frac{\lambda(\beta^2-\alpha^2)}{\alpha(\lambda\beta^2+(2\mu+\lambda)\alpha^2)}J_1(\alpha)\right] = 0.$$

In fact, in view of the equivalence $J_1'(x) = J_0(x) - \frac{1}{x}J_1(x)$, those lines can be seen to have points in common with the roots of the secular equation at $J_1(\alpha) = 0$, $J_1(\beta) = 0$, and at

$$J_1'(\beta) = 0, \qquad J_1'(\alpha) + \frac{\lambda(\beta^2-\alpha^2)}{\alpha(\lambda\beta^2+(2\mu+\lambda)\alpha^2)}J_1(\alpha) = 0.$$

(f) Substituting the expression for the lines of (e) into (9) shows that again additional intersections may be afforded by suitable roots of the cubic equation (8).

* The analogy to the corresponding intersections for the slab at the maxima and minima of $\sin\beta$ is noted by Lamb, ref. 2, p. 122, footnote.

** There are infinitely many such families of lines but none carries the analogy with the slab to the point of being independent of the elastic constants. The families used in (d) and (e) serve the present purpose as simply as any.

(g) In the region $\alpha = iA$, the analogous lines are given by

$$iJ_1(iA)J_1'(\beta) + J_1(\beta)\left[J_1'(iA) - \frac{\lambda(\beta^2 + A^2)}{A(\lambda\beta^2 - (2\mu + \lambda)A^2)}\, iJ_1(iA)\right] = 0,$$

with which intersections occur for the same values of $\beta/A$ as in the slab.

These results permit visualization of a counterpart to Fig. 1 for the cylindrical case. In it the critical lines radiating from the origin are the same. The horizontal lines, instead of being evenly spaced by $\pi/2$, are spaced as the zeros, maxima, and minima of $J_1(\beta)$. The vertical lines, again no longer evenly spaced, are replaced alternately by straight vertical lines $J_1(\alpha) = 0$ and by the curved "vertical" lines $J_0(\alpha) = \dfrac{2(\mu + \lambda)\alpha}{\lambda\beta^2 + (2\mu + \lambda)\alpha^2} J_1(\alpha)$ which lie between their straight companions and approach $J_0(\alpha) = 0$ as $\beta$ becomes large. Finally the confining lines, and the lines about which the branches oscillate, become the curves defined in (d) and (e), which can be seen to follow a course not dissimilar to the diagonal course of their predecessors, passing through the intersections of the new horizontals and verticals.

Again the branches do not intersect, except for pair-wise coincidence of cut-offs on one or another of the lines (d) when the elastic constants obey special relations. Thus the dispersion curves to which Hudson[5] assigns certain of Shear and Focke's data cannot be taken (and indeed Hudson does not suggest that they must be taken) as corresponding to higher branches of the longitudinal modes, since the former curves intersect one another, and the latter cannot unless anisotropy modifies their behavior qualitatively. The assignments could represent modes other than longitudinal. The more recent results of Hueter[9] show essentially the behavior of Fig. 3.

In view of the closeness of the analogy thus revealed, it may be taken as probable that qualitative correspondence will obtain quite generally between the longitudinal modes of the slab and those of the cylinder.

## References

1. Lord Rayleigh, *Proc. Lond. Math. Soc. 17*, 4 (1885).
2. H. Lamb, *Proc. Roy. Soc. Lond. A, 93*, 114 (1917).
3. L. Pochhammer, *J. reine angew. Math.* (Crelle) *81*, 33 (1875).
4. D. Bancroft, *Phys. Rev. 59*, 588 (1941).
5. G. E. Hudson, *Phys. Rev. 63*, 46 (1943).
6. S. K. Shear and A. B. Focke, *Phys. Rev. 57*, 532 (1940).
7. R. W. Morse, *Jl. Acous. Soc. Am. 20*, 833 (1948).
8. A. E. H. Love, "Mathematical Theory of Elasticity," Cambridge 1927, 4th Ed., p. 287.
9. T. F. Hueter, *Jl. Acous. Soc. Am. 22*, 514 (1950); *Zeit. angew. Phys. 1*, 274 (1949).

# Frequency Dependence of Elastic Constants and Losses in Nickel

By R. M. BOZORTH, W. P. MASON and H. J. McSKIMIN

The elastic constants of nickel crystals, and their variation with magnetic field ($\Delta E$ effect), have been measured by a 10-megacycle ultrasonic pulsing method. The constants of three crystals agree well with one another when the crystals are magnetically saturated, but vary with domain distribution when demagnetized. The maximum $\Delta E$ effect observed is much less (3%) than has been observed at lower frequencies (20%). By measuring the $\Delta E$ effect and the decrement of polycrystalline rods at low frequencies, it is shown that the small effect observed at 10 megacycles is due to a relaxation in the domain wall motion due to micro-eddy-current damping.

From the initial slope of the decrement-frequency curve, and also from the frequency of maximum decrement, the size of the average domain is found to be about 0.04 mm. Actual domains in single nickel crystals have been observed optically by Williams, who finds domain widths of 0.02 to 0.2 mm.

THE three elastic constants of nickel have been determined in several single crystals by measuring the velocity of pulses of elastic waves of frequency 10 mc/s and duration 0.001 sec. The method has been described by McSkimin[1] and the preliminary results on nickel have already been reported briefly.[2]

It is well known that Young's modulus, $E$, increases with magnetization, and changes in $E$ (the "$\Delta E$ effect") by 15 to 30 per cent have been observed at room temperature and changes by greater amounts at higher temperatures.[3] It was surprising to find then, in our own experiments at 10 mc, that the greatest change was only about 3 per cent. It then occurred to us that, at such a high frequency, relaxation of the domain wall motion by micro-eddy-current damping might be expected. This led to the investigation of the frequency dependence of $\Delta E$ and of the logarithmic decrement, $\delta$, in polycrystalline nickel, and the results obtained support the theory and give information about domain size, as described below. Calculations[4] based on the equations of domain wall motion give results which agree with the experiments.

A number of experiments[3] have already established the existence of micro-eddy-current losses in magnetic materials subjected to elastic vibrations. These losses have their origin in the local stress-induced changes of magnetization of the domains of which magnetic materials are composed. The change in magnetization of one domain will give rise to eddy-currents around it and in it, and the consequent loss in energy depends on the frequency $f$ and the resistivity $R$, and on the size and shape of the region in which the change in magnetization occurs. These losses are in addition to the *macro-*

eddy-current losses, due to the relatively uniform changes in magnetization of a magnetized specimen that occur during a change in stress.

Calculations of the logarithmic decrement, $\delta_i$, attributable to micro-eddy-currents, have been made by Becker and Döring[3] and by one of the writers.[4] According to these calculations when the material is composed of plate-like domains of thickness $l$, in which magnetization changes by boundary displacement, the decrement for nickel, which has its directions of easy magnetization parallel to [111] directions, is given by the relation

$$\delta = \frac{\mu_0 E_s \lambda_{111}^2}{5I_s^2}\left[\frac{5c_{44}}{c_{11} - c_{12} + 3c_{44}}\right]^2 \frac{f/f_0}{1 + f^2/f_0^2} \tag{1}$$

where $f_0$, the relaxation frequency for domain wall motion is $f_0 = \dfrac{\pi R}{24\mu_0 l^2}$, $I_s$ is the saturation magnetization, $E_s$ is the saturated value of Young's modulus, $\mu_0$ is the initial permeability, $R$ the electrical resistivity, $\lambda_{111}$ the saturation magnetostriction along the [111] direction, and $c_{11}$, $c_{12}$ and $c_{44}$ the three elastic constants of nickel which are evaluated in this paper. For low frequencies the initial slope of the decrement vs frequency curve is

$$\frac{\delta}{f} = \frac{24 E_s \mu_0^2 l^2 \lambda_{111}^2}{5\pi R I_s^2}\left[\frac{5c_{44}}{c_{11} - c_{12} + 3c_{44}}\right]^2 \tag{2}$$

As the frequency is increased the decrement rises to a maximum and then declines asymptotically to zero. Both the initial slope of the $\delta$ vs $f$ curve and the frequency at which the maximum occurs are measures of the domain size. The initial slope has already been used to evaluate the size of the domains in 68 Permalloy.[5] It is shown in the present work that the maximum occurs in polycrystalline nickel at a frequency consistent with the dimensions of domains observed by Williams and Walker[6] in single crystals of nickel.

## Elastic Constants and Damping in Single Crystals

The nickel crystals used here were grown by slow cooling of the melt in a molybdenum wound resistance furnace, by a method previously described.[7] They were cut with major surfaces parallel to (110) planes and were placed between two fused quartz rods as shown in Fig. 1. Measurements of the elastic constants were made as described in detail by McSkimin,[1] by measuring the velocity of propagation of 10 mc pulses. In order to obtain a number of reflections in the crystal, films of polystyrene approximately $\frac{1}{4}$ wavelength thick are placed between the rods and the nickel crystal. This has the effect of lowering the impedances next to the nickel to small values and hence nearly perfect reflections at the two surfaces are obtained. The frequency is varied until successive reflections occur in phase, and the velocity is then calculated from the frequency and the dimensions of the crystal.

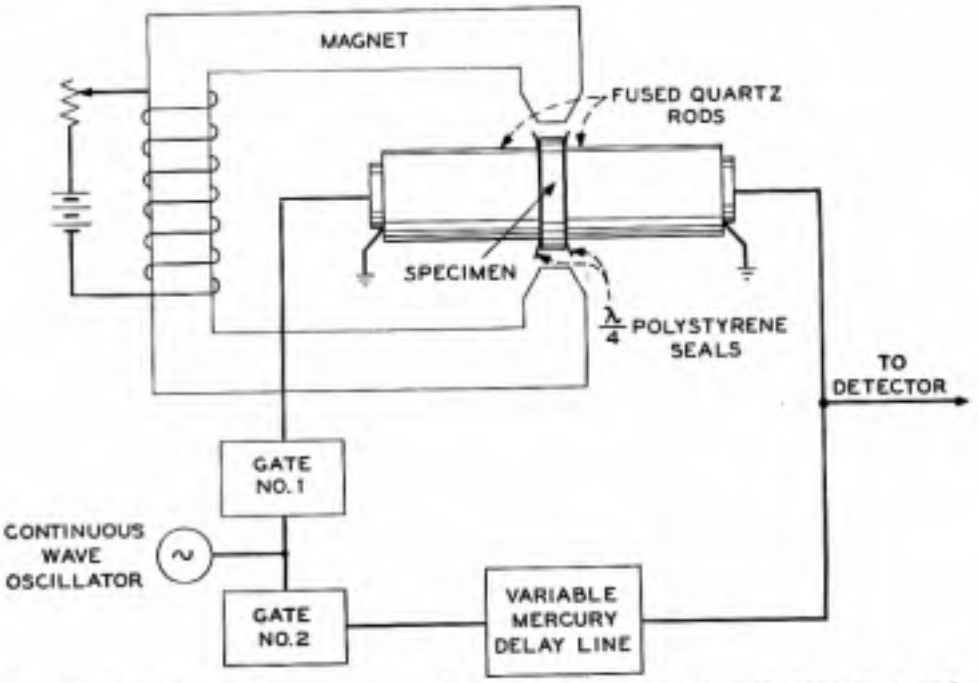



Fig. 1—Experimental arrangement for determining the elastic constants and $\Delta E$ effect in single nickel crystals.

TABLE I
ORIENTATION AND ELASTIC CONSTANTS

| Direction of Propagation | Direction of Particle Motion | Type of Vibration | Equation for Velocity | Measured Velocity (cm/sec) | Elastic constants (dynes/cm²) |
|---|---|---|---|---|---|
| 110 | 110 | Longitudinal | $v = \sqrt{(c_{11} + c_{12} + 2c_{44})/2\rho}$ | $6.03 \times 10^{5}$* | $c_{11} + c_{12} + 2c_{44} = 6.47 \times 10^{12}$ |
| 110 | $1\bar{1}0$ | Shear 1 | $v = \sqrt{(c_{11} - c_{12})/2\rho}$ | $2.26 \times 10^{5}$ | $c_{11} - c_{12} = 0.90 \times 10^{12}$ |
| 110 | 001 | Shear 2 | $v = \sqrt{c_{44}/\rho}$ | $3.65 \times 10^{5}$ | $c_{44} = 1.185 \times 10^{12}$ |

* A slight correction has been made for this value.

The velocities for one demagnetized crystal[1] were found to have the values shown by Table I. These values of velocity, and a density of 8.90 for the single crystal, give values for the demagnetized elastic constants of

$$c_{11} = 2.50, \qquad c_{12} = 1.60, \qquad c_{44} = 1.185 \tag{3}$$

all in $10^{12}$ dynes/cm².

To obtain the $\Delta E$ effect, the whole unit was placed between the jaws of a large electromagnet. Since the crystal was about 2.5 centimeters in diameter but only 0.472 cm thick, saturation could be obtained more easily along the long directions of the crystal. Figure 2 shows the changes in velocity of propagation along the [110] direction, caused by magnetization

along [001], for the shear mode with particle motion along [1$\bar{1}$0]. Fields of about 10,000 were attainable but a maximum field of about 6000 was usually used. The velocity increases by 2.6 per cent at the saturated value. On decreasing the field to zero the velocity drops below the original value

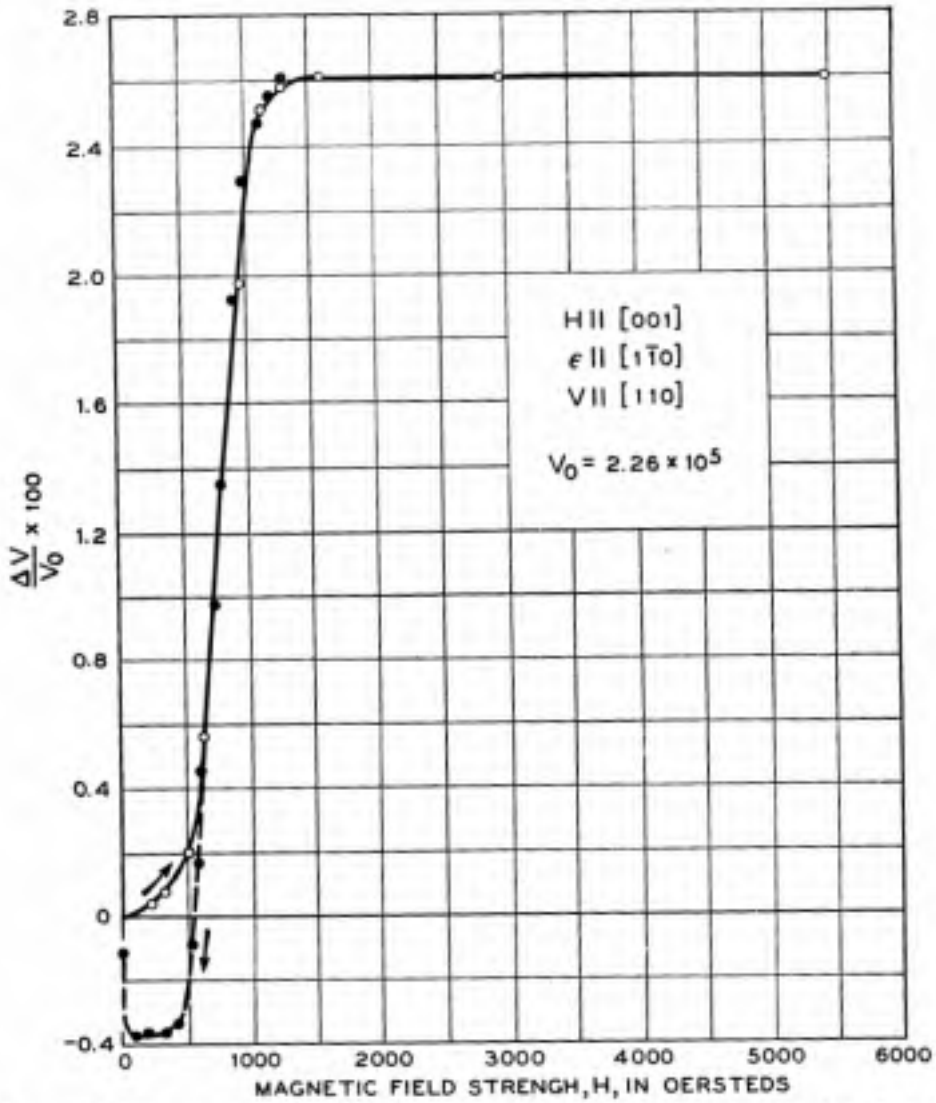



Fig. 2—Change in velocity in percent from demagnetized value as a function of the magnetizing field for a shear wave in a (110) section when the particle velocity $\epsilon$ is along the [1$\bar{1}$0] direction and the field $H$ along the [001] direction.

for the demagnetized state, but it has practically the initial value when the crystal is again demagnetized. The lower value of velocity for the return curve indicates that the free energy is lower for some arrangement of the elementary domains other than the demagnetized state.

Figure 3 shows the attenuation in decibels per trip as a function of mag-

netization. The loss drops from about 7 db to 1 db as the crystal becomes magnetized. The low value is the remanent loss caused by the energy lost to the terminations, so that one can say that the losses due to micro-eddy-current and micro-hysteresis are 6 db per trip or 12.7 db per centimeter for this mode of motion. The $Q$ of the crystal can be shown to be equal to

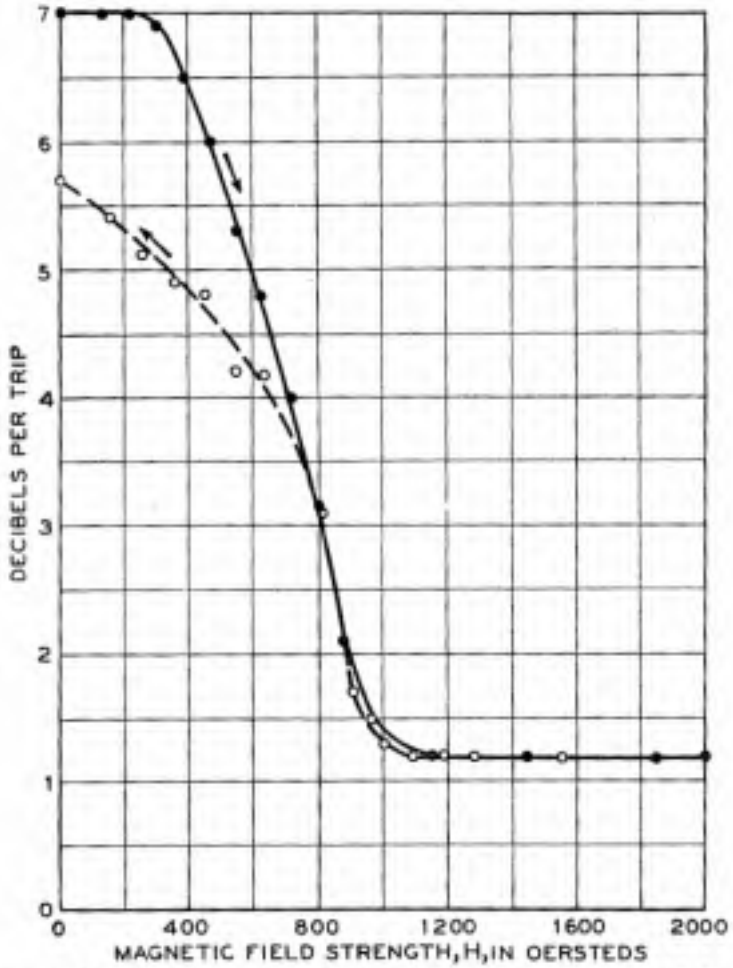



Fig. 3—Loss per trip (0.472 cm) as a function of magnetizing field for shear wave of Fig. 2.

the phase shift in radians divided by twice the attenuation in nepers per cm, or

$$Q = \frac{2\pi f/v}{2\ (db \text{ per cm})/8.68} \tag{4}$$

and our results give $Q = 94$, corresponding to a decrement of $\pi/Q = 0.033$.

Figure 4 shows a measurement of the same mode when the field is applied along the [110] direction. The velocity approaches a slightly different

limit on account of the "morphic" effect discussed in another paper.[8] If we average the two values the effective elastic constant for saturation becomes

$$c_{11}^s - c_{12}^s = 0.954 \times 10^{12} \text{ dynes/cm}^2. \quad (5)$$

Measurements for the field along the thickness did not produce saturation and are not shown.

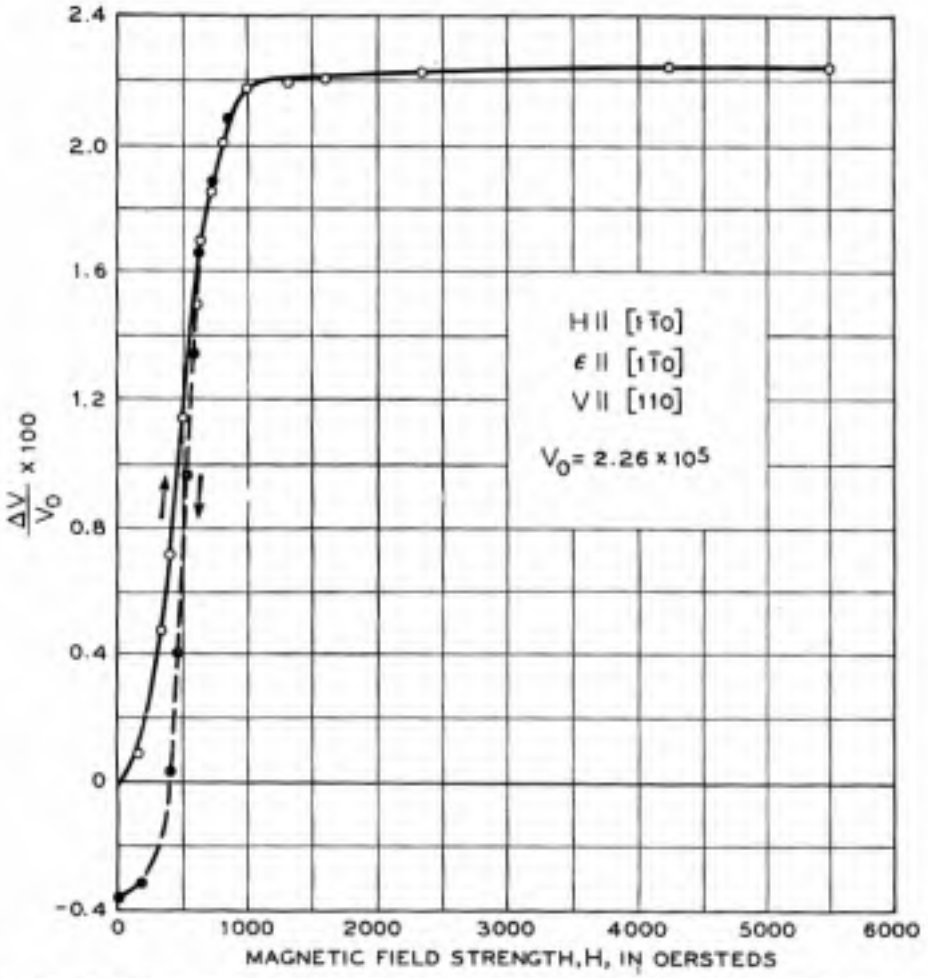



Fig. 4—Change in velocity in percent from demagnetized value as a function of the magnetizing field for a shear wave in a (110) section when the particle velocity is along the $[1\bar{1}0]$ direction and the field along the $[1\bar{1}0]$ direction.

Figures 5 and 6 show similar measurements for the other shear mode (Shear 2 of Table I) for two directions of the magnetic field. Averaging the two limiting values, the constant $c_{44}$ at saturation becomes

$$c_{44}^s = 1.22 \times 10^{12} \text{ dynes/cm}^2 \quad (6)$$

The $Q$ and decrement for this case become approximately 110 and 0.028.

Figures 7 and 8 show measurements for the longitudinal mode. Variations of about 0.6 per cent in the velocity are obtained, and the saturated elastic constants, $Q$ and decrement are

$$c_{11}^s + c_{12}^s + 2c_{44}^s = 6.55 \times 10^{12}, \qquad Q = 390, \qquad \delta = 0.008 \quad (7)$$

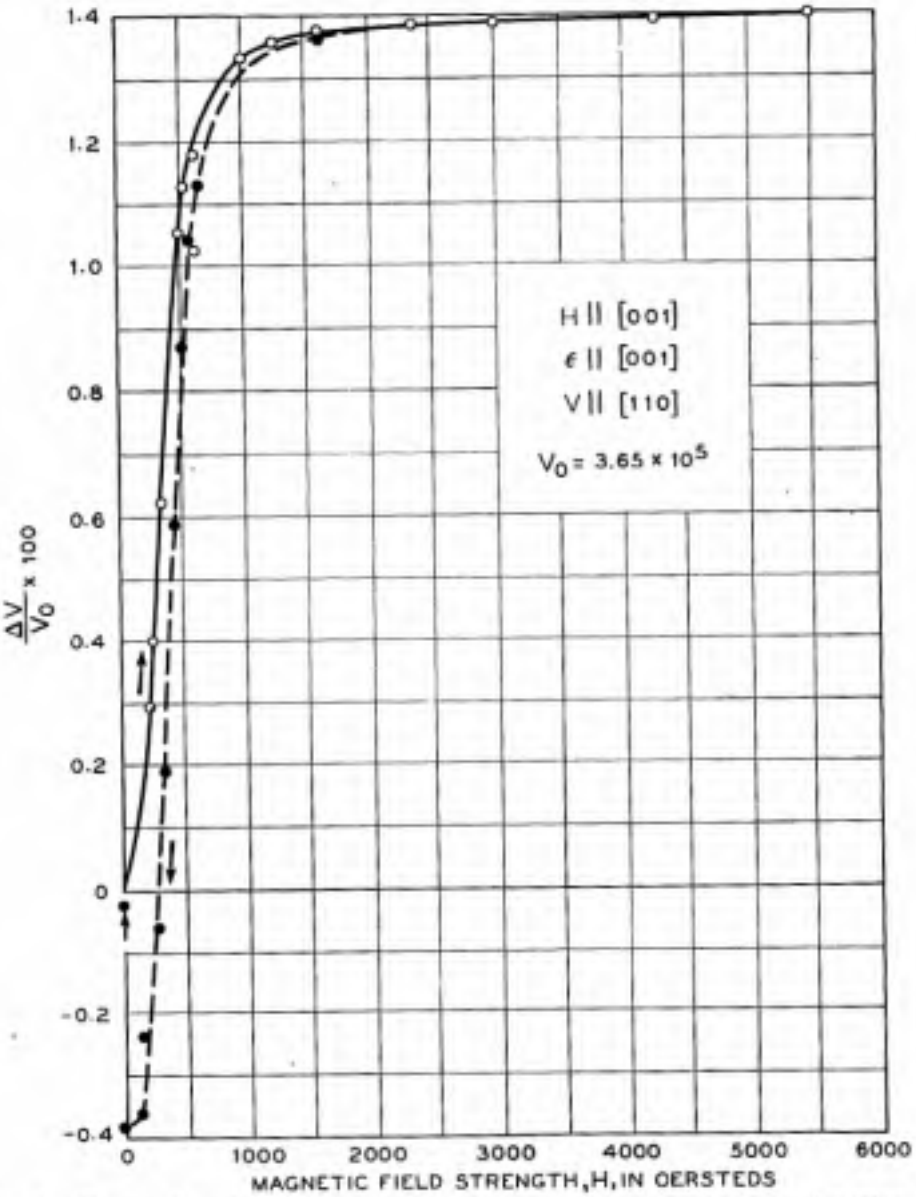



Fig. 5—Change in velocity in percent from demagnetized values as a function of the magnetizing field for a shear wave in a (110) section when the particle velocity is along the [001] direction and the field along the [001] direction.

Combining the elastic constants, the saturated elastic constants are evaluated:

$$c_{11}^s = 2.53, \qquad c_{12}^s = 1.58, \qquad c_{44}^s = 1.22, \quad (8)$$

all in $10^{12}$ dynes/cm$^2$.

It is obvious from the measurements of Figs. 4 to 8 that the changes in the elastic constants with magnetization are much smaller at the high frequencies (10 mc) than they are at the lower frequencies of 10 to 50 kilo-

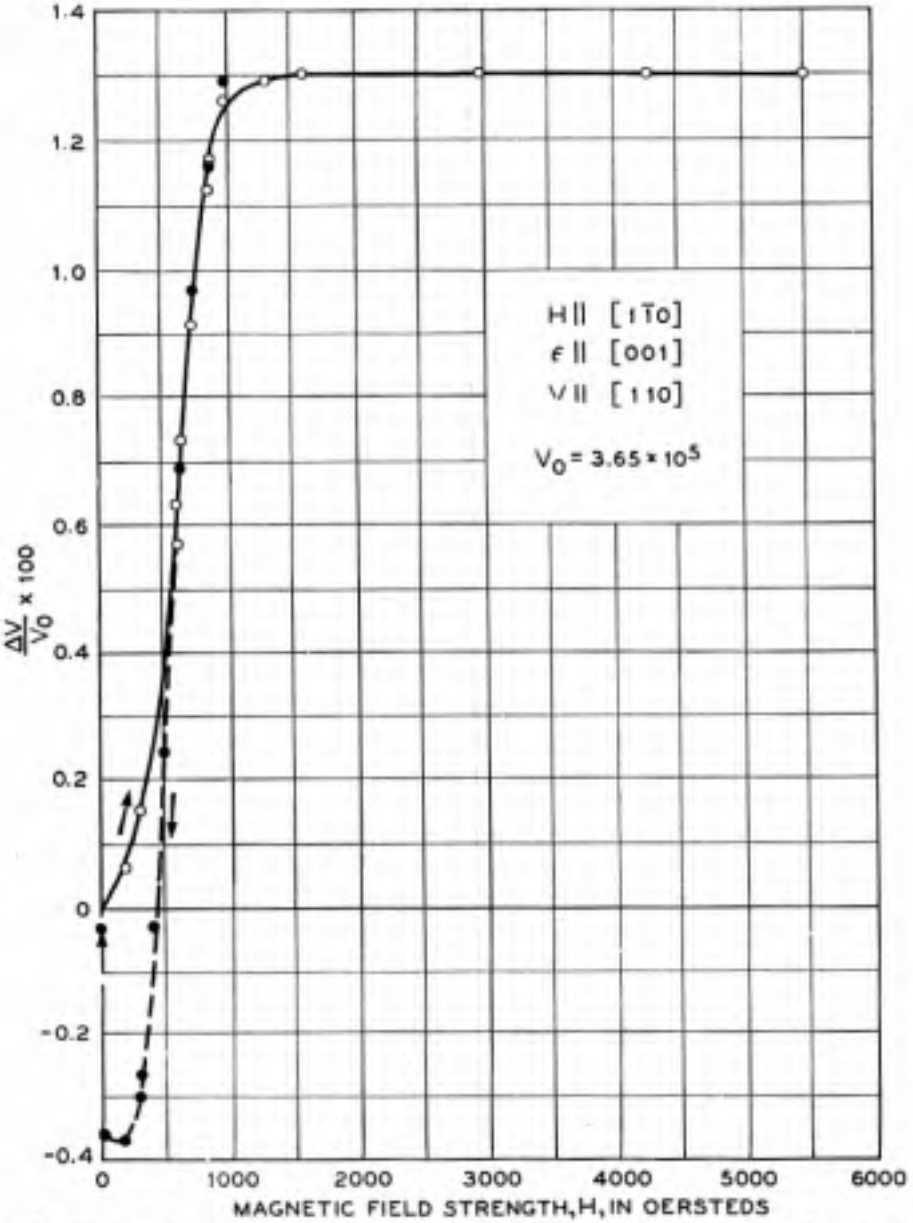



Fig. 6—Change in velocity in percent from demagnetized value as a function of the magnetizing field for a shear wave in a (110) section when the particle velocity is along the [001] direction and the field along the [1$\bar{1}$0] direction.

cycles where changes of 15 to 30 per cent have been observed in polycrystalline material.[3] A rough comparison of the low-frequency values with the 10 megacycle values can be obtained if we convert the observed changes in the $c$'s to the equivalent change in $E$. This can be done if we use the method

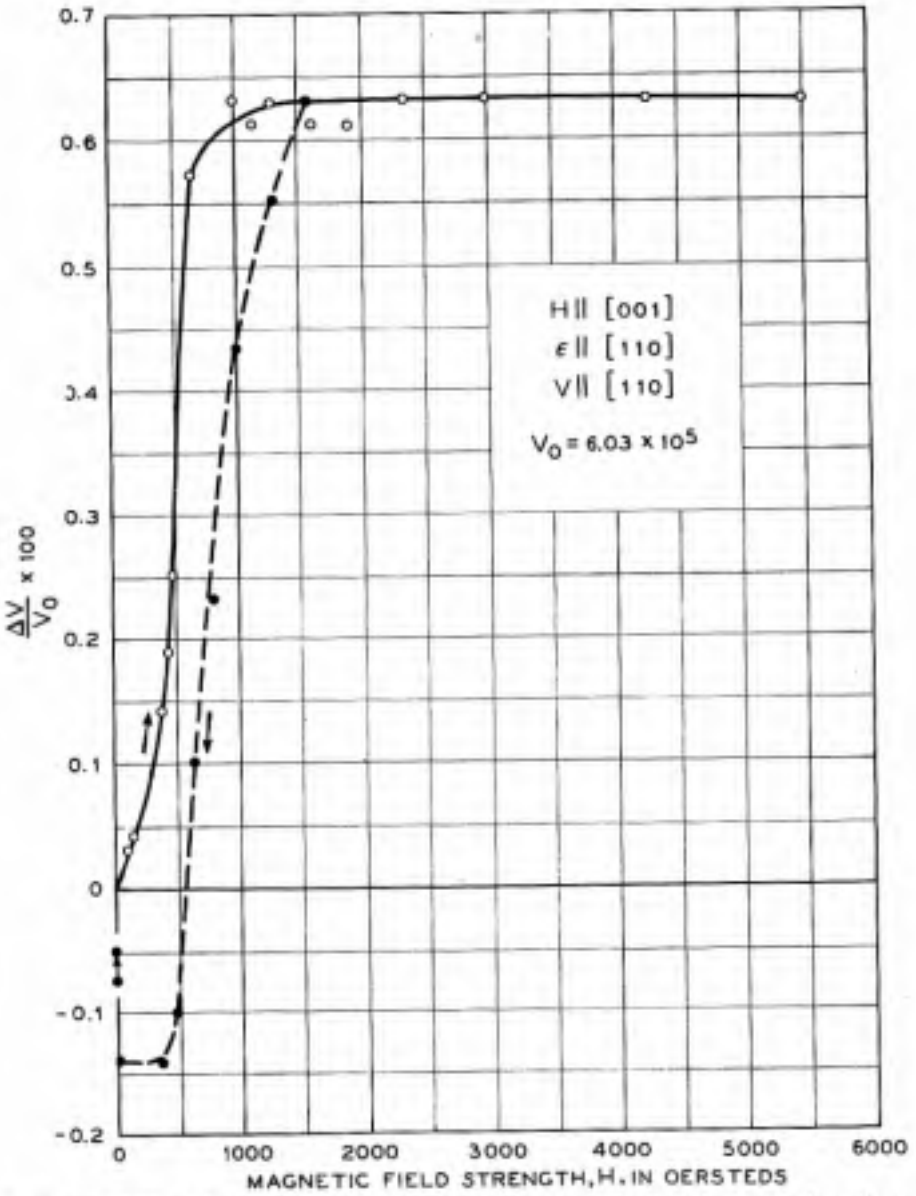



Fig. 7—Change in velocity in percent from demagnetized value as a function of the magnetizing field for a longitudinal wave in a (110) section when the particle velocity is along the [110] direction and the field along the [001] direction.

developed by one of the writers[9] for obtaining the elastic constants of a polycrystalline rod from the cubic elastic constants. In this case the Lamé elastic constants are given by the formulas

$$\lambda + 2\mu = \frac{3}{5}c_{11} + \frac{2}{5}c_{12} + \frac{4}{5}c_{44},$$
$$\mu = \frac{3}{5}c_{44} + \frac{c_{11} - c_{12}}{5}. \tag{9}$$

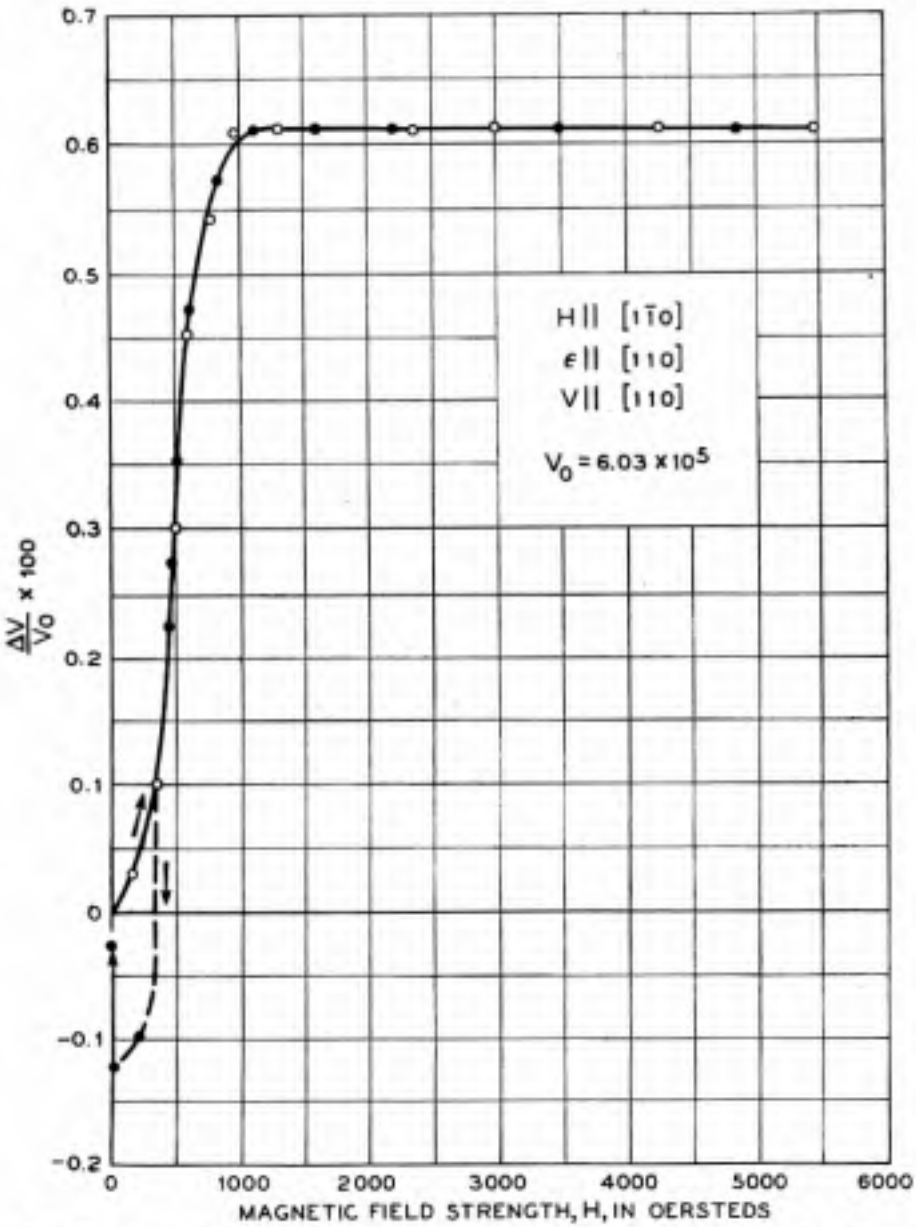



Fig. 8—Change in velocity in percent from demagnetized value as a function of the magnetizing field for a longitudinal wave in a (110) section when the particle velocity is along the [110] direction and the field along the [1$\bar{1}$0] direction.

Since in terms of the Lamé elastic constants, the value of Young's modulus is

$$E_0 = \mu \left( \frac{3\lambda + 2\mu}{\lambda + \mu} \right) \tag{10}$$

one finds that the difference between the saturated and demagnetized value of Young's modulus divided by the demagnetized value is

$$\frac{\Delta E}{E_0} = \frac{E_s - E_0}{E_0} = \frac{2.381 - 2.312}{2.312} = 0.03 = 3\% \qquad (11)$$

which is much smaller than that given by low frequency measurements.

To check our results, and be sure that the crystals were free from imperfections and strains, two other crystals were prepared and carefully annealed at 1100°C. The values found for the changes in elastic constants were considerably less for these crystals. The $Q$'s of the crystal were also higher. Table II shows the measured values and the equivalent $\Delta E/E$ values. The table shows also the measurements for the demagnetized crystal of two Japanese workers[10,11] and the equivalent $\Delta E/E$ assuming that the saturated elastic constants are the same as those found for the other three crystals. Since these vary by only $\pm 0.5$ per cent among themselves, this appears to be a good approximation.

TABLE II

ELASTIC CONSTANTS (IN $10^{12}$ DYNES/CM$^2$) AND $\Delta E$-EFFECT IN SINGLE CRYSTALS OF NICKEL

| Crystal | Magnetically Saturated | | | Demagnetized | | | $\Delta E/E$ |
|---|---|---|---|---|---|---|---|
| | $c_{11}$ | $c_{12}$ | $c_{44}$ | $c_{11}$ | $c_{12}$ | $c_{44}$ | |
| 1 | 2.53 | 1.58 | 1.22 | 2.50 | 1.60 | 1.185 | 0.03 |
| 2 | 2.524 | 1.538 | 1.23 | 2.52 | 1.54 | 1.229 | 0.0017 |
| 3 | 2.523 | 1.566 | 1.23 | 2.517 | 1.574 | 1.226 | 0.0046 |
| Yamamoto[11] | | | | 2.44 | 1.58 | 1.02 | 0.16 |
| Honda and Shirakawa[10] | | | | 2.52 | 1.51 | 1.04 | 0.11 |

The lower values of $\Delta E/E$ for the second and third crystals are probably due to larger domain sizes, caused by the longer anneal.

## DAMPING AND $\Delta E$-EFFECT IN POLYCRYSTALLINE RODS

To test the theory of micro-eddy-current shielding (see Introduction), the velocity and attenuation of elastic vibrations in well-annealed polycrystalline nickel rods were measured over the frequency range of 5 kilocycles to 150 kilocycles. In the method of measurement,[12] shown by Fig. 9, two matched piezoelectric crystals of resonance frequency corresponding to integral half wavelengths along the rod, are attached to the ends of the rod. Phase-amplitude balance was obtained by critical adjustment of frequency and output of the calibrated attenuator. The corresponding level was then compared with that obtained when the two crystals were cemented directly together. With little error, the velocity of propagation is given by

$$v = \frac{2fl_0}{n}, \qquad n = 1, 2, 3 \cdots \qquad (12)$$

The attenuation $A$ (and hence the $Q$ of the rod) was obtained by solving the equation

$$\sinh Al_0 = \frac{(r - \cosh Al_0)n\pi M_c}{Q_c M_R} \tag{13}$$

in which $r$ is the ratio of output with crystals together to output with specimen attached, $l_0$ is the length of rod, $M_c$ the mass of either crystal, $M_R$ the mass of the rod, and $Q_c$ the $Q$ of the crystal as determined by resonance response method.

For this equation to apply accurately, the terminating impedance presented to the rod by the crystals at resonance must be small compared to the characteristic impedance of the rod, and the $Q$ of the rod should be $>10$. This method may be used even when the total loss in the rod is so high that well defined resonances no longer exist. At the lower frequencies, however,

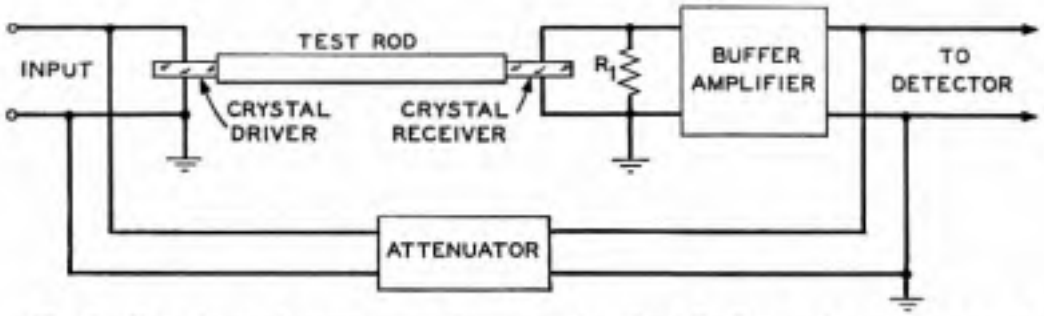



Fig. 9—Experimental arrangement for measuring the $\Delta E$ effect and associated loss in a polycrystalline rod at low frequencies.

a useful check may be made by the resonance response method of determining $Q$ which involves determining the frequency separation $\Delta f$ for two frequencies 3 db from the maximum response frequency, and using the formula

$$Q = \frac{f_{max}}{\Delta f} \tag{14}$$

Correction for the mass and dissipation of the piezoelectric crystals must of course be made. Both methods have been found to agree within about 10% —the probable error to be expected.

The Appendix lists formulae to be used when the resonance frequency of the crystal driver differs from the frequency at which phase balance is obtained. This condition of necessity occurred when the rods were subjected to a magnetic field, which caused an increase in the velocity of propagation.

Figure 10 shows a typical measurement of change in frequency and change in decrement with magnetizing field excited in a solenoid surround-

ing the nickel rod. Saturation is not quite obtained so that the $\Delta E$ effect measured is slightly lower than the true value, but for relative frequency comparisons this is not important.

The first rod measured was 0.320 cm in diameter and 10.16 cm long and was annealed at 1100°C. Five frequencies ranging from 22.5 kilocycles were

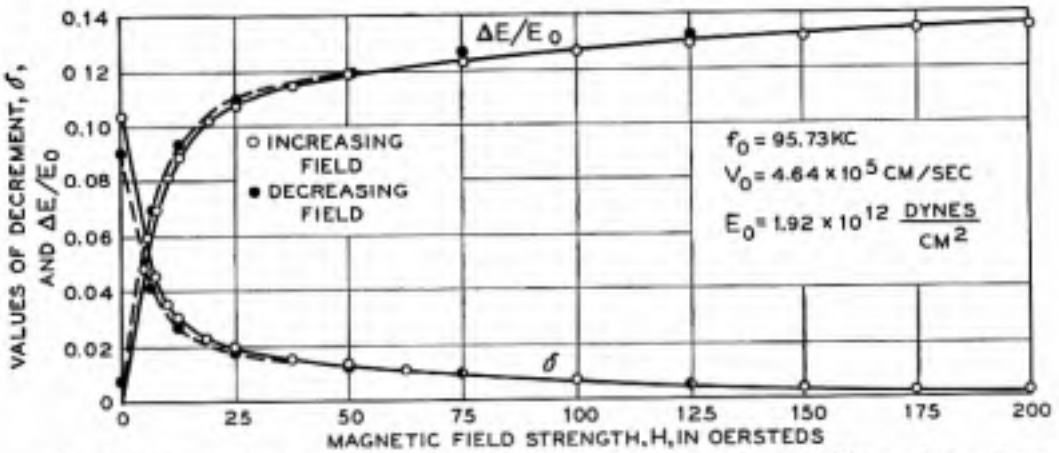



Fig. 10—Typical change in velocity and decrement of a polycrystalline rod as a function of the magnetizing field.

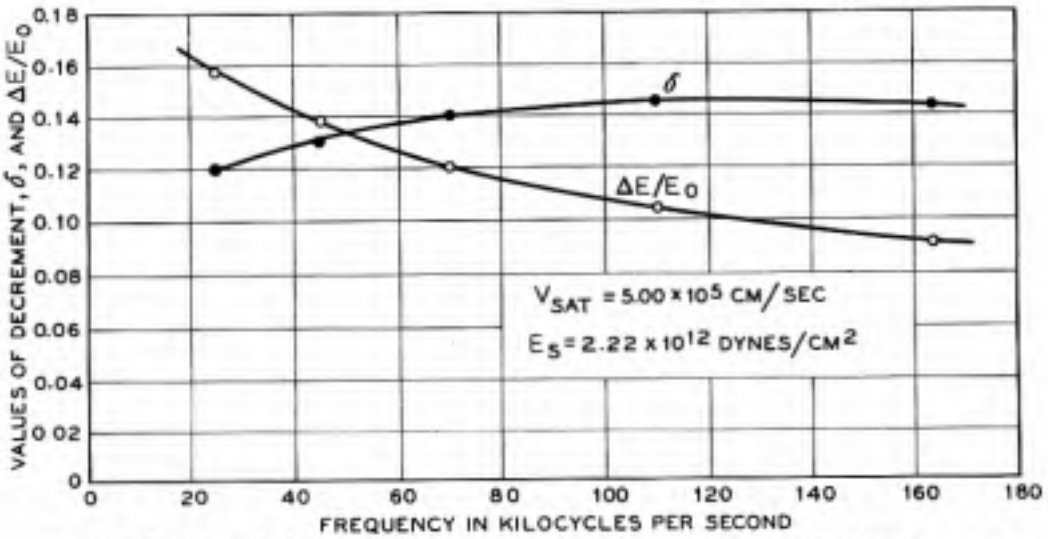



Fig. 11—Fractional change in Young's modulus, and the decrement, plotted as a function of frequency for rod No. 1.

used and the ratio of the change in Young's modulus to the value of Young's modulus for the demagnetized rod is shown plotted in Fig. 11. This figure shows also the decrement $\delta = \pi/Q$. It is obvious that the decrement eventually decreases as the frequency rises, and this is contrary to the simple theory of the micro-eddy-current effect,[3] which indicates that the decrement

should increase linearly with the frequency. The indicated maximum for this rod is below 120 kilocycles.

In order to obtain the first part of the decrement vs frequency curve, a rod of 46.05 cm length and 0.637 cm diameter was next used. This rod was annealed at 1050°C and presumably has a smaller average domain size than the first one, so that the important variations occur in a more favorable frequency range.

The changes in elastic constant and the decrement for this rod are shown by Fig. 12 for frequencies from 5 kilocycles to 96 kilocycles. At the lower frequencies the decrement increases in proportion to the frequency in

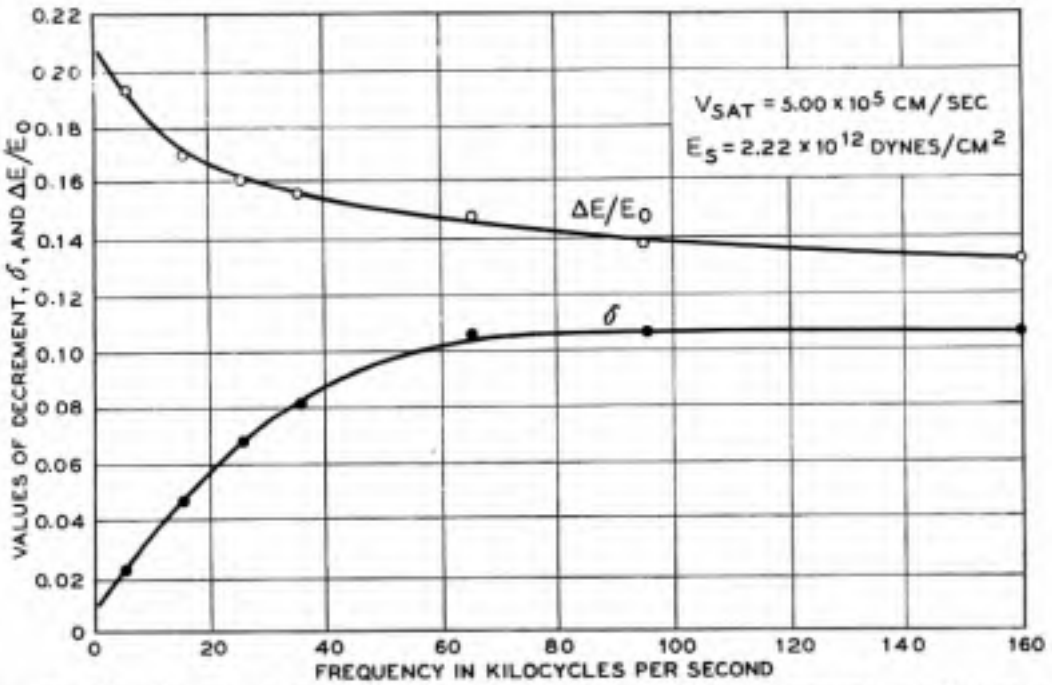



Fig. 12—Fractional change in Young's modulus, and the decrement, plotted as a function of frequency for rod No. 2.

agreement with the simple theory. By extending this curve down to zero frequency it is seen that a micro-hysteresis effect (which is independent of the frequency) gives an initial decrement of about 0.010. The decrement rises to an indicated maximum at somewhat more than 100 kilocycles and the change in elastic constant with saturation decreases with frequency.

The data on these two rods taken together indicate that there is a frequency of maximum decrement and for frequencies above and below this the decrement is smaller. The $\Delta E$ change in the elastic constant decreases as the frequency increases and for very high frequencies the $\Delta E$ effect becomes very small. As shown by the discussion in the next section, the fre-

quency of the maximum value of $\delta$ and the initial slope of the decrement frequency curve are connected with definite domain sizes which can be calculated approximately and compared with magnetic domain powder patterns.

## Discussion

Our determinations of the *elastic constants* may be discussed in relation to the values obtained by others. The results reported by Honda and Shirakawa[10] and Yamamoto[11] were unknown to us and unavailable at the time of our preliminary communication. The data of the Japanese, converted from $s$-constants to $c$-constants by the relations:

$$c_{11} = \frac{s_{11} + s_{12}}{s_{11}^2 + s_{11}s_{12} - 2s_{12}^2}$$

$$c_{12} = \frac{-s_{12}}{s_{11}^2 + s_{11}s_{12} - 2s_{12}^2} \tag{15}$$

$$c_{44} = 1/s_{44}$$

are included with our data in Table II.

As our experiments show, the 10 mc pulses that we used are so rapid that micro-eddy-currents largely prevent the stress-induced changes in magnetization from penetrating the domains. Therefore the constants determined by this method are those for material almost saturated. The values at saturation are independent of the initial domain distribution, and of the ease with which the magnetization in the separate domains can be changed by stress, consequently they are the more fundamental elastic constants of the material. The variety of values for unmagnetized nickel is made evident from the scatter in the ratios of $\Delta E/E$ that have been reported.[3] The variation in the values of the $c$-constants recently published is thus not surprising. The values at saturation of the three crystals examined by us are in substantial agreement, as shown in Table II. They cannot be compared directly with the results of the Japanese workers because the latter reported data for unmagnetized crystals only and then $E$ is sensitive to heat treatment and domain configuration.

As mentioned in the introduction, the damping of elastic vibrations by micro-eddy-currents is proportional to the frequency at low frequencies (Eq. 1) and it rises with frequency to a maximum and then declines toward zero. The frequency at which the maximum occurs has been calculated[4] by using the equation of domain wall motion and evaluating the constants from the initial permeability and the power loss caused by domain wall motion. The maximum value of $\delta$ comes at the same value as that calculated

for eddy current losses in sheets having the same thickness as the domain, namely

$$l^2\mu_0 f_m/R = 0.13 \tag{16}$$

$l$ being the thickness of the sheet, $\mu_0$ the initial permeability, and $R$ the resistivity of the material (all in c.g.s. units).

As noted below, the domain sizes calculated from the initial slope of the $\delta$ vs $f$ curve of Fig. 12, and from the frequency at which the maximum decrement occurs, are respectively 0.035 mm and 0.045 mm (for plates). These values agree quite well. The decrement curve is broader than would be calculated from equation (1) for a single domain size. This agrees with the optical measurements of domain size by Williams,[6] which are shown by Fig. 13. This indicates domain sizes from 0.01 to 0.2 mm.

The maximum value for the decrement calculated from equation (1), using the measured values, is 0.35 compared to the observed value of 0.11. Part of this is due to the broadening of the peak caused by a distribution of domain sizes, but part may also be due to the deviation of the actual domain shape from a sheet which has been assumed in making the calculations.

The calculations of domain size are made in more detail as follows:

According to Döring[13] the change in Young's modulus for nickel containing only small internal strains is related to the initial permeability, $\mu_0$, as follows:

$$\frac{\Delta E}{E_0} = -\frac{\lambda_{111}^2(\mu_0 - 1)E_s}{5\pi I_s^2}\left[\frac{5c_{44}}{c_{11} - c_{12} + 3c_{44}}\right]^2 \tag{18}$$

provided the averaging over all crystallites is carried out with constant strain. (If constant stress is assumed, the fraction in brackets, equal to 1.76, is omitted.) For nickel $\lambda_{111}$ is $25 \times 10^{-6}$, $I_s$ is 484, and the $c$'s are the elastic constants given in Table I. This equation holds for low frequencies at which the shielding in single domains is negligibly small. When the relaxation effect of domain wall motion is considered[4] equation (18) has to be multiplied by the factor

$$1/(1 + f^2/f_0^2) \tag{19}$$

The data of Fig. 12 give the values:

$$E_0 = 1.83 \times 10^{12}, \qquad E_s = 2.22 \times 10^{12}\ \text{dynes/cm}^2, \frac{\Delta E}{E} = 0.21 \tag{20}$$

for low frequencies. Using these in the above equation, the calculated value of $\mu_0$ is 320. A direct measurement* of $\mu_0$ has been made for this rod and found to be 340, in good agreement with that deduced from the $\Delta E$ effect.

* Measurements were made independently by Miss M. Goertz and Mr. P. P. Cioffi in order to check this unusually high value.

Since the permeability is much higher than can be accounted for by domain rotation it is obvious that domain boundary motion is occurring. Hence in determining the domain sizes from the slope of the decrement vs

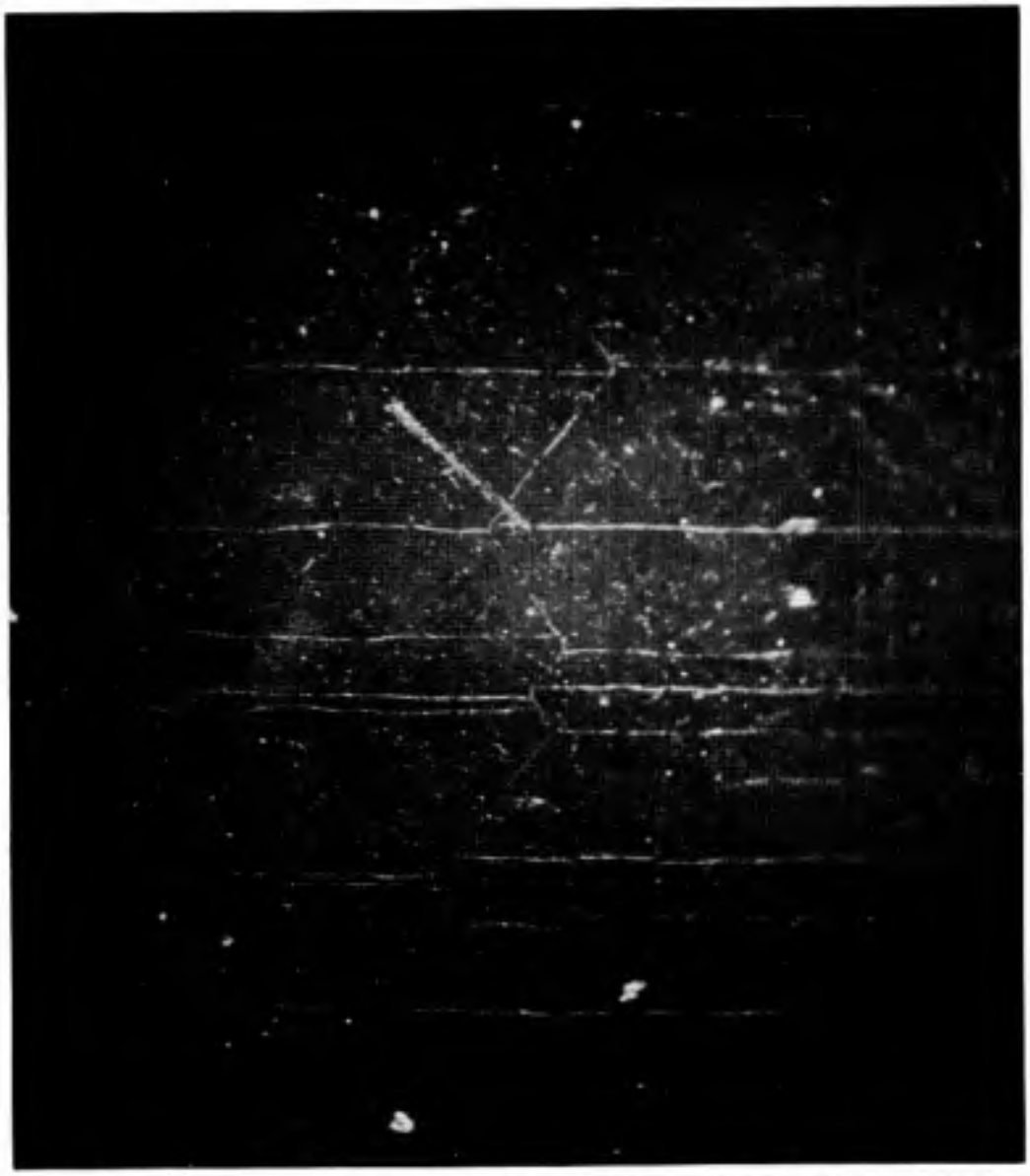


Fig. 13—Photograph of domains in a single nickel crystal (after Williams). Field of view, 0.5 mm.

frequency curve, equations (1) and (2) for domain boundary motion are appropriate.

When the data of Fig. 12 are extrapolated to zero frequency it appears that there is a microhysteresis loss (which is independent of the frequency) giving a decrement of 0.01. The initial slope of the $\delta_i$ vs $f$ curve is then about

$2.5 \times 10^{-6}$, and use of equation (2) with this and other appropriate values indicates that the domain size is

$$l = 0.035 \text{ mm}, \tag{21}$$

as reported above.

A check on this value can be obtained from the frequency, $f_m$, corresponding to the maximum of the $\delta$ vs $f$ curve. If we use equation (16),

$$l^2 f_m / R = 0.13, \tag{22}$$

with $f = 1.5 \times 10^5$ (Fig. 11), we find $l = 0.045$ mm, in reasonable agreement. An actual photograph of domains in a single crystal of nickel, taken by H. J. Williams and reproduced in Fig. 13, shows the presence of domains of various sizes ranging from about 0.01 to 0.2 mm. Any such range in domain sizes will naturally tend to flatten the maximum of the $\delta$ vs $f$ curve and, on account of the form of the $\delta$ vs $f$ function, will push the maximum to a higher frequency than that corresponding to the initial slope, and will give a lower maximum value to the decrement frequency curve.

The average domain size derived from our experiments is somewhat larger than that previously obtained in 68 Permalloy.[5] This may be expected, for nickel has a very high magnetostriction and the movement of domain boundaries by stress will be relatively large, possibly so large that the regions swept over by the domain walls will correspond to whole domains of the original domain structure, when the stresses are equal to those used in our experiments. The domain size which we have determined is based on this interpretation.

## APPENDIX

## METHOD OF MEASUREMENT—FORMULAE

From transmission line theory (see reference of footnote 12) the ratio of outputs, $r$, defined in the text and applicable to the circuit of Fig. 9 is given by

$$r = \cosh \theta l_0 + \frac{1}{2}\left(\frac{Z_T}{Z_0} + \frac{Z_0}{R_T}\right) \sinh \theta l_0 \tag{24}$$

where $\theta = A + jB$ = propagation constant

$Z_0 = \dfrac{jS\rho\omega}{A + jB}$ = characteristic impedance of rod

$Z_T$ = resistive terminating impedance provided by crystals

$S$ = area of rod

This expression may be expanded into real and imaginary parts and the latter term set to zero in accordance with the condition of phase balance.

Assuming that $|Z_0| \gg Z_T$ and that $Q = \frac{B}{2A} > 10$, the ratio $r$ which is now a real number determines the attenuation $A$ in accordance with equation (13) of the text.

If the crystal resonance frequency $f_c$ is slightly different from the balance frequency $f_s$ obtained with the rod specimen in place, correction may be made by considering a new terminating resistance $Z'_T$ formed by the crystal driver and a small section of the rod sufficient to make the combined resonance equal to $f_s$. A slightly different length, $l'_0$, of rod is then used to compute velocities and attenuation. Also a different mass $M'_R$ and ratio $r'$ result. The equation applicable provided $f_s = f_c$, is

$$\sinh Al'_0 = \frac{nZ'_T\,(r' - \cosh Al'_0)}{f_s M'_R} \tag{25}$$

where

$$\begin{aligned} r' &= \frac{r}{1 + \frac{Q_c}{2Q}\left(\frac{f_c}{f_s} - \frac{f_s}{f_c}\right)} \\ Z'_T &= M_c \pi f_s \left[\frac{f_c}{Q_c f_s} + \frac{1}{2Q}\left(\frac{f_c^2}{f_s^2} - 1\right)\right] \\ M'_r &= M_r\left[1 - \frac{M_c}{M_r}\left(\frac{f_c^2}{f_s^2} - 1\right)\right] \\ l'_0 &= l_0\left[1 - \frac{M_c}{M_r}\left(\frac{f_c^2}{f_s^2} - 1\right)\right] \end{aligned} \tag{26}$$

In the above a sufficiently accurate value of $Q$ is ordinarily obtained by assuming $f_c = f_s$. Further accuracy, if needed, can be obtained by recalculation, using the corrected value of $Q$.

We are glad to acknowledge the cooperation of Mr. J. G. Walker in growing and processing the single crystals used,[7] and in preparing also the polycrystalline specimens.

## References

1. H. J. McSkimin, *Jl. Acous. Soc. Amer.*, *22*, 413 (1950), particularly method K.
2. R. M. Bozorth, W. P. Mason, H. J. McSkimin, J. G. Walker, *Phys. Rev.*, *75*, 1954 (1949).
3. Summarized by (a) R. Becker, and W. Döring, Ferromagnetismus, Springer, Berlin (1939), and by (*b*) R. M. Bozorth, Ferromagnetism, Van Nostrand, New York (1951).
4. W. P. Mason, *Phys. Rev. 83*, 683 (1951).
5. H. J. Williams and R. M. Bozorth, *Phys. Rev.*, *59*, 939 (1941), and reference 3b.
6. H. J. Williams and J. G. Walker, *Phys. Rev. 83*, 634 (1951).

7. J. G. Walker, H. J. Williams and R. M. Bozorth, *Rev. Sci. Instruments, 20,* 947 (1949) and reference 2.
8. W. P. Mason, *Phys. Rev. 82,* 715, (1951).
9. W. P. Mason, Piezoelectric Crystals and Their Application to Ultrasonics, Van Nostrand, New York (1950).
10. K. Honda and Y. Shirakawa, Nippon Kinzoku Gakkai-Si (*Jl. Inst. Metals,* Japan) *1,* 217 (1937), and *Sci. Rep. Res. Inst.,* Tohoku Univ. *1,* 9 (1949).
11. M. Yamamoto, *Jl. Inst. Metals,* Japan *6,* 331 (1942) and *Phys. Rev., 77,* 566 (1950).
12. This method is a modification of one described in "Electromechanical Transducers and Wave Filters," W. P. Mason, D. Van Nostrand (1942) pages 244–247. For frequencies above 20 kc., 45° Z-cut ammonium dihydrogen phosphate crystals were used. Below 20 kc the crystals were cemented on square brass rods which were placed at the ends of the nickel rods in place of the piezoelectric crystals alone.
13. W. Döring, *Z. Physik, 114,* 579 (1939).

# Hot Electrons in Germanium and Ohm's Law

By W. SHOCKLEY

The data of E. J. Ryder on the mobility of electrons in electric fields up to 40,000 volts per cm are analyzed. The mobility decreases many fold due to the influence of scattering by optical modes and due to increases of electron energy. It is estimated that electron "temperatures" as high as 4000°K have been produced in specimens having temperatures of atomic vibration of 300° K. The critical drift velocity above which there are deviations from Ohm's law is about $2.6 \times 10^6$ cm/sec. This is three times higher than the elementary theory and an explanation in terms of complex energy surfaces is proposed.

TABLE OF CONTENTS



## 1. INTRODUCTION: FUNDAMENTAL DEVIATIONS FROM OHM'S LAW

THE starting point of many branches of physics is a linear relation. Among the most prominent of these are Hooke's law, which relates stress and strain for solid bodies, Newton's second law of motion F = ma and Ohm's law. In all of these cases, the linear relation is only an approximation that may be regarded as the first term in a Taylor's expansion of the functional relationship between the two variables. Important physical principles are brought to attention when the nonlinear range is reached.

Of the three laws mentioned, Newton's is, of course, the one in which the failure of linearity is the most significant representing as it does the entrance of relativistic effects into the laws of motion.

The failure of Hooke's law may be of either a primary or secondary form.

If a solid contains voids, then under a certain pressure it will crumble and fill the voids. This is a secondary effect. If the sample is homogeneous, however, high pressures will produce fundamental deviations from Hooke's law, these deviations arising from the nonlinearity of the forces between atoms. Studies of these nonlinear effects by Bridgman have, among other things, put on a firmer basis the understanding of the forces between ions in ionic crystals and the pressures of electron gases in metals.

Deviations from Ohm's law for electronic conduction in semiconductors are almost the rule rather than the exception, but the most familiar cases are secondary rather than primary. The primary linear relation for the conduction process is that between the drift velocity of an electron, or hole, and the electric field that drives it. This relationship is

$$v_d = \mu_0 (T) E, \tag{1.1}$$

where the mobility $\mu_0(T)$ is a function of the temperature $T$ of the specimen. By a *fundamental deviation from Ohm's law* we shall mean a deviation in this linear relationship arising from the largeness of $E$ rather than other causes.

Thermistor action is typical of a secondary deviation from Ohm's law. A thermistor is usually a two-terminal circuit element in which the current flows through an electronic semiconductor. The semiconductor has the property that its resistance decreases rapidly as the temperature increases; and the physical basis for this decrease is an increase in the number of conducting electrons (or holes or both) with increasing temperature. The passage of current heats the thermistor and its resistance changes; consequently the linear relation between current and voltage fails and in fact there may result a decrease of voltage with increasing current so that a differential negative resistance is observed. The electric fields are so low, however, that equation (1.1) is valid provided the dependence of $\mu_0$ on the temperature is taken into account. An experimental proof that no fundamental deviation of Ohm's law occurs is furnished by applying a small a-c. test signal on top of a d-c. bias that produces heating. If the frequency is much higher than the thermal relaxation rate, the a-c. resistance is found to be simply that expected for the observed temperature.

The principal nonlinearities of crystal rectifiers, or varistors, and of transistors are also secondary and are associated with changing numbers of current carriers.

In this article we shall discuss some experimental evidence of fundamental deviations in Ohm's law for electrons in $n$-type germanium obtained by E. J. Ryder of Bell Telephone Laboratories.[1] We shall describe his ex-

[1] E. J. Ryder and W. Shockley, *Phys. Rev. 81*, 139 (1951).

perimental techniques and results briefly in the next section and shall then present some aspects of the quantitative theory that explains them, leaving the bulk of the mathematical manipulations for the appendices.

Before discussing Ryder's results, we may indicate why his procedure succeeded whereas previous attempts, of which there have apparently been a number, largely unpublished, have failed. Ryder's work takes advantage of three factors: (1) the availability of electrical pulses of microsecond duration, (2) the high resistivity of germanium, and (3) the high mobility of electrons in germanium. Because of (3), it is possible to deliver energy to electrons at relatively high rates by electric fields. In effect this "heats" the electrons above the temperature of the crystal and lowers their mobility. The generalized equation is then

$$v_d = \mu(T, E)E \tag{1.2}$$

where the fact that $\mu$ depends on $E$ represents the fundamental nonlinearity. We shall show that Ryder's techniques raise the "temperature" of the electrons by a factor of about thirteen fold to above 4000°K. Since the resistivity is high, say 10 ohm cm, the power delivered to the specimen is sufficiently low that the heating in one pulse is negligible. The pulse repetition rate is then kept so low that accumulated heat is negligible also.

These conditions are enormously more favorable than those met with in metals. In a metal the average electron energy is several electron volts; in order to double this energy, each electron would acquire an added energy roughly equal to the cohesive energy per atom of the crystal. Furthermore, in a metal there is about one conduction electron per atom, compared with $10^{-7}$ per atom in Ryder's samples. Thus the stored energy due to "hot" electrons in a metal would be enough to vaporize it, whereas in germanium, or a similar semiconductor, a temperature of 10,000°K for the electrons would be enough to raise the crystal less than 0.01°K. From this reasoning it appears that it will be extremely difficult, if not impossible, to produce significant fundamental deviations from Ohm's law in metals and certainly impossible to produce effects of the magnitude described below.

It should be pointed out that the behavior of electrons in crystals in fields so high that equation (1) fails have been subject to both experimental and theoretical investigation in connection with dielectric breakdown.[2] The work does not apply to cases in which the specimens obey Ohm's law at low fields, however, and the experiments do not permit accurate deter-

[2] See, for example, H. Fröhlich and F. Seitz, *Phys. Rev. 79*, 526 (1950) and F. Seitz, *Phys. Rev. 76*, 1376 (1950). Much of the treatment presented in the Appendices is essentially equivalent to that given in Seitz. In our Appendices, however, we give much more emphasis to the low field case. The Seitz paper also contains a review of the literature to which the reader is referred.

minations of $v_d$ as a function of $E$. From the theoretical side also the emphasis has been on fields so high that the linear range is neglected so that the transition from linear to nonlinear is not stressed.

The current theories of dielectric breakdown are based on the principle of "secondary generation" or "electron multiplication." Thus if an electron acquires enough energy from the electric field, it will be capable of producing secondaries by collision with bound electrons, and the repetition of this process will lead to an avalanche. Our theory indicates that in germanium, even at fields as high as 200,000 volts/cm, few electrons will have enough energy to produce secondaries. At about those fields, however, another phenomenon occurs.

In 1934 C. Zener[3] proposed that dielectric breakdown was due to a primary effect: the field induced generation of hole-electron pairs. His mathematical theory is similar to that for field emission from cold metal points and to that for radioactive decay. It involves the "tunnelling" of electrons through regions in which their wave functions are attenuated, rather than running, waves.

Zener's theory does not seem to apply to breakdown; however, it does apply to the high electric fields produced in rectifying $p$-$n$ junctions in germanium when these are biased in the reverse direction. Under these conditions fields of the order of 200,000 volts/cm are produced. The mobilities of electrons, or holes, in these fields have not been measured. It has been shown,[4] however, that secondary production is very small. At these fields a sort of "breakdown" effect occurs and above a critical value of the voltage a very rapid increase in current is observed. This current appears to be of the nature predicted by Zener. It is stable at a given voltage, has a small temperature coefficient and will probably be useful in semiconductor analogues of "voltage regulator tubes" and protective devices.

As is shown in the treatment given in qualitative terms in Section 3 and in more detail in the Appendices, the explanation of the fundamental deviations from Ohm's law is based on the theory of electron waves. The investigations described in this paper may thus be regarded as furnishing evidence for the wave nature of conduction electrons in germanium and are thus related to the researches of C. J. Davisson, to whom this volume is dedicated, and his collaborator, L. H. Germer. The Davisson-Germer experiments were concerned chiefly with electron waves in free space and with high energy electrons in crystals. Both of these cases are simpler than that dealt with in this paper. Electrons in the conduction band in germanium appear to behave as though they were in a multiply refracting medium in which they may have

[3] C. Zener, *Proc. Roy. Soc. 145*, 523 (1934).
[4] K. C. McAfee, E. J. Ryder, W. Shockley and M. Sparks, *Phys. Rev. 83*, 650 (1951).

several velocities of propagation for any specified direction of propagation and frequency. It is to be hoped that more detailed analyses of the data obtained by Ryder, together with quantitative interpretations of certain observations of magnetoresistance, may lead to a unique evaluation of the "refractive constants" of the medium for the electron waves; this possibility is discussed briefly in Section 5. The phenomena in the Zener current range afford another new opportunity to study electron waves in crystals. The waves involved in this effect are those with energies in the energy gap between the conduction band and the valence band; waves in this range have received little attention from either the experimental or theoretical side.

## 2. E. J. Ryder's Results

One of germanium's most noted attributes is its ability to give amplification of electrical signals when made into a transistor. The basic phenomenon for many types of transistors is that of "carrier injection." As is

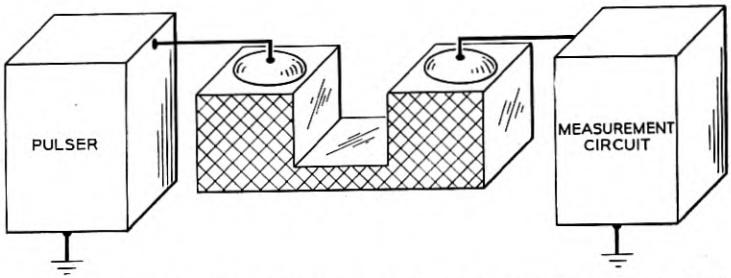



Fig. 1—The principles of E. J. Ryder's technique for observing conductivity in high electric fields.

well known, germanium may carry current either by the electron mechanism in which case it is called *n*-type germanium, or by the mechanism of hole conduction in which case it is called *p*-type. If a suitably prepared electrode is placed on an *n*-type specimen and current is caused to flow in the sense that removes electrons from the specimen, then the process may cause "hole injection." In this case in addition to removing conduction electrons from the germanium, electrons are removed from the valence bonds so that holes are injected. This leads to nonlinear effects because, as the current passes through the specimen, the number of carriers in the specimen changes and so does its resistivity.

In order to avoid the secondary deviations from Ohm's law due to carrier injection, Ryder has designed specimens of the form shown in Fig. 1. These specimens have large ends to which the metal electrodes are attached. The resistance arises chiefly from a thin section of the material connecting the

large ends. Since the fields at the large ends are small, carrier injection is largely suppressed; furthermore, the electric fields are applied for such a short time during the pulse that, even if holes were injected at one of the ends, they would not have time to reach the narrow section of the bridge and modulate its conductivity during the period of the pulse.

Further causes of non-linearity can arise from inhomogeneities in the germanium material itself. For example grain boundaries in polycrystalline germanium are known to have added electrical resistance. Difficulties due

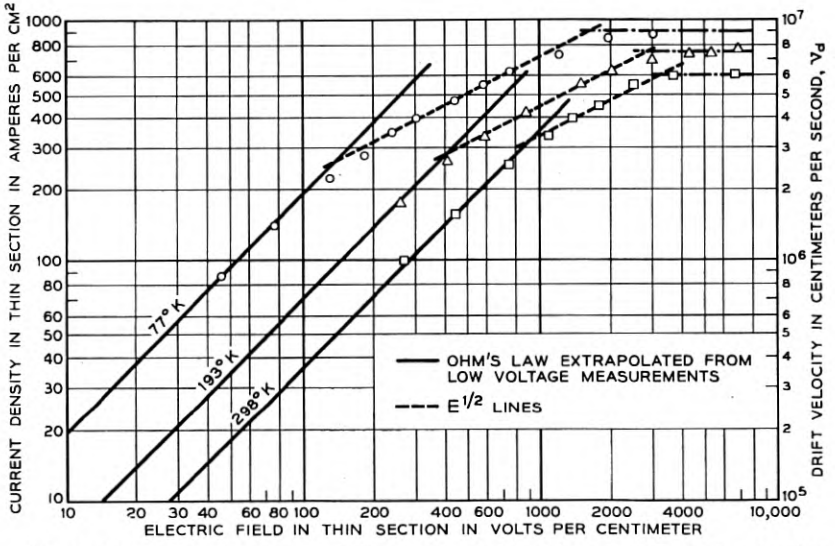



Fig. 2—Currents and estimated drift velocities deduced from E. J. Ryder's pulse data on a specimen of *n*-type germanium of 2.7 ohm-cm resistivity. [The fact that the numerical values of current density and drift velocity have the same digits is a consequence of the accident that $\sigma = 1/2.7 = 0.37$ is almost exactly $10^{-4}$ times the mobility.]

to inhomogeneity have largely been eliminated in these experiments by the use of highly homogeneous single-crystal germanium material furnished by G. K. Teal and his collaborators.

Other experimental precautions are necessary, such as assuring a smooth polished surface on the filament; if this is not done, apparently holes are injected from the surface irregularities of the thin section between the large ends. It is also necessary to make corrections for end effects since some of the resistance arises within the large blocks themselves.

Some of the data obtained by Ryder are shown in Fig. 2. The drift velocity, plotted as ordinate, is not measured directly but is inferred from the measured currents through the specimen by the following reasoning:

In a specimen at room temperature the drift velocity of electrons is given by the equation

$$v_d = 3600\, E \text{ cm/sec} \tag{2.1}$$

when $E$ is expressed in volts per cm.[5] At 100 volts per cm, for example, (which is below the non-linear range) the velocity of electrons should be $3.6 \times 10^5$. This establishes the drift velocity scale for room temperature. From other measurements of the germanium specimen of Fig. 2, it is concluded that the number of electrons available for conduction is substantially independent of temperature down to liquid air temperatures. Consequently, for this temperature range the drift velocity should be directly proportional to the current in the specimen. The other two sets of data are accordingly simply scaled in proportion to their currents.

On the figure we also show extrapolated lines at 45 degrees corresponding to Ohm's law. From these we see that decreases in mobility of tenfold or more have been produced in these experiments for high field conditions.

The data for each temperature fall approximately on three lines: The low field or Ohm's law region, an intermediate region over which $v_d$ is proportional to $E^{1/2}$ and $\mu$ is proportional to $E^{-1/2}$, and a saturation region. The break at low fields comes at drift velocity of about $3 \times 10^6$ cm/sec for all three cases. This break, according to theory, should come when the drift velocity is several times the speed of sound in germanium, the speed of sound being about $5.4 \times 10^5$ cm/sec. The limiting drift velocity at higher fields is associated with the energy required to excite a particular type of atomic vibration, called an "optical mode." It comes at approximately the value of drift velocity predicted by theory.[6] The theoretical curves, computed in the appendices, do not break into the sharp line section suggested on Fig. 2. However, they show the distinct influences of separate causes and fit the data reasonably well, as we shall show below in connection with Fig. 5.

## 3. Theory of Deviations from Ohm's Law

### 3a. *Electrons in n-Type Germanium*[7]

The specimens we shall consider are of $n$-type germanium and have resistivities of several ohm cm. The conductivity arises from the presence

[5] J. R. Haynes and W. Shockley, *Phys. Rev. 81*, 835 (1951).

[6] The observation and explanation of these general features was presented in our first publications: E. J. Ryder and W. Shockley *Phys. Rev. 81*, 139 (1951) and *82*, 330 (1951).

[7] The material under this heading is treated in more detail in the author's book, "Electrons and Holes in Semiconductors," D. van Nostrand (1950), Chapter I. This book will be referred to subsequently as *E and H in S*.

of donors: chemical impurities such as arsenic or antimony. These donors substitute themselves for germanium atoms in the crystal structure, form electron-pair bonds with their four neighbors and release their fifth valence electron to the conduction band. The density of donors is about $10^{15}/cm^3$ or one per cube $10^{-5}$ cm = 1000 Å on an edge. The donors are fixed positive charges and do not move in electric fields. Their charges neutralize those of the electrons. The electrostatic energy of interaction between electrons and donors leads to a deflection of the electron's motion. For the temperatures of Fig. 2, however, this effect is unimportant compared to the effect of thermal vibrations of the atoms.

The electrons in the conduction band move in accordance with a wave equation. They may, however, be thought of as particles. The justification is that, under many experimental conditions, the wave functions will actually be wave packets. These wave packets, it can be shown, behave much as particles and can be dealt with as particles, at least provided the phenomena considered do not involve distances smaller than the size of the wave packet.

Under conditions of thermal equilibrium we may think of the $10^{15}$ electrons in each cubic centimeter as an electron gas with the electrons (as wave packets) moving at random with an average kinetic energy of motion of $(3/2)kT$.

If the atoms of the crystal were held rigidly at rest in a perfectly regular crystal structure, an electron wave, and a wave packet too, would be transmitted through it with no scattering. At 300°K the vibrations are such that the wave packet moves for only 2500 Å before being scattered. At low temperatures, the mean free path is longer and at liquid hydrogen temperatures it is so long that thermal vibrations are less important than the fields of the ionized donors. As stated above, however, we may neglect this scattering by ions over the temperature range of Fig. 2.

Before proceeding with the discussion of the interactions of electrons and thermal vibration we shall point out that two problems must be solved before the dependence of mobility upon electric field can be explained:

First, the mechanisms of the individual processes must be analyzed. This is the basic physical problem. In order to solve it we must apply quantum mechanics to the model representing the electron moving in the crystal and determine the probabilities of various types of transitions and some appropriate averages.

Second, the statistical consequences must be worked out. On the basis of the individual processes, the statistics of the assemblage of electrons must be analyzed and a steady state solution found.

The first problem poses the more physical problems and is given the most

attention. The second problem is more difficult mathematically. It is given only an approximate treatment which is adequate, however, to indicate that the solution to the first problem contains the necessary features to explain the experiments.

3b. *The Phonons*

We must next consider how to describe the thermal vibrations and to evaluate their interactions with the electrons. We shall present only the principal results of the mathematical analysis here, leaving the details for the appendices. The earliest treatment of thermal vibration in a crystal was that of Einstein, who considered each atom to be a separate harmonic oscillator. This model was improved on by Debye who treated the crystal as an elastic continuum that could support running waves. Debye's method is regarded as essentially correct and we, therefore, resolve the atomic motions into a set of running waves, or normal modes. There are three times as many independent normal modes as there are atoms in the crystal, or one per degree of freedom, and any possible atomic motion of the crystal may be made up as a sort of Fourier series in these normal modes.

Each normal mode must be treated as a Planck oscillator and has a system of energy levels with values

$$(n + 1/2)h\nu \tag{3.1}$$

where $\nu$ is its frequency of vibration. Each quantum of energy is referred to as a *phonon*; if a normal mode makes a transition with $\delta n = +1$, we say a phonon has been emitted and if $\delta n = -1$, we say one has been absorbed.

The description of the crystal in terms of phonons is in close analogy with the description of electromagnetic waves in a cavity in terms of *photons*. For the case of light, the electromagnetic state of the cavity is determined by finding the normal modes, which are treated as quantized oscillators, and transitions with $\delta n = \pm 1$ correspond to photon emission and absorption.

The normal modes for the crystal are unlike those for light. For low frequencies the waves are essentially the microscopic transverse and longitudinal waves of a solid. As the wave length becomes shorter, however, the sound velocity varies and there is a limiting minimum wave length which is about twice the spacing between atoms. In order to understand the energy losses of electrons in high fields, we must consider the role of this minimum wave length. For this purpose we shall describe the dependence of frequency upon wave length for a longitudinal mode.

Accordingly we consider the frequency of the normal modes corresponding to a longitudinal wave propagating along a cube axis. Rather than

using the wave length $\lambda$ as a variable, we use the wave number or $(1/\lambda)$. For long waves the frequency is simply

$$\nu = c/\lambda = c(1/\lambda). \tag{3.2}$$

This corresponds to the straight line portion for low frequencies in Fig. 3. This portion extends to a wave length equal to twice the lattice constant $a$ of the crystal.

Figure 3 shows another curve which has a high frequency even for $(1/\lambda) = 0$ or infinite wave length. The presence of this branch of the "vibrational spectrum" is due to the fact that the diamond structure has two

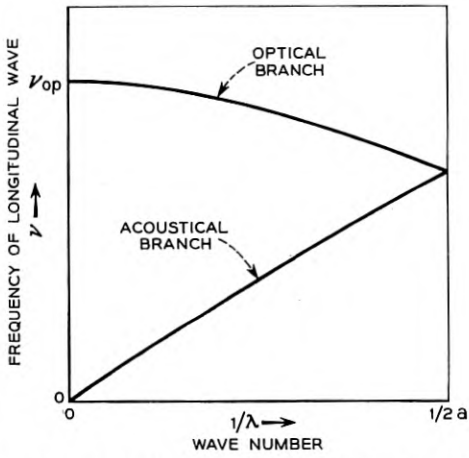



Fig. 3—Frequency of longitudinal vibrations in [100] direction in the diamond structure. (In this particular direction of propagation the acoustical and optical branches join smoothly at the same frequency; for other directions, there is a discontinuity in frequency. The dependence of $\nu$ upon $1/\lambda$ is approximated by a sine wave.)

atoms per unit cell. (The diamond structure is made of two face centered cubic arrays of atoms, juxtaposed so that each atom of one array is centrally situated in respect to a tetrahedron of four atoms of the other array, with which it forms four electron-pair bonds. The unit cell contains one atom of each array.) As a consequence of this it is possible to have a vibration in which one atom vibrates in the plus $x$ direction while the other atom vibrates oppositely and to have this same motion occur in phase in every unit cell. Such a vibration is considered to have infinite wave length, since every unit cell does the same thing at the same time. It has the highest possible frequency since the pattern of motion involves directly opposed motions of nearest neighbors. If the motion is modified so as to have dif-

ferent phases in adjoining unit cells, and thus to correspond to a finite wave length, the frequency drops.

The opposed motions are referred to as "optical modes" by analogy with polar crystals. In a crystal of sodium chloride there is one $Cl^-$ and one $Na^+$ per unit cell. In the opposed type of motion, ions of like sign move one way and opposite to those of the other sign. This relative motion of charge polarizes the crystal and phonons of this type of vibration can absorb or emit light. Because of this optical activity the mode is termed optical.

The name "optical" is carried over to valence crystals to describe the opposed form of motion, although no polarization accompanies the displacement in the latter case.

### 3c. *The Selection Rules*

We shall next consider the laws which govern the interchange of energy between an electron and the phonons. There are two important laws, closely analogous to the laws of conservation of energy and momentum for two masses in collision. The quantity analogous to momentum for the phonon is a vector, called $\vec{P}_\gamma$, directed along the direction of propagation of the phonon, and having a magnitude given by the relationship between momentum and wavelength

$$P_\gamma = h(1/\lambda) = h/\lambda, \qquad \vec{P}_\gamma \parallel \text{propagation direction} \tag{3.3}$$

In a transition in which an electron exchanges energy with the phonons, and changes its momentum from $\vec{P}_1$ to $\vec{P}_2$, so that $\delta n_\gamma = \pm 1$ for one of the modes, one selection rule requires that

$$\vec{P}_2 - \vec{P}_1 + \delta n_\gamma \vec{P}_\gamma = 0. \tag{3.4}$$

This is analogous to conservation of momentum; actually it is based on far more subtle effects. The conservation of energy requires that

$$\mathcal{E}_2 + \delta n_\gamma h\nu_\gamma = \mathcal{E}_1 \tag{3.5}$$

where

$$\mathcal{E}_2 = P_2^2/2m, \qquad \mathcal{E}_1 = P_1^2/2m \tag{3.6}$$

are the electron's energies before and after collision. The mass $m$ need not be the mass of an electron but may instead be the "effective mass," a mass-like quantity of the same order as the electron mass which takes into account the influence of the periodic potential of the crystal structure upon the electron wave packet.

The effective mass concept represents a simplification that may not

necessarily be correct. In a cubic semiconductor, the electron waves can be "refracted" as are the longitudinal and transverse acoustical waves. The deviations from Ohm's law of Fig. 2 furnish evidence that the simplified assumption of equation (3.6) must in fact be replaced by the more general possibility. We shall return briefly to this point in Section 5.

In addition to the conservation of energy and momentum, there are two other approximate selection rules which, while not exact, are so nearly fulfilled that no appreciable error is introduced by using them:

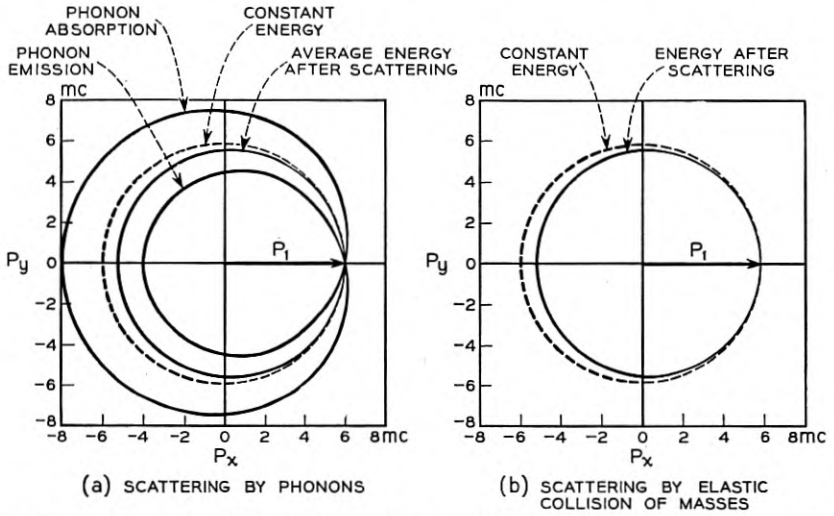



Fig. 4—Comparison of the scattering by acoustical phonons with the scattering of a small mass in elastic collision with a larger mass.

Only $\delta n_\gamma = \pm 1$ is allowed.

For the acoustical modes, only the longitudinal modes interact with electrons. (This restriction does not apply to optical phonons.)

Figure 4 shows the allowed transitions for an electron with initial momentum $P_1$ in the $x$-direction. If the energy of the phonons were zero, the allowed transitions would be to points on the sphere (or circle in Fig. 4) with $P_2 = P_1$. Since the energy of a phonon is

$$h\nu_\gamma = hc/\lambda = cP_\gamma = c\,|\,\vec{P}_2 - \vec{P}_1\,|\,, \tag{3.7}$$

however, the end points lie on the surfaces shown.

These surfaces do not differ much from the sphere, as may be seen by considering the final energy for an electron that reverses its motion by

phonon absorption. For this case

$$P_\gamma = P_2 + P_1 \tag{3.8}$$

and

$$(P_2^2 - P_1^2)/2m = cP_\gamma \tag{3.9}$$

so that the change in magnitude of momentum is

$$P_2 - P_1 = 2mc. \tag{3.10}$$

For an electron with energy $kT$ and "thermal velocity"

$$v_T = (2kT/m)^{1/2}, \tag{3.11}$$

corresponding to $10^7$ cm/sec at $300°K$, the fractional change in momentum is

$$(P_2 - P_1)/P_1 = 2c/v_T = 2 \times 5.4 \times 10^5/10^7 = 0.11. \tag{3.12}$$

Thus the phonon absorption surface lies only 11% outside the constant energy sphere. Figure 4 is drawn for the case of $P_1 = 6\ mc$, or $v_1 = v_T$ for 32°K, for purposes of exaggerating the differences in the surfaces. (A further discussion of Fig. 4 is given in the appendices.)

For transitions with optical phonons, in the range of interest, $h\nu_\gamma$ is nearly independent of $P_\gamma$. Furthermore, the optical phonons have $h\nu_\gamma \doteq k\ 520°K$ so that they are nearly unexcited at room temperature and have $n_\gamma = 0$ so that only $\delta n_\gamma = +1$ is allowed. For this case transitions can occur only if $\mathcal{E}_1$ is greater than $h\nu_{op}$ and the end surface is a sphere with

$$\mathcal{E}_2 = \mathcal{E}_1 - h\nu_{op}. \tag{3.13}$$

We shall neglect the role of the optical phonons until after a comparison between acoustical phonon processes and experiment has been made. We shall then show that they play an essential role in explaining Ryder's data.

### 3d. *Energy Exchange and The Equivalent Sphere Problem*

We shall here give in brief some results derived in the Appendices which permit us to show the equivalence of the problem of *acoustical phonon* scattering to a problem in gas discharges. This has two advantages: it enables us to take over the solution to the statistical problem from gas discharge theory, the second problem mentioned at the end of Section 3a, and to concentrate on the problem of the mechanism. In addition the equivalence makes it much easier to visualize the mechanism of energy losses.

According to the theory of phonon scattering, an electron is equally likely to be scattered from its initial direction of motion to any other. This implies that after an interaction the electron is equally likely to end in any unit

area of the surfaces of Fig. 4. The probability of being scattered per unit time is simply

$$1/\tau_1 = v_1/\ell \tag{3.14}$$

where $v_1$ is the speed and $\ell$ the mean free path; according to the theory $\ell$ is a function of the temperature $T$ of the phonon system and is independent of $v_1$. The time $\tau_1$ is the mean free time or average time between collisions.

Figure 4 shows the average energy after collision. The Figure represents a case in which the average energy is somewhat smaller than the initial energy. This will be the case for a high energy electron, that is one with an energy greater than $kT$ for the acoustical modes. The average loss in energy for a high energy electron is found to be

$$\langle \delta \mathcal{E} \rangle = -c^2 P_\gamma^2/2kT \tag{3.15}$$

where $P_\gamma$ is the momentum change in the collision. This formula is analogous to the formula for energy loss if a light mass $m$ strikes a heavy stationary mass $M$ and transfers a momentum $\vec{P}_2 - \vec{P}_1 = \vec{P}_\gamma$ to it. The energy transfer is given by (3.15) if

$$M = kT/c^2 \tag{3.16}$$

since then the kinetic energy of the large mass is simply

$$P_\gamma^2/2M = c^2 P_\gamma^2/2kT. \tag{3.17}$$

The value of the mass which satisfies equation (3.16) for room temperature may be calculated from the previously quoted values of $v_T$ and $c$:

$$M = kT/c^2 = mv_T^2/2c^2 = 170m, \tag{3.18}$$

a value which may certainly be considered large compared to $m$.

Equation (3.15) is not the complete expression for average energy change for a collision with momentum change $P_\gamma$ and another term representing energy gain also occurs. If the complete expression is averaged over all final directions of motion, it is found that the average change of energy, which is obviously the average energy change per collision, is

$$\langle \delta \mathcal{E} \rangle = 4mc^2(1 - P_1^2/4mkT) = (4mkT/M) - (P_1^2/M). \tag{3.19}$$

This is the correct expression for the average gain in energy per collision of a light mass $m$ colliding with a heavy mass $M$ which is moving with the thermal energy appropriate to temperature $T$. This corresponds to a thermal velocity of

$$v_{TM} = (2kT/M)^{1/2} = 2^{1/2}c \tag{3.20}$$

for the large mass. The second term in (3.19) is just the average of (3.15) over all directions of motion after collision and represents the energy loss that would arise if $M$ were initially stationary. The first term represents an energy transfer from $M$ to $m$ due to the thermal motion of the large mass.

Furthermore, if the light and heavy masses are perfectly elastic spheres, the scattering of $m$ will be isotropic, just as is the case for the phonons. This shows that there is an almost perfect correspondence between the two mechanisms of scattering so that we are justified in using previously derived results for the sphere case and applying them to the phonon case.

To complete the equivalence we should introduce a density of large spheres so as to get the correct mean free path. There is nothing unique about this procedure, as there is about the mass $M$ and temperature $T$, and we may make a large selection of choices for number of $M$-spheres per unit volume and radii of interaction so as to obtain the desired mean free path. Once any choice is made, of course, it will give the same mean free path independent of electron energy and may be held constant independent of the electric field.

### 3e. *Acoustical Phonons and Electric Fields*

We shall next give a very approximate treatment of mobility in low and high electric fields. The emphasis will be upon the interplay of the physical forces, the mathematical details being left to the Appendices or to references.

In the Ohm's law range, the field $E$ is so small that the electrons have the temperature of the lattice. They have a velocity of motion of approximately

$$v_T = (2kT/m)^{1/2} \tag{3.21}$$

and a mean free time between collisions of

$$\tau = \ell/v_T . \tag{3.22}$$

The electric field accelerates the electron at a rate

$$a = qE/m \tag{3.23}$$

and imparts a velocity $a\tau$ in one mean free time. Since the collisions are spherical, the effect of the field is wiped out after each collision. The drift velocity is thus approximately

$$v_d = a\tau = (q\ell/mv_T)E. \tag{3.24}$$

An exact treatment which averages over the Maxwellian velocity distribution gives a value smaller by 25% and leads to

$$\mu_0 = 4q\ell/3\pi^{1/2}mv_T . \tag{3.25}$$

Since theory shows that $\ell$ varies as $T^{-1}$, the mobility should vary as $T^{-3/2}$. This prediction is in good agreement with experimental findings over the range of conditions for which the dominant scattering processes are those considered here.

Next we consider the effect of very large fields. Under these conditions an electron drifting in the direction of the field with drift velocity $v_d$ acquires energy from the field at an average rate

$$(d\mathcal{E}/dt)_{\text{due to } E} = v_d q E. \tag{3.26}$$

If this power is large enough, the electrons will be unable to dissipate energy sufficiently rapidly to the phonons that they can maintain their normal temperature. As a result their average energy mounts, after the field is initially applied, until they can furnish energy to the phonons fast enough to maintain a steady state. Under these conditions the sum of the two rates is zero

$$(d\mathcal{E}/dt)_{\text{due to } E} + (d\mathcal{E}/dt)_{\text{due to phonons}} = 0. \tag{3.27}$$

If the field is high enough, there may be no steady state solution. This can occur if the ability of the phonons to remove energy decreases with increasing energy. Such cases play an essential role in the theory of dielectric breakdown.[8] In them it is concluded that electrons will gain sufficient energy from the field so that they can produce secondary electrons which repeat the process thus producing avalanches. For the cases with which we are concerned, theory indicates that the energy losses increase rapidly with the energy of the electron while the power input decreases because of decreasing mobility so that a steady state will thus occur.

In order to estimate the drift velocity for the steady state we must introduce expressions for the two powers involved. For this purpose we assume that an electron has on the average a speed $v_1$ and we calculate the power to phonons as the average energy loss per collision for this velocity times the rate of collision, $v_1/\ell$. For $v_1 \gg v_T$, we can neglect the effect of motion of the $M$ spheres and thus obtain from (3.19)

$$(d\mathcal{E}/dt)_{\text{phonons}} = -(v_1/\ell)\, m^2 v_1^2/M. \tag{3.28}$$

The mobility will be less because of the higher collision rate so that the drift velocity in the field will be approximately

$$v_d = (q\ell/mv_1)E. \tag{3.29}$$

The power furnished by $E$ will be

$$(d\mathcal{E}/dt)_{\text{due to } E} = (q^2\ell/mv_1)E^2. \tag{3.30}$$

[8] See the references to Fröhlich and Seitz in Section 1.

The steady state condition then leads to

$$v_1 = (q\ell E/m)^{1/2}(M/m)^{1/4} \tag{3.31}$$

and to

$$\begin{aligned} v_d &= (q\ell E/m)^{1/2}(m/M)^{1/4} \\ &= (\sqrt{2}cq\ell E/mv_T)^{1/2} \cong (c\mu_0 E)^{1/2}. \end{aligned} \tag{3.32}$$

The treatment[9] based on accurate statistics for the equivalent sphere model leads to

$$v_d = 1.23(c\mu_0 E)^{1/2}. \tag{3.33}$$

The transition between the high field behavior and low field behavior should occur in the neighborhood of a critical field $E_c$ at which both limiting forms give the same $v_d$ :

$$v_d = \mu_0 E_c = 1.23(c\mu_0 E_c)^{1/2}, \tag{3.34}$$

leading to

$$E_c = 1.51\ c/\mu_0 \tag{3.35}$$

and to a drift velocity, which shall be referred to as the *critical velocity*, of

$$v_{dc} = 1.51c \tag{3.36}$$

if Ohm's law held to a field as high as $E_c$. The drift velocity can be expressed in terms of $E_c$ by the equation

$$v_d = \mu_0(EE_c)^{1/2} \tag{3.37}$$

for values of $E$ much greater than $E_c$.

It is interesting to note that this initial field is just that which would give electrons a drift velocity corresponding to the thermal motion of $M$-masses. This seems a natural critical field. For it the effect of random motion of $M$ would be suppressed by the systematic drift velocity so that the transfer of thermal energy to the electrons would be much reduced. This value of $E_c$ corresponds to much smaller initial fields than are sometimes proposed. For example, one frequently encounters proposals that Ohm's law should hold up to the condition that $v_d = v_T$. This would correspond to 10 times higher field at 300°K than that obtained. Another criterion is that the energy gained in one mean free path, $q\ell E$, should be equal to $kT$. This is substantially equivalent and corresponds to $v_d = v_T/2$.

[9] *Druyvesteyn Physica 10*, 61, 1930. This paper is reviewed by S. Chapman and T. G. Cowling in "The Mathematical Theory of Non-Uniform Gases," Cambridge at the University Press, 1939, page 347. (The factor is 0.897 $(18\pi/8)^{1/4} = 1.23$.)

For comparison with experiment we note that for a value of

$$E = 4E_c = 6.04c/\mu_0 \tag{3.37}$$

such that if Ohm's law held

$$v_d = 6.04c, \tag{3.38}$$

the value of $v_d$ should be less than half the value predicted by Ohm's law. We shall shortly discuss the discrepancy between this prediction and experiment.

The "temperature" of the electrons may be conveniently expressed in terms of the ratio $E/E_c$. Since the electrons for the high field case are not in a Maxwellian distribution of velocities, one cannot define their "temperature" unambiguously. As a measure of their temperature we shall take their average kinetic energy divided by $k$. This leads to a ratio of electron temperature $T(E)$ to crystal temperature $T$ of $2v_1^2/3v_T^2$. Since to a first approximation the ratio of mobilities at low and high fields is $4v_1/3\pi^{1/2}v_T$, the ratio of temperatures is

$$\frac{T(E)}{T} = \frac{3\pi}{8}\left[\frac{\mu_0}{v_d(E)/E}\right]^2 = \frac{3\pi E}{8E_c}. \tag{3.40}$$

This ratio may also be thought of in terms of the square of the ratio of drift velocity on the extrapolated Ohm's law line to the drift velocity on the $E^{1/2}$ line:

$$T(E)/T = (3\pi/8)[\mu_0 E/v_d(E)]^2. \tag{3.41}$$

Either of these equations may be used to estimate electron temperature from the data in the range in which the $E^{1/2}$ formula is a good approximation.

## 4. Comparison Between Theory and Experiment

### 4a. *Discrepancy in Critical Field*

In Fig. 5 we repeat Ryder's data of Fig. 2 together with data on an additional sample at 77°K. This new sample is considered more reliable than the first since its low field resistivity varies in just the proper ratio [see (3.25) and subsequent text] of $(298/77)^{3/2}$ compared to its value at room temperature. Also we show the theoretical curves that will be discussed below.

The deviations of the data from Ohm's law do not occur at fields as low as those predicted in Section 3e. For $c = 5.4 \times 10^5$ cm/sec., the critical drift velocity should be

$$v_{dc} = 1.51c = 8.2 \times 10^5 \text{ cm/sec.} \tag{4.1}$$

It is seen that Ohm's law is followed to several times higher velocities with negligible deviations. The deviations should be a factor of 2 at the field corresponding to

$$v_d = 6.04c = 3.26 \times 10^6 \text{ cm/sec.} \tag{4.2}$$

on the Ohm's law line. The deviations are actually much less.

Another important difference between the data and the theory of Section 3 is that the experimental points do not continue on a straight line with slope 1/2 but instead tend to flatten out with a roughly constant drift velocity.

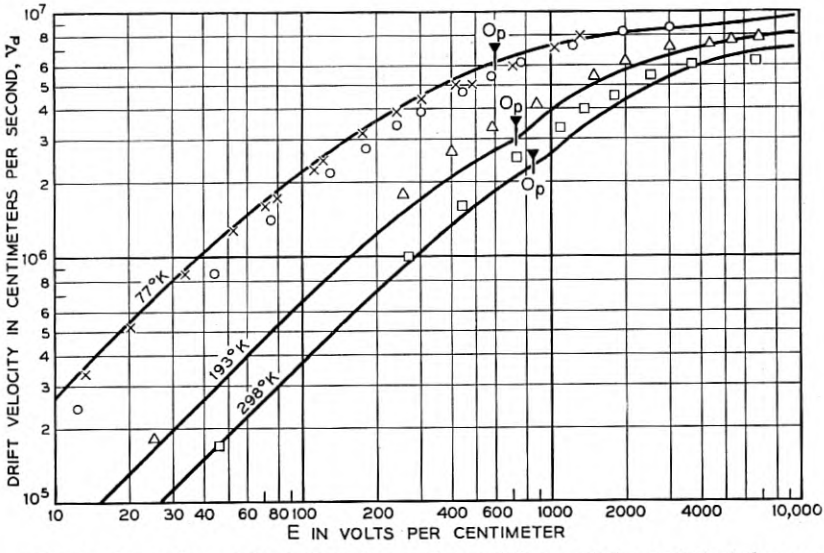



Fig. 5—Comparison of E. J. Ryder's experimental data and the statistical theory of Appendix A7.

Three theoretical curves are shown. These are based on an approximate treatment that includes the effect of the optical transitions. Due to the approximation, the optical modes are neglected below the points marked Op on the Figure. This approximation also leads to a discontinuity in slope at these points; in a more accurate treatment, this bump would be smoothed out. The optical modes play the least role for the curve at 77°K and for this theory fits experiment within experimental accuracy if a value of

$$v_{dc} = 2.6 \times 10^6 \text{ cm/sec.} \tag{4.3}$$

is used. This value is 3.2 times larger than the value given by equation (3.36) using $c = 5.4 \times 10^5$ cm/sec, the value appropriate for longitudinal phonons.[10] The interpretation of this discrepancy, which we refer to as the "low field" discrepancy, is discussed in Section 5. It does not, of course, imply an error in the value of the sound velocity, but instead an error in the theory leading to the formula for critical velocity in terms of sound velocity.

Although an exact theory along the lines discussed in Section 5 has not been developed, it appears evident that its chief effect will be to increase energy interchange with the phonons by a factor of 3.2 squared or approximately 10. This increase can be effected in a mathematically equivalent way by introducing an *effective velocity* for dealing with phonon energies which is 3.2 times larger than the true velocity of longitudinal waves. The approximate theory in the Appendices uses this procedure.

We may remark in passing that only two constants were arbitrarily chosen to fit the curves to the data. One of these was the effective velocity $c = 1.73 \times 10^6$ cm/sec. which is 3.2 times larger than the speed of longitudinal waves. The other was the mobility of electrons at room temperature. Three other constants were chosen from independent estimates of the properties of the crystal. One of these is $h\nu$ for the optical modes for which a value of $k520°K$ was used; another is the effective electron mass, for which the free electron mass was used; and a third was the interaction constant for optical modes, which was set equal to that for the acoustical modes. The meaning of these terms is discussed in the Appendices.

### 4b. *The Effect of the Optical Modes*

We shall next discuss briefly the role of the optical modes before remarking on a theory of the low field discrepancy.

As discussed above the optical modes can act only if $\mathcal{E}_1 > h\nu_{op}$. Theory indicates, however, that when they do come into action they are much more effective than the acoustical modes. On the basis of these ideas, we can see how they can act to give a limiting drift velocity that does not increase with increasing electric field. For purposes of this illustration we shall imagine that so high an electric field is applied that an electron may be accelerated from $P_1 = 0$ to $P_2 = (2mh\nu_{op})^{1/2}$, at which its energy equals $h\nu_{op}$, in so short a time that it is not scattered by acoustical modes. As soon as it reaches $P_2$, we assume that it is scattered by the optical modes, loses all its energy and returns to zero energy. This process then repeats, the period being

[10] This is the velocity of longitudinal waves in the [110] direction as reported by W. L. Bond, W. P. Mason, H. J. McSkimin, K. M. Olsen and G. K. Teal, *Phys. Rev. 73*, 549 (1948). See "Electrons and Holes in Semiconductors," page 528, for the reason for using this wave.

$P_2/qE$ since $dP_2/dt = qE$. The average momentum is evidently $P_2/2$ and the average velocity is

$$v_d = P_2/2m = (h\nu_{op}/2m)^{1/2} \tag{4.4}$$

and is independent of $E$. The optical modes correspond to a temperature of about 520°K and this leads to

$$v_d = 6.3 \times 10^6 \text{ cm/sec.} \tag{4.5}$$

in general agreement with the observed value.

The theory in the appendices indicates that both optical and acoustical modes are active simultaneously and their interplay leads to the theoretical curves shown. (In the Appendices a further discussion is presented and some additional data are compared with theory.)

The tendency of the theoretical curves to fall below the data for 193°K and 298°K for field values below the optical point is thought to arise largely from the approximations employed in the theory. The approximations neglect the ability of the optical modes to enable the electrons to lose energy for fields below the indicated value. Actually some electrons will be scattered by the optical modes and this will contribute in an important way to holding the temperature down and the mobility up. A correct treatment would, therefore, raise the theoretical curve appreciably in the region where it deviates most from the data.

### 4c. *Electron "Temperatures"*

From the theory it follows that the average electron energies correspond to about 520°K at the points marked Op on Fig. 5. The highest point on the 298°K curve corresponds to ~700°K and the highest point on the 77°K curve corresponds to 550°K. For this last case the electron temperature is more than seven times as high as that of the atomic vibrations. In the appendix we quote some other earlier data of Ryder's that indicates electron temperatures of about 4000°K while the crystal itself remains at room temperature.

## 5. An Explanation of the Low Field Discrepancy

The failure to deviate from Ohm's law at the low fields predicted indicates that the electrons can dissipate their excess energy more effectively than would be expected on the basis of their mobility. This conclusion is forced on us by the observation that they apparently retain their thermal distribution and normal mobility to higher fields than predicted. It is not possible to explain the discrepancy by assuming a large or a small value for the effective mass, since the value of the effective mass does not enter into the final comparison with experiment.

It is possible, however, to explain the discrepancy by assuming that the effective mass is not single valued. This assumption corresponds to the case in which the surface in the Brillouin zone belonging to a single energy is not a sphere but instead a complex surface of two or three sheets. Such surfaces have been found as a result of numerical calculations for certain crystals[11] and it has also been shown that such surfaces are to be expected in general[12] if the energy at the bottom of the conduction band is degenerate. It appears necessary to assume that such complex surfaces occur in order to explain magnetoresistance effects.[13]

In terms of Fig. 4, this theory replaces the circular energy contours by deeply re-entrant curves. Transitions from peak to peak of the curves result in large energy transfers to the phonons and hence more effective energy losses. This effect can occur without a compensating change in the effective mass involved in the mobility and, as a result, the critical field may be increased by a large factor. A preliminary analysis indicates that in order to increase the critical field by a factor of 3 a value of about 3 is also required for the ratio of maximum to minimum momentum for the energy surface. A similar analysis of magnetoresistance leads to a factor about 50% larger in order to account for the increases in transverse resistance of about 7-fold observed by Suhl.[14] At the time of writing, therefore, the author feels that both the critical field data at low fields and the magnetoresistance data require a modification of the effective mass picture and that the same modification may well explain both sets of data.

I am indebted to E. J. Ryder, whose experimental results provoked the analysis presented in this paper, to F. Seitz and J. Bardeen for several helpful discussions, to Gregory Wannier for an introduction to the analogous case in gas discharge theory and to Esther Conwell for help with the manuscript.

I shall also take this opportunity to express my appreciation to C. J. Davisson. The opportunity to work in his group was a large factor in my decision to come to Bell Telephone Laboratories, where I enjoyed his stimulating companionship while assigned to his group, and later as well.

## APPENDICES

### *A.1* Introduction and Notation

The problem of energy exchange between the electrons and the phonons requires a somewhat more sophisticated treatment than does the problem of mobility at low fields. In order to present the theory of energy exchange,

[11] W. Shockley, *Phys. Rev. 50*, 754 (1936).
[12] W. Shockley, *Phys. Rev. 78*, 173 (1950).
[13] W. Shockley, *Phys. Rev. 79*, 191 (1950).
[14] H. Suhl, *Phys. Rev. 78*, 646 (1950). Suhl finds increases in resistance in transverse fields as high as 7-fold.

it is necessary to reproduce a large amount of the material dealt with in ordinary conductivity theory. We do this in a somewhat abbreviated form expanding the exposition on the points particularly pertinent to the theory of energy losses.

For convenience we reproduce here a number of the more important symbols. The references indicated refer to places where they are discussed in the text.

$a$ = lattice constant; Fig. 3.
$c$ = speed of longitudinal acoustical wave; Equ. (3.2).
$c_{ll}$ = average longitudinal elastic constant; Equ. (A4.1).
$e$ = base of Naperian logarithms.
$E$ = electric field.
$\mathcal{E}$ = energy.
$\mathcal{E}_{1n}$ and $\mathcal{E}_{2n}$; Equ. (A4.1) and (A7.9).
$h = 2\pi\hbar$ = Planck's constant.
$k$ = Boltzmann's constant.
$\ell$ = mean free path for electron due to scattering by acoustic phonons; (A4.3).
$\ell_{op}$ = describes scattering by optical phonons; (A7.19).
$m$ = effective mass of electron.
$M$ = mass in equivalent mass treatment; (A5.8).
$\vec{P}$ = "crystal momentum" of electron = $h$ times its wave number.
$V$ = volume of crystal.
$\Delta$ = dilation; (A4.1) and (A7.10).
$\nu$ = frequency of normal mode.
$\nu_{op}$ = frequency of optical mode (used in Section 4 only); Equ. (4.5).

### *A.2* The Probability of Transition into Energy Range $\delta\mathcal{E}_2$

In this section we consider an electron initially with energy $\mathcal{E}_1$ and momentum $\vec{P_1}$, which for convenience we take to be along the $P_z$-axis, and we evaluate the probability that it make a transition to states with energies in the range $\mathcal{E}_2$ to $\mathcal{E}_2 + \delta\mathcal{E}_2$. We shall assume that the crystal is elastically isotropic so that for the spherical energy surface approximation employed, i.e. equation 3.6, the scattering will be symmetrical about the $P_z$-axis. The end states, $\vec{P_2}$, may, therefore, be considered in groups lying in the range $d\mathcal{E}_2$, $d\theta$ where $\theta$ is the angle between $\vec{P_1}$ and $\vec{P_2}$. These states lie in a ring in $\vec{P}$-space whose volume is

$$2\pi P_2 \sin\theta\, P_2\, d\theta\, dP_2 = 2\pi m P_2 \sin\theta\, d\theta\, d\mathcal{E}_2 \qquad \text{(A2.1)}$$

The number of end states in $d\theta\, d\mathcal{E}_2$ space is thus[15]

$$(V/h^3)\, 2\pi m P_2 \sin\theta\, d\theta\, d\mathcal{E}_2 \equiv \rho\, d\theta\, d\mathcal{E}_2 \tag{A2.2}$$

(The density $\rho$ introduced above is used below in calculating the transition probability; since spin is conserved in the transitions of interest, the density of possible end states in phase space is $1/h^3$ instead of $2/h^3$.)

The transitions will occur between states of the entire system, electron plus phonons, which conserve energy. The transition of the electron from $\vec{P}_1$ to $\vec{P}_2$ requires a compensating change in the phonon field.[16] The conservation laws allow two possibilities: (I) *phonon emission*; the longitudinal acoustical mode with

$$P_\alpha \equiv \hbar \vec{k}_\alpha = -(\vec{P}_2 - \vec{P}_1) \tag{A2.3}$$

undergoes a change

$$n_\alpha \rightarrow n_\alpha + 1 \tag{A2.4}$$

with a change in energy for the electron of

$$\mathcal{E}_2 - \mathcal{E}_1 = -\hbar\,\omega_\alpha = -h\,c/\lambda = -cP_\alpha \tag{A2.5}$$

where $c$ is the velocity of the longitudinal phonons that are chiefly responsible for the scattering. These relationships lead to conservation of the sum of $\vec{P}$ for the electron plus $\sum n_\alpha \vec{P}_\alpha$ for the phonons. The other possibility is (II) *phonon absorption*, for this case

$$\vec{P}_\beta \equiv \hbar \vec{k}_\beta = (\vec{P}_2 - \vec{P}_1) \tag{A2.6}$$

$$n_\beta \rightarrow n_\beta - 1 \tag{A2.7}$$

$$\mathcal{E}_2 - \mathcal{E}_1 = +cP_\beta \tag{A2.8}$$

again with conservation of the sum of $P$ vectors.

If we denote by $\mathcal{E}$ the energy of the electron after collision plus the change in phonon energy, then the requirement of equality for energy before and after collision gives

$$\mathcal{E} = \mathcal{E}_2 + \delta n_\gamma c P_\gamma = \mathcal{E}_1 \tag{A2.9}$$

where $\delta n_\gamma = +1$ is the phonon emission or $\alpha$ surface and $\delta n_\gamma = -1$ is the phonon absorption or $\beta$ surface of Fig. 4.

[15] The notation in this appendix follows closely that of W. Shockley, "Electrons and Holes in Semiconductors," D. van Nostrand (1950) to which we shall refer as *E and H in S*. See page 253 for a similar treatment of $\rho$.

[16] This condition is analogous to one for the conservation of momentum but has a different interpretation. See for example *E and H in S*, p. 519 equation (15).

We shall next insert these symbols into the conventional expression for transition probability. We consider a system described by one or more sets of quantum numbers, say $x_1, x_2, \cdots, x_n$ which may take on discrete but closely spaced values so that the numbers of states lying in a range $dx_1, \cdots, dx_n$ is

$$\rho(x_1, \cdots, x_n)\, dx_1 \cdots dx_n. \tag{A2.10}$$

The system may make a transition from an initial state $\varphi_0$ and energy $\mathcal{E}_0$ to another state $\varphi_i$ of the same energy between which there is a matrix element $U_{0i}$. The total probability of the system making a transition per unit time to the range of quantum numbers $dx_2, dx_3, \cdots, dx_n$ is then[17]

$$W_{0i}\, dx_2 \cdots dx_n = (2\pi/\hbar)\, |\, U_{0i}\,|^2\, [\rho/(\partial \mathcal{E}_i/\partial x_1)]\, dx_2, \cdots, dx_n \tag{A2.11}$$

where $\partial \mathcal{E}_i/\partial x_1$ is evaluated where $\mathcal{E}_i = \mathcal{E}_0$; if for the range $dx_2, \cdots, dx_n$ of the other quantum numbers there is no $x_1$ value that gives $\mathcal{E}_i = \mathcal{E}_0$, then the transition does not occur.

We shall apply this to our case letting $\theta = x_1$ and $\mathcal{E}_2 = x_2$. The expression $\partial \mathcal{E}_i/\partial x_1$ then becomes

$$\frac{\partial \mathcal{E}_i}{\partial x_1} \equiv \left(\frac{\partial \mathcal{E}}{\partial \theta}\right)_{\mathcal{E}_2} = \delta n_\gamma\, c \left(\frac{\partial P_\gamma}{\partial \theta}\right)_{\mathcal{E}_2} \tag{A2.12}$$

where

$$\begin{aligned} \frac{\partial}{\partial \theta} P_\gamma &= \frac{\partial}{\partial \theta} |\, \vec{P}_2 - P_1\,| = \frac{\partial}{\partial \theta} (P_2^2 + P_1^2 - 2P_1 P_2 \cos\theta)^{1/2} \\ &= P_1 P_2 \sin\theta / P_\gamma. \end{aligned} \tag{A2.13}$$

We then obtain, for $W_{12}\, d\mathcal{E}_2$, the probability per unit time of transition of the electron from $\vec{P}_1$ to states with energies between $\mathcal{E}_2$ and $\mathcal{E}_2 + d\mathcal{E}_2$, the expression

$$\begin{aligned} W_{12}\, d\mathcal{E}_2 &= (2\pi/\hbar)\, |\, U\,|^2\, (V/h^3) 2\pi m P_2 \sin\theta \times (P_\gamma/c\delta n_\gamma P_1 P_2 \sin\theta)\, d\mathcal{E}_2 \\ &= (V/2\pi\hbar^4) m\, |\, U\,|^2\, (P_\gamma/c\delta n_\gamma\, P_1)\, d\mathcal{E}_2 \\ &= (V/2\pi\hbar^4) m\, |\, U\,|^2\, (P_\gamma/P_1)(-dP_\gamma); \end{aligned} \tag{A2.14}$$

where the negative coefficient of $dP_\gamma$ is without significance except for its relationship to the selection of the limits of integration. In subsequent equations we shall disregard the sign convention which relates $d\mathcal{E}_2$ to $dP_\gamma$; no

[17] See L. I. Schiff "Quantum Mechanics," McGraw-Hill Book Co. (1949), equation (29.12). The additional factor $1/(\partial \mathcal{E}_i/\partial x_1)$ converts the $\rho$ used here to that of Schiff, which latter is number of states per unit energy range.

error is introduced provided the subsequent integrations are always in the direction of increasing values for the variables concerned.

### *A.3* The Allowed Ranges for $P_\gamma$

An electron with an energy corresponding to room temperature can change its energy by only a small fraction in a one phonon transition. The extremes occur for $\theta = \pi$ corresponding to complete reversal of direction. For this case we have

$$\begin{aligned} \mathcal{E}_2 - \mathcal{E}_1 &= (P_2^2 - P_1^2)/2m = -\delta n_\gamma c P_\gamma \\ &= -\delta n_\gamma c(P_2 + P_1) \end{aligned} \tag{A3.1}$$

so that

$$P_2 - P_1 = -\delta n_\gamma \, 2mc \equiv -\, 2\delta n_\gamma \, P_0. \tag{A3.2}$$

Thus the limiting values of $P_2$ differ from $P_1$ by

$$\pm 2P_0 = \pm\, 2mc \tag{A3.3}$$

in keeping with the results shown in Fig. 4. [For $v_1 = P_1/m = 10^7$ cm/sec, corresponding to $\mathcal{E}_1 = 0.025$ electron volts, and $c = 5.4 \times 10^5$ cm/sec, it is seen that $P_2$ and $P_1$ differ by 10%.] For this case the range of $P_\gamma$ is

$$\text{phonon emission, } \delta n_\alpha = +1,\ P_\alpha \text{ from 0 to } 2(P_1 - P_0) \tag{A3.4}$$

$$\text{phonon absorption, } \delta n_\beta = -1,\ P_\beta \text{ from 0 to } 2(P_1 + P_0). \tag{A3.5}$$

A singularity occurs for $P_1 = P_0$. For this case phonon emission becomes impossible and the inner curve of Fig. 4 shrinks to zero; in Fig. A1 we show the sequence of shrinkage. The value of $\mathcal{E}_1$ for this condition corresponds to thermal energy for a temperature of less than 1°K. Under the conditions for which we shall compare theory and experiment, a negligible number of electrons lie in this range. Accordingly we shall use the above limits in calculations and neglect the small errors introduced.

### *A.4* The Matrix Element and the Mean Free Path

The matrix element may be written in the form[18]

$$|\, U \,|^2 = \mathcal{E}_{1n}^2 \, (2n_\gamma + 1 + \delta n_\gamma) P_\gamma c / 4V c_{ll} \tag{A4.1}$$

where $\mathcal{E}_{1n}$ is the derivative of the edge of the conduction band in respect to dilatation of the crystal and $c_{ll}$ is the elastic constant for longitudinal

[18] See *E and H in S*, page 528, equation 31. The second expression in equation 31 of this reference is in error by omission of a factor $\hbar\omega_{k\alpha} = cP_\gamma$.

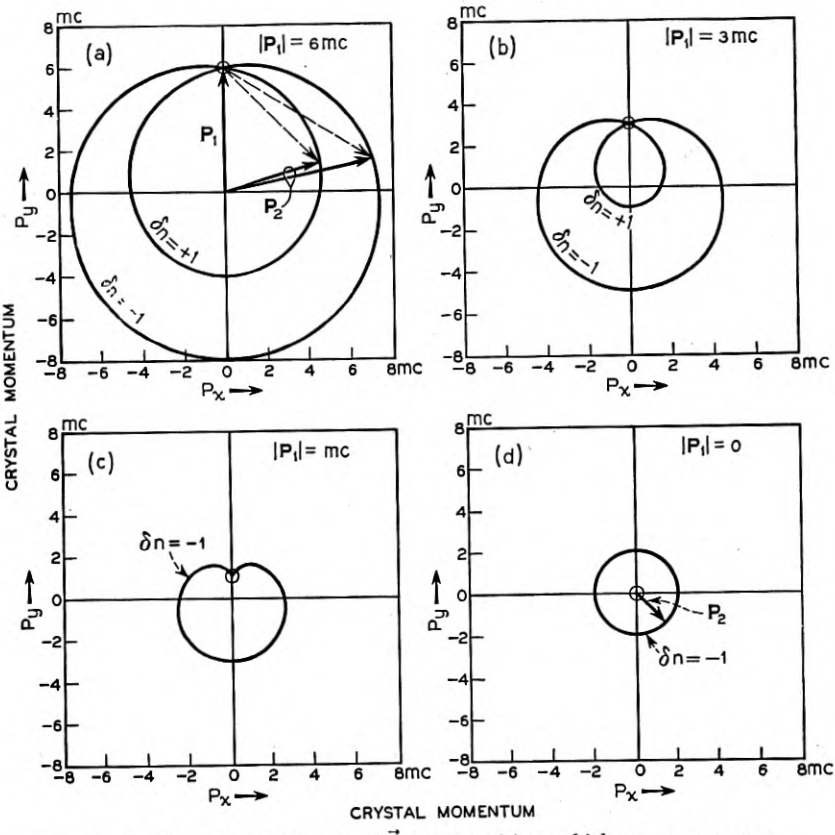



Fig. A1—Initial and Final Values of $\vec{P}$ for Transitions which conserve energy.
(a) The two nearly spherical surfaces for $|\vec{P_1}| = 6$ mc.
(b) The two surfaces for $|\vec{P_1}| = 3$ mc.
(c) For an electron with $|\vec{P_1}| =$ mc, energy loss is impossible.
(d) The case of $|\vec{P_1}| = 0$.

waves. Inserting this in the expression (A2.14) for $W_{12}\, d\mathcal{E}_2$, we obtain

$$W_{12}\, d\mathcal{E}_2 = (V/2\pi\hbar 4)\, m\, [\mathcal{E}_{1n}^2 P_\gamma c/4Vc_{\ell t}] \times (2n_\gamma + 1 + \delta n_\gamma)(P_\gamma/P_1)\, dP_\gamma \quad \text{(A4.2)}$$

$$= (1/\ell)(1/8mP_1)(2n_\gamma + 1 + \delta n_\gamma)(cP_\gamma/kT)P_\gamma\, dP_\gamma$$

where we have used the symbol $\ell$ to represent

$$\ell = \pi\hbar^4 c_{\ell t}/m^2\mathcal{E}_{1n}^2\, kT \quad \text{(A4.3)}$$

because, as we shall shortly show, $\ell$ is the mean free path for electrons.

For the cases with which we shall be concerned, the values of $n_\gamma$ may be approximated by classical equipartition. This may be seen from the fact that largest energy phonons correspond approximately to an energy of

$$\begin{aligned} cP_\gamma &\doteq 2cP_1 = (4cm/P_1)P_1^2/2m \\ &\doteq (4c/v_1)kT. \end{aligned} \tag{A4.4}$$

Their energies will, therefore, be considerably less than $kT$. For large increases in electron energy in high fields, however, this approximation may not be adequate. At room temperature $cP_\gamma/kT \doteq 0.2$ and the critical range will correspond approximately to an increase of about 10 fold of electron temperature above the temperature of the crystal. Under these conditions $cP_\gamma$ deduced from equation (A4.4) will be about $2kT$ for the most energetic phonons; for this condition, however, $\vec{P}_\gamma$ lies at the edge of the Brillouin zone and dispersive effects must be considered. In this treatment we shall not investigate further these limits and shall in general assume that $cP_\gamma < kT$.

We shall next derive an expression for the mean free path and verify that the scattering is isotropic. These results can be derived more simply and directly from the matrix element by neglecting $(P_0/P_1)$ and $(cP_\alpha/kT)$ from the outset. For the treatment of energy losses that follows, we cannot make these approximations. We shall, however, make them in the remainder of this section thus establishing that our more general formulation reduces correctly to the more convenient and simpler formulation usually used.

For the condition under which equilibrium applies we may approximate $n_\gamma$ as follows:

$$n_\gamma = 1/[(\exp cP_\gamma/kT) - 1] \doteq (kT/cP_\gamma) - \tfrac{1}{2}. \tag{A4.5}$$

Then, for the phonon emission or $\alpha$ case, the contribution to $W_{12}\, d\mathcal{E}_2$ becomes

$$W_{12}\, d\mathcal{E}_2 = (4m\ell P_1)^{-1}[1 + (cP_\alpha/2kT)]P_\alpha\, dP_\alpha \tag{A4.6a}$$

and for the phonon absorption case, it becomes

$$W_{12}\, d\mathcal{E}_2 = (4m\ell P_1)^{-1}[1 - (cP_\beta/2kT)]P_\beta\, dP_\beta. \tag{A4.6b}$$

We shall use these expressions later for the calculation of energy exchange, in which case the terms in $cP_\alpha/2kT$, which favor phonon emission, play an important role. In order to check the expression for mean free path we neglect these terms, however, and also approximate the integral of $P_\alpha\, dP_\alpha$ by $2P_1^2$ rather than $2(P_1 - P_0)^2$. The total probability of transition from state

$\vec{P}_1$, which should be taken to be $1/\tau_1$ where $\tau_1$ is the mean free time, is then

$$1/\tau_1 = W_1 = (4m\ell P_1)^{-1}\, 4P_1^2 = (P_1/m)/\ell = v_1/\ell. \tag{A4.7}$$

This is just the relationship appropriate to the interpretation of $\ell$ as a mean free path between collisions.[19] It does not follow that $\tau_1$ is the relaxation time for the current, however, unless the average velocity after collision is zero. We shall next show that the average velocity after collision is zero to the same degree of approximation used above by showing that scattering into any solid angle of directions is simply proportional to the solid angle, i.e. the direction of motion after collision is random.

The conclusion that the scattering is nearly isotropic follows from the approximate proportionality of probability to $P_\gamma dP_\gamma$. Since $P_2$ is nearly equal to $P_1$ and substantially independent of $\theta$, we may write

$$\begin{aligned} P_\alpha\, dP_\alpha &= \tfrac{1}{2}\, d(P_\alpha)^2 = \tfrac{1}{2}\, d\,(2P_1^2 - 2P_1^2 \cos\theta) \\ &= -\,P_1^2\, d\cos\theta = P_1^2 \sin\theta\, d\theta. \end{aligned} \tag{A4.8}$$

The last term is simply proportional to the solid angle lying in range $d\theta$; hence the end states are distributed with uniform probability over all directions and the scattering is isotropic.

### *A5.* Approximate Equivalence to Elastic Sphere Model

In this section we shall show that on the average the energy exchange between the electron and the phonons when the electron is scattered through an angle $\theta$ is very similar in form to that corresponding to elastic collisions between spheres with the phonons represented by a mass much greater than the electrons. If $dP_\alpha$ and $dP_\beta$ correspond to scattering through angles between $\theta$ and $\theta + d\theta$, then the energy loss for phonon emission is $cP_\alpha$ and the energy gain for absorption is $cP_\beta$. The relative probabilities of loss and gain are given by equations (A4.6) and from these it is found that the average energy gain is

$$\begin{aligned} \langle \delta\mathcal{E} \rangle &= \frac{cP_\beta[1 - (cP_\beta/2kT)]P_\beta\, dP_\beta - cP_\alpha[1 + (cP_\alpha/2kT)]P_\alpha\, dP_\alpha}{[1 - (cP_\beta/2kT)]P_\beta\, dP_\beta + [1 + (cP_\alpha/2kT)]P_\alpha\, dP_\alpha} \\ &= \frac{c[P_\beta^2\, dP_\beta - P_\alpha^2\, dP_\alpha] - (c^2/2kT)(P_\beta^3\, dP_\beta + P_\alpha^3\, dP_\alpha)}{[1 - (cP_\beta/2kT)]P_\beta\, dP_\beta + [1 + (cP_\alpha/2kT)]P_\alpha\, dP_\alpha}. \end{aligned} \tag{A5.1}$$

Since we are concerned chiefly with cases in which $P_0 \ll P_1$, $P_2 \doteq P_1$, and $cP_\beta/kT \ll 1$, we may set

$$P_\alpha \doteq P_\beta \doteq P_\gamma \equiv 2P_1 \sin(\theta/2) \tag{A5.2}$$

[19] See *E and H in S*, Chapter 11.

where $P_\gamma$ corresponds to neglecting the energy change of the electron on collision. This approximation is not good enough for the first term in (A5.1) which involves $P_\beta - P_\alpha$. In order to evaluate this first term we note that the relationships

$$P_\alpha = P_\gamma[1 - (P_0 P_\gamma / 2P_1^2)] \tag{A5.3a}$$

$$P_\beta = P_\gamma[1 + (P_0 P_\gamma / 2P_1^2)] \tag{A5.3b}$$

may be derived up to the first order in $P_0$. This permits us to write

$$\begin{aligned} P_\beta^2\, dP_\beta - P_\alpha^2\, dP_\alpha &\doteq (\tfrac{1}{3})\, d\,[(P_\beta - P_\alpha)(3P_\gamma^2)] \\ &= d[P_0 P_\gamma^4 / P_1^2] = 4(P_0 P_\gamma^2 / P_1^2)\, P_\gamma\, dP_\gamma . \end{aligned} \tag{A5.4}$$

Hence

$$\langle \delta\mathcal{E} \rangle = 2(cP_0 P_\gamma^2 / P_1^2) - c^2 P_\gamma^2 / 2kT. \tag{A5.5}$$

The second term is proportional to the (average change in momentum)$^2$ for collision by angle $\theta$. It thus corresponds to energy which would be transferred to an initially stationary mass $M \gg m$ by the colliding electron provided we take

$$M = kT/c^2. \tag{A5.6}$$

That this mass is much greater than the electron's mass may be seen in terms of $v$, the velocity of a thermal electron:

$$mv_1^2/2 = kT. \tag{A5.7}$$

From this we obtain

$$M = m(v_1/c)^2/2 \doteq 170m \tag{A5.8}$$

at room temperature where $v_1 \doteq 10^7$ cm/sec while $c \doteq 5.4 \times 10^5$ cm/sec.

The first term may then be interpreted as follows:

$$2cP_0 P_\gamma^2 / P_1^2 = 2c^2 m P_\gamma^2 / P_1^2 = 2(kTm/M)(P_\gamma^2 / P_1^2) \tag{A5.9}$$

If this is averaged over all angles $\theta$, the $P_\gamma^2/P_1^2$ term becomes 2; the energy gain is then just that picked up by a mass $m$ colliding with a mass $M$ moving with a Maxwellian distribution at temperature $T$ as may be seen as follows: For this case the velocity $v_M$ of $M$ parallel to the line of centers on collision imparts an added velocity $2v_M$ to the electron and, on the average, an energy

$$(\tfrac{1}{2})m\,\langle (2v_M)^2 \rangle = 2m\,\langle v_M^2 \rangle = 2\,mkT/M \tag{A5.11}$$

since $M\langle v_M^2\rangle/2 = kT/2$. In addition, however, there is an effect due to relative velocity: when $v_M$ is directed so as to increase the closing velocity, the probability of collision is increased. Due to this effect, collisions with higher relative velocity are favored and as a result the energy transferred due to the $v_M$ effect is just doubled[20] leading to a total contribution of

$$(\text{energy transfer due to } v_M) = 4\,mkT/M. \tag{A5.10}$$

This is just to the first term of (A5.5) when averaged over all values of $\theta$.

Thus the average of the gain in energy term in (A5.5) is just that corresponding to interactions with heavy masses $M$ in thermal agitation. The difference is that in the sphere model the average energy gain term is independent of $\theta$, whereas in the phonon case it varies as $P_\gamma^2/P_1^2$ and approaches zero for forward scattering so that the dependence upon angle is different. The energy loss term, however, is correctly represented by the sphere model.

The average value of $\langle\delta\mathcal{E}\rangle$ averaged over all values of $\theta$ is denoted by $\langle\delta\mathcal{E}\rangle_{P_1}$. Since $P_\gamma^2$ averaged over $\theta$ is $2P_1^2$, we obtain

$$\begin{aligned}\langle\delta\mathcal{E}\rangle_{P_1} &= 4cP_0 - c^2P_1^2/kT \\ &= 4mc^2\,(1 - P_1^2/4\,mkT).\end{aligned} \tag{A5.12}$$

From this expression it is seen that an electron with energy $P_1^2/2m = 2kT$ keeps the same energy on the average after $M$ collision. We shall use expression (A5.12) in Section A.6.

For high electric fields, the electron energies are higher than thermal and the loss terms predominate. Furthermore, the scattering in both cases is nearly isotropic if $M \gg m$ and the colliding particles are spheres. Hence, the analysis of kinetic theory of ionized gases can be applied to a high degree of approximation to estimate electron behaviors.

It should be stressed that several approximations are involved in this treatment. In particular it is assumed that (I) $cP_\gamma < kT$ and that (II) $P_1 \gg P_0$. If this is true, then

$$v_1/c = P_1/P_0 \gg 1 \tag{A5.13}$$

so that the approximation used in considering the heavy spheres to be moving slowly holds and the mass of the heavy spheres is much greater than the electron mass:

$$M/m \doteq (kT/c^2)/(2kT/v_1^2) = v_1^2/2c^2. \tag{A5.14}$$

If conditions (I) and (II) are not satisfied, the approximations leading to (A5.5) will require revision.

[20] If $v_{1n}$ is the velocity towards centers, then the probability of collision is weighted by $[1 + (v_M/v_{1n})]$ and the term linear in $v_M$ in the energy, which is $(\frac{1}{2}m)(4v_{1n}v_M)$, contributes $2\langle v_M^2\rangle m$ to the average transfer.

### *A.6* Approximate Treatments of Mobility in High Fields

A correct treatment of mobility in high electric fields $\vec{E}$ is based upon finding the steady state distribution function $f(\vec{P}, \vec{E}, T)$ which satisfies the Boltzmann equation, i.e. a function for which the rate of change due to acceleration by $\vec{E}$ just balances that due to scattering. This method leads at once to rather formidable mathematics which may tend to obscure somewhat the physical forces at work. We shall in this section derive relationships between drift velocity and $\vec{E}$ on the basis of simpler models and shall compare the results with the exact treatment as given in the literature for the case of a gas.

For this purpose, we shall first suppose that the field in effect raises the electrons to a temperature $T_e$ which is greater than the temperature $T$ of the phonon distribution. The electrons with temperature $T_e$ will have collisions at a rate $(T_e/T)^{1/2}$ greater than before and their mobility will be reduced from its equilibrium value $\mu_0$ to a new value $\mu$

$$\mu = (T/T_e)^{1/2}\,\mu_0. \tag{A6.1}$$

The average rate at which the electric field does work on an electron is then

$$(d\mathcal{E}/dt)_{\text{field}} = \text{Force} \times \text{Speed} = q\mu E^2. \tag{A6.2}$$

For steady state conditions this must be equal to minus the average rate at which an electron gains energy from the phonons. Denoting this by $(d\mathcal{E}/dt)_{\text{phonons}}$ we have

$$(d\mathcal{E}/dt)_{\text{field}} + (d\mathcal{E}/dt)_{\text{phonons}} = 0. \tag{A6.3}$$

In order to calculate the average rate of energy loss to the phonons, we consider the average energy gain of an electron of momentum $P_1$. As given by (A5.12):

$$\langle \delta\mathcal{E} \rangle_{P_1} = 4mc^2\,[1 - (mv_1^2/4kT)]. \tag{A6.4}$$

According to our assumption, the number of electrons in the velocity range $v$ to $v + dv$ is $N(v)\,dv = A \exp(-mv^2/2kT_e)v^2 dv$. These electrons suffer collision at a rate $v/\ell$. Hence the average rate of energy gain is

$$\begin{aligned}(d\mathcal{E}/dt)_{\text{phonons}} &= \int \langle \delta\mathcal{E} \rangle_P (v/\ell) N(v)\,dv \Big/ \int N(v)\,dv \\ &= (8/\sqrt{\pi})(mc^2 v_e/\ell)[1 - (v_e/v_T)^2]\end{aligned} \tag{A6.5}$$

where we have introduced

$$v_T = (2kT/m)^{1/2}, \qquad v_e = (2kT_e/m)^{1/2}. \tag{A6.6}$$

We see that for thermal equilibrium, with $T = T_e$ and $v = v_e$, this equation gives correctly the result that there is no net interchange of energy between electrons and phonons.

The equation for mobility for the case of phonon scattering is

$$\mu_0 = 4q\ell/3\sqrt{\pi}v_T m. \tag{A6.7}$$

This expression may be used to reduce the steady state equation:

$$\begin{aligned} 0 &= (d\mathcal{E}/dt)_{\text{field}} - (d\mathcal{E}/dt)_{\text{phonons}} \\ &= (v_T/v_e)q\mu_0 E^2 + (8/\sqrt{\pi})(mc^2 v_T/\ell)(v_e/v_T)\,[1 - (v_e/v_T)^2] \end{aligned} \tag{A6.8}$$

to

$$(v_e/v_T)^4 - (v_e/v_T)^2 = (3\pi/32)(\mu_0 E/c)^2. \tag{A6.9}$$

This equation may be solved for $v_e/v_T$:

$$(v_e/v_T = \sqrt{(\tfrac{1}{2})}(1 + [1 + (3\pi/8)(\mu_0 E/c)^2]^{1/2})^{1/2} \tag{A6.10}$$

The drift velocity then becomes

$$v_d = \mu E = \mu_0 E\,\sqrt{2}/(1 + [1 + (3\pi/8)(\mu_0 E/c)^2]^{1/2})^{1/2}. \tag{A6.11}$$

For $E \ll c/\mu_0$, we have

$$v_d = \mu_0 E. \tag{A6.12}$$

For $E \gg c/\mu_0$, we have[21]

$$v_d = (32/3\pi)^{1/4}\,(\mu_0 Ec)^{1/2} = 1.36(\mu_0 Ec)^{1/2}. \tag{A6.13}$$

These equations define a critical field $E_c$ at which the two limiting cases would give the same mobility:

$$E_c = (32/3\pi)^{1/2} c/\mu_0 = 1.84c/\mu_0. \tag{A6.14}$$

At this critical field, the drift velocity is less than the value on either limiting form by a factor of

$$v_d/\mu_0 E_c = \sqrt{2}\,/\,[1 + 5^{1/2}]^{1/2} = 0.785. \tag{A6.15}$$

This reduction in mobility corresponds to an electron temperature of

$$T_e = T/(0.785)^2 = 1.62T. \tag{A6.16}$$

Expressed in terms of $c$ the drift velocity given by the limiting forms is

$$v_c = v_d\ (\text{extrapolated}) = 1.84c \tag{A6.17}$$

[21] This relationship has been published in *Bulletin of Am. Phys. Soc.*, Vol. 26, No. 1, paper 55.

and the actual drift velocity is

$$v_d(E_c) = 0.785 \times 1.84c = 1.45c. \tag{A6.18}$$

If we introduce $E_c$ into the steady state equation (A6.9), and express $v_e/v_T$ as a mobility ratio, we obtain

$$(\mu_0/\mu)^4 - (\mu_0/\mu)^2 = (E/E_c)^2. \tag{A6.19}$$

We give this equation in order to show the similarity of the Maxwellian distribution case to the cruder case considered next.

A similar treatment may be given for a hypothetical distribution of electrons such that all have the same energy $\mathcal{E}$ and speed $v$. The effective mobility of such a distribution is[22]

$$\begin{aligned} \mu &= (q\tau/m)\,(1 + (\tfrac{1}{3})\,d\ell n\tau/d\ell nv) \\ &= 2q\ell/3vm = \pi^{1/2}\,\mu_0 v_T/2v. \end{aligned} \tag{A6.20}$$

The steady state equation deduced from (A6.19) and (A6.4) is

$$(v^2/2v_T^2)\,[(v^2/2v_T^2) - 1] = (3\pi/64)\,(\mu_0 E/c)^2 \tag{A6.21}$$

For high fields we find that this steady state condition gives

$$v_d = (\pi/3)^{1/4}\,(c\mu_0 E)^{1/2} = 1.01\,(c\mu_0 E)^{1/2}, \tag{A6.22}$$

a value somewhat smaller than that obtained from the Maxwellian approximation. This leads to a critical field of

$$E_c = (\pi/3)^{1/2}\,c/\mu_0 = 1.03\,c/\mu_0 \tag{A.623}$$

at which $v_d$ given by (A6.22) extrapolates to give the same value as $\mu_0 E$. We may use this distribution and make it give identical results with the Maxwellian distribution. If we use the steady state condition for low fields, we find $v = 2^{1/2}\,v_T$ and

$$\mu = \mu_0' = \sqrt{\pi/8}\,\mu_0 = 0.625\,\mu_0. \tag{A6.24}$$

In terms of this $\mu_0'$, the steady state equation becomes

$$(\mu_0'/\mu)^2\,[(\mu_0'/\mu)^2 - 1] = (E/E_c')^2. \tag{A6.25}$$

with

$$E_c' = (8/3)^{1/2}\,c/\mu_0' = 1.63\,c/\mu_0'. \tag{A6.26}$$

It is evident that if we chose modified values of $c'$ and $\ell'$ so as to make $\mu_0'(\ell')$ become equal to $\mu_0$ and $E_c'(c', \ell') = E_c(c, \ell)$, then the monoenergetic

[22] See *E and H in S* problem 8 page 293.

approximation will give the same results as the Maxwellian distribution. We use this procedure in the following section.

As pointed out in Section A.5, the problem treated here is closely analogous to electronic conduction in gasses. For this case, an exact treatment has been given for high fields such that the motion of the large masses $M$ may be neglected. The drift velocity is found to be[23]

$$v_d = (\tfrac{4}{3})^{3/4}(\pi/4)^{1/2}(m/M)^{1/4}(qE\ell/m)^{1/2}/\Gamma(\tfrac{3}{4}). \tag{A6.27}$$

Substituting $kT/c^2$ for $M$ and using (A6.7) to eliminate $\ell$, we obtain

$$\begin{aligned} v_d &= (\pi^3/6)^{1/4}\Gamma(\tfrac{3}{4})(\mu_0 Ec)^{1/2} \\ &= 1.23\ (\mu_0 Ec)^{1/2}. \end{aligned} \tag{A6.28}$$

This value lies intermediate between the two simple approximations considered above and leads to a critical field of

$$E_c = 1.51\ c/\mu_0 \tag{A6.29}$$

at a velocity of

$$v_{dc} = \mu_0 E_c = 1.51\ c. \tag{A6.30}$$

Since the exact case gives $v_d = \mu_0 E$ for low fields and $v_d = \mu_0(EE_c)^{1/2}$ for large fields, it is evident that either the Maxwellian or single energy distribution will approximate it well (provided suitable choices of $\mu_0$ or $\mu_0'$ and $E_c$ or $E_c'$ are made) except for a small error near $E_c$.

The distribution in energy for the gas case leads to a probability of finding the electron in a range $v$ to $v + dv$ proportional to

$$\left[\exp - \int_0^v mv\ dv/(kT + 3M(q\ell E/3mv)^2)\right] v^2\ dv \tag{A6.31}$$

For high fields this reduces to

$$[\exp - (3m/4M)(v^2 m/q\ell E)^2]v^2\ dv$$

It is seen that this distribution weights the low energies less heavily than does the Maxwellian for the same average energy and thus gives a lower mobility. It weights them more heavily than does the single energy and, therefore, gives a higher mobility than it.

[23] *Druyvesteyn Physica 10*, 61 (1930). See also S. Chapman and T. G. Cowling, "The Mathematical Theory of Non-Uniform Gases," Cambridge at the University Press, 1939, page 351.

## *A.7* The Effect of the Optical Modes

### *A.7a. Introduction*

The treatment presented above is based entirely upon interaction with the longitudinal acoustical modes of the crystal. Since the diamond structure has two atoms per unit cell, it is also possible to have "optical modes" in which the two atoms vibrate in opposite directions. Such modes may have long wave lengths; for example, do the same thing in each unit cell, and thus correspond to small values of $\vec{P}_\gamma$. Hence they may interact with the electron waves with $\vec{P}_\gamma \doteq \vec{P}_1$. Their frequencies correspond roughly to a wave length of about $(\frac{1}{2})$ the lattice constant in the [100] direction and hence to a frequency of about

$$(4.92 \times 10^5)((\tfrac{1}{2})5.6 \times 10^{-8})^{-1}\,(2/\pi) = 1.12 \times 10^{13}\ \text{sec}^{-1} \qquad \text{(A7.1)}$$

The last factor of $(2/\pi)$ is a crude allowance for dispersion, which leads to a decreasing phase velocity at short wave lengths. This corresponds to an energy

$$h\nu = k\ 520°K. \qquad \text{(A7.2)}$$

These optical phonons thus contain much more energy than the acoustical phonons; the latter having at room temperature an energy of about

$$cP_1 = 2(c/v_1)\ (v_1P_1/2) = (1/10)k\ 300° = k\ 30°K \qquad \text{(A7.3)}$$

[or about $k$ 100°$K$ if we use the higher effective value of $c$ discussed in Section 4]. Furthermore, the optical phonons will be only slightly excited; thus collisions with them will in general involve phonon emission so that a collision will on the average result in an energy loss of nearly $500k$. This is very large compared, for example, to the average loss of about 8 $mc^2 \doteq k$ 12° per collision for electrons with energies of $4kT$.

A difficulty with the optical modes is that for the approximation of spherical energy bands, which we have used in earlier parts of this paper, they should have matrix elements which vanish when $P_1 = 0$. This conclusion is reached by considering the deformation potentials corresponding to an optical displacement: this may be thought of as moving one of the face-centered cubic sublattices of the diamond structure in respect to the other. Since the initial position is one of tetrahedral symmetry, there can be no first order change in the energy $\mathcal{E}_c$ at the bottom of the conduction band. There may be a distortion of the energy spheres for higher energies, however, and this can lead to matrix elements proportional to $P_1^2$ and transition probabilities proportional to $P_1^4$.[25]

[25] This view is in disagreement with the position stated by F. Seitz in his two papers on mobility. *Phys. Rev. 73,* 550 (1948) and *76,* 1376 (1949). See for example the text between equations (14) and (15) of the latter paper.

On the other hand, if the band is degenerate and has energy surfaces consisting, for example, of three sheets, then an optical displacement may split the degeneracy and the shift in energy for $P_1 = 0$ can be linear in the displacement. (For example consider one wave function with angular dependence of the form $(\cos\theta + \sin\theta\cos\varphi + \sin\theta\sin\varphi)$ i.e. a $p$-type wave function with its axis along [111]. Its energy should certainly change for relative displacements of the two sublattices along the [111] direction since positive and negative displacements are unsymmetrical in their distortion in respect to this line.)

Chiefly from evidence on magneto-resistance, the writer is convinced that the electron energy band is complex in form. Thus it would be expected that the optical modes would have matrix elements for low values of $P_1$. We shall assume that this is the case but shall not endeavor to deal with a non-spherical band. This an inherently inconsistent approach, but should give at least a semiquantitative agreement with an exact theory.

In order to illustrate the effect of the optical modes, we shall assume for the moment that for high fields they are the dominant mechanism of scattering and that the scattering is isotropic. If the mean free time between collisions is $\tau$, then the power input is

$$(d\mathcal{E}/dt)_{\text{field}} = q^2\tau E^2/m. \tag{A7.4}$$

If the temperature is so low that the optical modes are only slightly excited, the transitions will in general absorb an optical phonon so that

$$(d\mathcal{E}/dt)_{\text{op}} = h\nu/\tau. \tag{A7.5}$$

The steady state condition leads to the surprising but simple result that

$$v_d = q\tau E/m = (h\nu/m)^{1/2} \tag{A7.6}$$

so that the drift velocity is independent of $E$.

If we insert $k\,520°K$ for $h\nu$ and the free electron mass for $m$, $v_d$ becomes the velocity of an electron with $k\,260°K$ of energy giving

$$v_d = 0.88 \times 10^7 \text{ cm/sec.} \tag{A7.7}$$

The limiting value on Fig. 2 for $298°K$ corresponds to extrapolating Ohm's law to about 1500 volts/cm or a drift velocity of $1500 \times 3600 = 0.54 \times 10^7$. The limiting value for $77°K$ is about twice as high. These values are seen to be in reasonable agreement with the predicted value.

It should be pointed out, however, that the answer obtained for $v_d$ depends implicitly on the assumption that a relaxation time may be used in the simple way employed above. To illustrate that the result is not completely general we refer the reader to equation (4.4) which was obtained

on the basis of a somewhat different treatment. According to (4.4)

$$\begin{aligned} v_d &= (\tfrac{1}{2})\,(2h\nu/m)^{1/2} = (h\nu/2m)^{1/2} \\ &= 0.63 \times 10^7 \text{ cm/sec.} \end{aligned} \tag{A7.8}$$

a result smaller than (A7.6) by a factor of $2^{1/2}$.

A correct treatment of the optical modes together with acoustical modes would involve solving the Boltzmann equation to find the steady state distribution. This will obviously present problems of considerable complexity. In particular scattering will suddenly begin to increase when the electron acquires an energy, $\mathcal{E}_1 > h\nu$ and it seems unlikely that analytic solutions can be obtained. Even the Maxwellian distribution leads to somewhat complicated integrals.

*A.7b. Estimate of the Matrix Element*

In order to proceed further we must estimate the order of magnitude of the optical scattering matrix element. For this purpose we introduce a "deformation potential" coefficient for the optical modes by the equation

$$\mathcal{E}_{2n} = \partial\mathcal{E}_c/\partial(x/x_0) \tag{A7.9}$$

where $x$ is the displacement parallel to the $x$-axis of one sublattice in respect to the other and $x_0$ is the $x$-component of relative displacement of the sublattices for equilibrium conditions. The same reasoning as used in treating mobility by deformation potentials[26] may then be applied and the matrix element evaluated by analogy with the dilatation waves. For the latter the matrix element may be written in the form

$$|\, U_\Delta \,|^2 = \mathcal{E}_{1n}^2 \langle \Delta^2 \rangle / 2 \tag{A7.10}$$

where $\Delta$ is the dilatation and $\langle \Delta^2 \rangle$ is the average (dilatation)$^2$ for the mode before and after transition.[27] Since half the energy in a running wave is potential

$$\begin{aligned} \tfrac{1}{2}\, c_{ll} \langle \Delta^2 \rangle\, V &= \tfrac{1}{2}\, h\nu \times [\text{average of } (n + \tfrac{1}{2})] \\ &= h\nu\,(2n + 1 + \delta n)/4. \end{aligned} \tag{A7.11}$$

This leads to the form introduced in equation (A4.1). By analogy, for the optical modes we should take

$$|\, U_{\text{op}} \,|^2 = \mathcal{E}_{2n}^2 \langle (x/x_0)^2 \rangle / 2 \tag{A7.12}$$

[26] W. Shockley and J. Bardeen *Phys. Rev. 77*, 407–408 (1950) and J. Bardeen and W. Shockley, *Phys. Rev. 80*, 72 (1950).

[27] See *E and H in S*, page 528.

where the stored energy for deformation $(x/x_0)$ is

$$\tfrac{1}{2}\, c_{00}\, (x/x_0)^2\, V \tag{A7.13}$$

and the average value for a transition from $n = 0$ to $n = 1$ is $h\nu$ for the total energy and $(\tfrac{1}{2})h\nu$ for potential. This leads to

$$|\, U_{\text{op}}\,|^2 = \mathcal{E}_{2n}^2\, h\nu/2c_{00}\, V \tag{A7.14}$$

As a first approximation we may take the stiffness between the planes of atoms separated by $x_0$ as the same as the macroscopic value. This leads to

$$c_{00} = c_{tt} \tag{A7.15}$$

Furthermore, equal relative displacements of neighbors are produced by equal values of $\Delta$ and $x/x_0$. Hence approximately equal changes in energy may occur so that we may assume that

$$\mathcal{E}_{2n} \doteq \mathcal{E}_{1n}. \tag{A7.16}$$

Under these conditions

$$|\, U_{\text{op}}\,|^2 \doteq (h\nu/kT)\;|\, U_{\Delta}\,|^2. \tag{A7.17}$$

Since the energy of the optical modes is a maximum for $P_\gamma = 0$, it will change only a small fraction for values of $P_\gamma$ comparable to $P_1$.[23] (See Fig. 3.) Consequently, conservation of energy leads to transitions between $\vec{P}_1$ and a sphere with $P_2 = [2m(\mathcal{E}_1 - h\nu)]^{1/2}$. The probability is equal to each point on the sphere and the transition probability is readily found[29] to be

$$1/\tau_{\text{op}} = (V\;|\, U_{\text{op}}\,|^2 m^2/\pi\hbar^4) v_2 \tag{A7.18}$$

where $v_2 = P_2/m$ is the speed *after collision*. If we assume relationship (A7.17) between matrix elements and introduce $\ell$ as defined in (A4.3), then

$$1/\tau_{\text{op}} = (h\nu/kT)\; v_2/\ell \equiv v_2/\ell_{\text{op}} \tag{A7.19}$$

where $\ell_{\text{op}}$ is a sort of mean free path for optical scattering.

The dependence of $v_2$ upon the velocity before collision $v_1$ is obtained as follows:

$$\mathcal{E}_2 = \mathcal{E}_1 - h\nu \tag{A7.20}$$

$$\begin{aligned} v_2 &= [2(\mathcal{E}_1 - h\nu)/m]^{1/2} \\ &= (v_1^2 - v_\nu^2)^{1/2} \end{aligned} \tag{A7.21}$$

[28] See F. Seitz, "Modern Theory of Solids," McGraw-Hill Book Co., 1940, p. 122.
[29] See *E and H in S* of A.2, p. 493, for a similar treatment.

where $v_\nu$ is the velocity corresponding to $h\nu$:

$$v_\nu = (2h\nu/m)^{1/2}. \tag{A7.22}$$

The rate of energy loss to the optical modes is simply

$$h\nu v_2/\ell_{op}. \tag{A7.23}$$

*A.7c. Approximate Steady State Treatment*

In order to test whether or not the role of optical modes can explain Ryder's data, we shall use a very crude method. We shall assume that the electrons all have the same energy and shall calculate their mobility on the basis of the mean free time at that energy; from this we calculate the power input. We shall also calculate the power loss in the same way. It is obvious that this treatment is a very poor approximation to the actual situation. An electron which loses energy to the optical modes will, under most circumstances, have only a small fraction of its energy left afterwards; thus to assume a monoenergetic distribution is unrealistic. However, the treatment does bring into the analytic expressions the principal mechanisms and, as we shall show, appears to account for the main experimental features.

The collision frequency or relaxation time for transitions involving the optical modes is given in (A7.19) and the energy loss in (A7.23). We must introduce corresponding expressions for the effect of the acoustical modes. Since the single energy distribution is to be used over the entire range of electric fields, we must introduce some approximations like those discussed in connection with (A6.25) in order to make it converge on the correct behavior at $E = 0$. The particular choice selected is a compromise between the energy loss formulae for the Maxwellian and single energy distributions:

$$(d\mathcal{E}/dt)_{\text{acous.phonons}} = (4\, mc^2 v_1/\ell)[1 - (v_1/v_T)^2]. \tag{A7.24}$$

A simplified expression is also used for the mobility:

$$\mu = q\ell/mv_1 = \mu_0 v_T/v. \tag{A7.25}$$

The relationship between $\mu_0$ and $\ell$ given by this differs by 25 per cent from the correct relationship (A6.7); since $\mu_0$ is an adjustable parameter in the comparison between theory and experiment, (A7.25) does not introduce any error at low fields. Equations (A7.24) and (A7.25) cause $\mu$ to converge on $\mu_0$ and $\mathcal{E}_1$ on $kT$ as $E$ approaches zero. (It is probable that a slightly better fit to the data would be obtained by using the procedure described with equation (A6.25); the calculations based on (A7.24) and (A7.25) were made before the (A6.25) procedure was worked out, however, and it was not considered worth while to rework them for this article.)

Rewriting the equation for mobility in terms of the collision frequencies $1/\tau = v_1/\ell$ and $1/\tau_{op}$, the power input from the electric field is

$$\begin{aligned}(d\mathcal{E}/dt)_{\text{field}} &= q\mu E^2 = q^2E^2/m\,[(1/\tau) + (1/\tau_{op})] \\ &= q\mu_0E^2/(v_1/v_T)\,[1 + (\ell/\ell_{op})(v_2/v_1)]\end{aligned} \tag{A7.26}$$

where the $v_2$ term is omitted if $v_1 < v_v$. The power delivered by phonons is

$$\begin{aligned}(d\mathcal{E}/dt)_{\text{phonons}} &= 4mc^2\,(v_1/\ell)\,[1 - (v_1/v_T)^2] - h\nu v_2/\ell_{op} \\ &= (4qc^2/\mu_0)(v_1/v_T) \\ &\quad \times [1 - (v_1/v_T)^2 - (h\nu/4mc^2)(\ell/\ell_{op})(v_2/v_1)].\end{aligned} \tag{A7.27}$$

The two coefficients of $(v_2/v_1)$ are both larger than unity according to the analysis given above. We shall introduce the symbols $A$ and $B$ for them: Accordingly

$$A = \ell/\ell_{op} = h\nu/kT, \tag{A7.28}$$

the last equality following from (A7.19), and

$$B = h\nu/4\,mc^2. \tag{A7.29}$$

If we take $h\nu = k\ 520°K$, $m =$ the electron mass and $c = 5 \times 10^5$ cm/sec, we find

$$B = 87. \tag{A7.30}$$

As discussed in the text, the losses appear to be larger than can be accounted for by these values of $m$ and $c$. The critical drift velocity used in the fit of Fig. 4 was $2.6 \times 10^6$ cm/sec and this corresponds to a value of $c$ of

$$c = v_c/1.51 = 1.72 \times 10^6 \text{ cm/sec} \tag{A7.31}$$

according to the exact treatment based on the sphere model. (As stated in the text this means an effectiveness of energy interchange about $(1.72 \times 10^6/5 \times 10^5)^2 \doteq 10$ times larger than the simple theory.)

Our simplified energy loss equation (A7.27) leads to

$$v_c = 2c \tag{A7.32}$$

so that we shall take

$$c = 1.3 \times 10^6 \tag{A7.33}$$

in this section so as to agree with the critical velocities observed in Fig. 2. This leads to a value for $B$ of

$$\begin{aligned} B &= k\ 520°K/4m\ (1.3 \times 10^6)^2 \\ &= 12.8 \end{aligned} \tag{A7.34}$$

For $A$ we shall take

$$A = 520/T \tag{A7.35}$$

The only other adjustable parameter is $\mu_0$. For this we shall use the value based on Haynes' drift mobility and acoustical scattering.

$$\begin{aligned} \mu_0(T) &= \mu_0\ (298°K)(298°K/T)^{3/2} \\ &= 3600\ (298°K/T)^{3/2} \end{aligned} \tag{A7.36}$$

This value automatically fits the room temperature data in the Ohm's law range. The $T^{-3/2}$ dependence then extrapolates it to the other ranges.

The steady state condition may then be written in the form

$$x^2(1 + Ay)(ABy + Ax^2 - 1) = z^2 \tag{A7.37}$$

where

$$v_\nu = (2h\nu/m)^{1/2} \tag{A7.38}$$

$$x = v_1/v_\nu,\ y = v_2/v_1 = (1 - x^{-2})^{1/2} \tag{A7.39}$$

$$A = h\nu/kT = (v_\nu/v_T)^2 \tag{A7.40}$$

$$B = 12.8 \tag{A7.41}$$

$$z = \mu_0\ (T)E/2cA^{1/2}. \tag{A7.42}$$

This form lends itself to calculation of $z$ as a function of $x$. The drift velocity is then found to be given by

$$\begin{aligned} u &= v_d/2c = z/x(1 + Ay) \\ &= [(ABy + Ax^2 - 1)/(1 + Ay)]^{1/2} \end{aligned} \tag{A7.43}$$

If $A \gg 1$, there are three distinct ranges of behavior for $u$ versus $z$:

*Range (I)* $u \doteq z/A^{1/2}$

For $z \to 0$, $x^2 \to 1/A$, $y = 0$ and consequently,

$$u = zA^{1/2} = v_d/2c = \mu_0\ E/2c \tag{A7.44}$$

so that the low field relationship

$$v_d = \mu_0 E \tag{A7.45}$$

is correctly given.

*Range II, $Ax^2 \gg 1$ and $x < 1$*

In this range the electrons are at high temperature but not high enough to excite the optical modes. For it

$$z^2 = Ax^4, \qquad x = z^{1/2}/A^{1/4} \tag{A7.46}$$

$$u = z^{1/2}A^{1/4} \qquad \text{or} \qquad v_d = (2c\mu_0 E)^{1/2}. \tag{A7.47}$$

This corresponds to the square root range with a critical field of $E_c = 2c/\mu_0$ and $v_c = 2c$.

*Range III, $x > 1$*

When $x$ is greater than unity, the optical modes enter the picture. For the three cases considered the values of $A$ and $AB$ are:

$$\begin{aligned} 77°K,\ A &= 6.75, \qquad AB = 87, \\ 193°K,\ A &= 2.69, \qquad AB = 34.5, \\ 296°K,\ A &= 1.74, \qquad AB = 22.3. \end{aligned} \tag{A7.48}$$

The large value of $AB$ means that as soon as $y$ is appreciably greater than zero, say 0.5 corresponding to $x = 1.15$, energy losses to optical modes dominate. As $y$ approaches unity, the value of $u$ is approximately

$$\begin{aligned} u &\doteq [AB/(1 + A)]^{1/2} \\ &= (h\nu/4mc^2)^{1/2}/(1 + A^{-1})^{1/2} \end{aligned} \tag{A7.49}$$

leading to

$$\begin{aligned} v_d(A, B) &= (h\nu/m)^{1/2}/(1 + A^{-1})^{1/2} \\ &= v_v/[2(1 + A^{-1})]^{1/2}. \end{aligned} \tag{A7.50}$$

For the values of $A$ and $B$ given above, the ranges are not completely separated. In Fig. A2 we show the theoretical curves used in Fig. 5, together with the limiting lines just discussed.

For the middle or 193°K curve, we also show the fit that would be obtained if $c = 5 \times 10^5$ cm/sec, corresponding to $B = 87$ as for (A7.30). It is seen that this deviates much more from the data than does the curve based

on $c = 1.3 \times 10^6$ corresponding to $B = 12.8$. The deviation between theory and experiment would be still worse at 77°K, for which temperature the curve of Figure A2 fits the data well, as is shown in Figure 5.

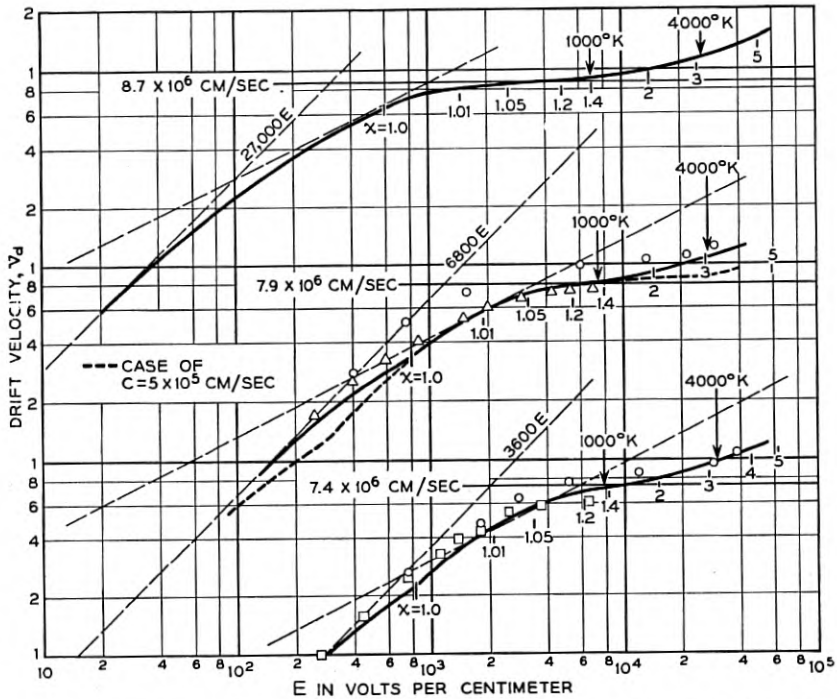



Fig. A2—The theoretical curves and their limiting forms. Some of the data from Fig. 2 are repeated here and some additional data from the original publication (Ryder and Shockley loc. cit.) are shown by crosses. A scale of values of $x$ and of approximate "temperatures" is also shown. The dashed curve for $T = 193°K$ is drawn for the case of the simple theory with $c \doteq 5 \times 10^5$ cm/sec; the change at 77°$K$ would be even more marked.

Above range III, the $Ax^2$ term makes an appreciable contribution. When $Ax^2$ becomes large compared to $B$, the approximation of taking the acoustical modes to be fully excited becomes questionable. This effect may be estimated by comparing $kT$ and the energy in an acoustical transition. The ratio is approximately

$$\frac{P_1 c}{kT} = \frac{mxv_v c}{kT} = \frac{h\nu}{kT} \cdot \frac{2c}{v_v} \cdot x$$
$$\doteq \frac{520}{300} \cdot \frac{2 \cdot 1.3 \cdot 10^6}{1.26 \cdot 10^7} x \doteq x/3. \qquad \text{(A7.51)}$$

On Fig. A2 we show the values of $x$. Equ. (A7.51) leads to values of $x < 3$ for 298°$K$ and $x < 2$ for 193°$K$. Although the approximation of $h\nu$ (acoustical) $< kT$ breaks down, extrapolating the acoustical scattering into the higher range involves some compensating effects.

The effective "temperature" $T_e$ of the electrons may be taken on the basis of this approximate treatment to be proportional to $v_1^2$. In order to make $T_e$ become equal to $T$ for zero field, we define $T_e$ by the equation

$$\begin{aligned} T_e &= Tv_1^2/v_T^2 = x^2\, h\nu/k \\ &= 520\, x^2. \end{aligned} \tag{A7.52}$$

Some temperatures deduced from this equation are also shown on Fig. A2.

Some of the data of Fig. 2 are also repeated in Fig. A2. In addition some earlier data[30] are also shown. These data extend to a somewhat higher range and appear to show the upward tendency predicted by the theory. The scale of temperatures indicates that for the extreme conditions experienced electron "temperatures" of about 4000°$K$ have been produced.

[30] E. J. Ryder and W. Shockley, *Phys. Rev. 81*, 139 (1950).

## Published writings of C. J. Davisson

Note on Radiation due to Impact of Beta Particles upon Solid Matter. Abstract.
*Phys. Rev.*, v. 28, ser. 1, pp. 469–470, June 1909.

Positive Thermions from Salts of Alkali Earths.
*Phil. Mag.*, v. 23, pp. 121–139, Jan. 1912.

Role Played by Gases in the Emission of Positive Thermions from Salts.
*Phil. Mag.*, v. 23, pp. 139–147, Jan. 1912.

Dispersion of Hydrogen and Helium on Bohr's Theory.
*Phys. Rev.*, v. 8, ser. 2, pp. 20–27, July 1916.

Gravitation and Electrical Action.
*Science*, v. 43, p. 929, June 30, 1916.

The Emission of Electrons from Oxide-coated Filaments under Positive Bombardment. (with Germer, L. H.)
*Phys. Rev.*, v. 15, ser. 2, pp. 330–332, April 1920.

The Electro-magnetic Mass of the Parson Magnetron. Abstract.
*Phys. Rev.*, v. 9, ser. 2, pp. 570–571, June 1917.

The Emission of Electrons from Oxide-Coated Filaments. (with Pidgeon, H. A.)
*Phys. Rev.*, v. 15, ser. 2, pp. 553–555, June 1920.

The Relation between the Emissive Power of a Metal and its Electrical Resistivity. (with Weeks, J. R.) Abstract.
*Phys. Rev.*, v. 17, pp. 261–263, February 1921.

Scattering of Electrons by Nickel. (with Kunsman, C. H.)
*Science*, v. 54, pp. 522–524, Nov. 25, 1921.

The Scattering of Electrons by Aluminum. (with Kunsman, C. H.) Abstract.
*Phys. Rev.*, v. 19, ser. 2, pp. 534–535, May 1922.

Secondary Electron Emission from Nickel. (with Kunsman, C. H.) Abstract.
*Phys. Rev.*, v. 20, ser. 2, page 110, July 1922.

Thermionic Work Function of Tungsten. (with Germer, L. H.)
*Phys. Rev.*, v. 20, ser. 2, pp. 300–330, Oct. 1922.

Scattering of Electrons by a Positive Nucleus of Limited Field.
*Phys. Rev.*, v. 21, ser. 2, pp. 637–649, June 1923.

Scattering of Low Speed Electrons by Platinum and Magnesium. (with Kunsman, C. H.)
*Phys. Rev.*, v. 22, ser. 2, pp. 242–258, Sept. 1923.

The Thermionic Work Function of Oxide-coated Platinum. (with Germer, L. H.)
Abstract.
*Phys. Rev.*, v. 21, ser. 2, page 208, February 1923.

Note on the Thermodynamics of Thermionic Emission.
*Phil. Mag.*, v. 47, ser. 6, pp. 544–549, Mar. 1924.

The Relation between Thermionic Emission and Contact Difference of Potential.
Abstract.
*Phys. Rev.*, v. 23, page 299, January 1924.

Relation between the Total Thermal Emissive Power of a Metal and its Electrical Resistivity. (with Weeks, J. R.)
*Opt. Soc. Am., Jl.* and *Rev. Sci. Instruments*, v. 8, pp. 581–605, May 1924.

Thermionic Work Function of Oxide-coated Platinum. (with Germer, L. H.)
*Phys. Rev.*, v. 24, ser. 2, pp. 666–682, Dec. 1924.

Note on Schottky's Method of Determining the Distribution of Velocities among Thermionic Electrons.
*Phys. Rev.*, v. 25, pp. 808–811, June 1925.

Are Electrons Waves?
*Bell Lab. Record*, v. 4, no. 2, pp. 257–260, Apr. 1927.

Diffraction of Electrons. (with Germer, L. H.)
*Phys. Rev.*, v. 30, pp. 705–740, Dec. 1927.

Note on the Thermionic Work Function of Tungsten. (with Germer, L. H.)
*Phys. Rev.*, v. 30, ser. 2, pp. 634–638, Nov. 1927.

Scattering of Electrons by a Single Crystal of Nickel. (with Germer, L. H.)
*Nature*, v. 119, pp. 558–560, Apr. 16, 1927.

Are Electrons Waves?
*Franklin Inst., Jl.*, v. 205, pp. 597–623, May 1928.

Attempt to Polarize Electron Waves by Reflection. (with Germer, L. H.)
*Nature*, v. 122, p. 809, Nov. 24, 1928.

Diffraction of Electrons by a Crystal of Nickel.
*Bell Sys. Tech. Jl.*, v. 7, pp. 90–105, Jan. 1928.

Reflection and Refraction of Electrons by a Crystal of Nickel. (with Germer, L. H.)
*Nat'l. Acad. Sci., Proc.*, v. 14, pp. 619–627, Aug. 1928.

Reflection of Electrons by a Crystal of Nickel. (with Germer, L. H.).
*Nat'l. Acad. Sci., Proc.*, v. 14, pp. 317–322, Apr. 1928.

"Anomalous Dispersion" of Electron Waves by Nickel. (with Germer, L. H.)
*Phys. Rev.*, v. 33, pp. 292–293, Feb., 1929.

Electron Waves.
*Franklin Inst., Jl.*, v. 208, pp. 595–604, Nov. 1929.
Electrons and Quanta.
*Opt. Soc. Amer., Jl.*, v. 18, pp. 193–201, Mar. 1929.
Scattering of Electrons by Crystals.
*Sci. Monthly*, v. 28, pp. 41–51, Jan. 1929.
Test for Polarization of Electron Waves by Reflection. (with Germer, L. H.)
*Phys. Rev.*, v. 33, pp. 760–772, May 1929.
Wave Properties of Electrons.
*Science*, v. 71, pp. 651–654, June 27, 1930.
Sir Chandrasekhara Venkata Raman, Nobel Laureate.
*Bell Lab. Record*, v. 9, pp. 354–357, Apr. 1931.
Conception and Demonstration of Electron Waves.
*Bell Sys. Tech. Jl.*, v. 11, pp. 546–562, Oct. 1932.
Diffraction of Electrons by Metal Surfaces. (with Germer, L. H.)
Abstract.
*Phys. Rev.*, v. 40, p. 124, Apr. 1932.
Electron Lenses. (with Calbick, C. J.)
Letter to the editor.
*Phys. Rev.*, v. 42, p. 580, Nov. 15, 1932.
Electron Particles as Waves. (with Germer, L. H.)
Abstract.
*Science*, v. 75, supp. pp. 10, 12, Mar. 4, 1932.
Electron Microscope. (with Calbick, C. J.)
Abstract.
*Phys. Rev.*, v. 45, p. 764, May 15, 1934.
Electron Optics.
*Sci. Monthly*, v. 39, pp. 265–268, Sept. 1934.
What Electrons can Tell us About Metals.
*Jl. Applied Phys.*, v. 8, pp. 391–397, June 1937.
Discovery of Electron Waves. Nobel Lecture.
*Bell Sys. Tech. Jl.*, v. 17, pp. 475–482, July 1938.
Laureation in Stockholm.
*Bell Lab. Record*, v. 16, pp. VII–XII, Feb. 1938.
Theory of the Transverse Doppler Effect.
*Phys. Rev.*, v. 54, pp. 90–91, July 1, 1938.
Double Bragg Reflections of X-rays in a Single Crystal. (with Haworth, F. E.)
Letter to the Editor.
*Phys. Rev.*, v. 66, pp. 351–352, Dec. 1 & 15, 1944.
Double Bragg Reflections of X-rays in a Single Crystal.
Letter to the Editor.
*Phys. Rev.*, v. 67, page 120, February 1 and 15, 1945.

## Contributors to This Issue

JOSEPH A. BECKER, A.B., Cornell University, 1918; Ph.D., Cornell University, 1922. National Research Fellow, California Institute of Technology, 1922–24; Assistant Professor of Physics, Stanford University, 1924. Engineering Department, Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Dr. Becker has worked in the fields of X-rays, magnetism, thermionic emission and adsorption, particularly in oxide coated filaments, and the properties of semiconductors as applied in varistors, thermistors and transistors.

R. M. BOZORTH, A.B., Reed College, 1917; U. S. Army, 1917–19; Ph.D. in Physical Chemistry, California Institute of Technology, 1922; Research Fellow in the Institute, 1922–23. Bell Telephone Laboratories, 1923–. As Research Physicist, Dr. Bozorth is engaged in research work in magnetics.

C. J. CALBICK, B.Sc. in E.E., State College of Washington, 1925; M.A. in Physics, Columbia, 1928. Bell Telephone Laboratories, 1925–. Here he has been engaged in the study of thin films on thermionic cathodes, electron diffraction problems, electron optics and microscopy, and in the development of high quality cathode-ray tubes for television reception. Member of American Physical Society, American Crystallographic Association, the Electron Microscope Society of America, the New York Microscopical Society and the I.R.E.

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. As Research Physicist, Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

L. H. GERMER, B.A., Cornell, 1917; M.A., Columbia, 1927; Ph.D., Columbia, 1927. Bell Telephone Laboratories, 1917–. With the Research Department, Dr. Germer has been concerned with studies in electron scattering and diffraction, surface chemistry, order-disorder phenomena, contact physics and physics of arc formation. Member of American Physical Society, the American Crystallographic Society of which he was president in 1944, the A.A.A.S., the New York Academy of Sciences and Sigma Xi.

FRANK GRAY, B.S., Purdue, 1911; M.A., Wisconsin University, 1913; Ph.D., 1916. U. S. Navy, 1917–19; Bell Telephone Laboratories, 1919–. His work has been chiefly research in microphone and relay contacts, gas discharge tubes, television, electron beam tubes, microwave tubes, PCM systems, and transistors. Fellow of American Physical Society and the A.A.A.S.; member of Gamma Alpha and Sigma XI; and Associate, I.R.E.

R. D. HEIDENREICH, B.S., Case School of Applied Science, 1938; M.S., 1940. Dow Chemical Company, 1940–45; Bell Telephone Laboratories, 1945–. Here he has worked chiefly on problems of surface metallurgy. Fellow of American Physical Society; member of A.A.A.S., the Electron Microscope Society of America and Sigma Xi.

A. N. HOLDEN, B.S., Harvard, 1925. Bell Telephone Laboratories, 1925–. In the Research Department his work has been chiefly in chemistry and solid state physics, primarily in originating new piezoelectric materials and in perfecting methods of growing crystals.

A. G. JENSEN, E.E., Royal Technical College, Copenhagen, 1920; instructor, 1921. Bell Telephone Laboratories, 1922–. He has been occupied chiefly in radio receiving studies, short-wave transatlantic telephony, coaxial cable development and television research. Fellow of I.R.E. and member of Society of Motion Picture and Television Engineers.

M. J. KELLY, B.S., Missouri School of Mines and Metallurgy, 1914; M.S., University of Kentucky, 1915; Ph.D., University of Chicago, 1918. Joining Bell Telephone Laboratories in 1918, Dr. Kelly became Director of Vacuum Tube Development in 1928; Director of Research, 1936; Executive Vice President, 1944; President, 1951. For the past year he has also served in an advisory capacity to the Air Force to assist in organizing its research and development. He holds honorary doctors' degrees from the University of Kentucky and the University of Missouri. In 1944 Dr. Kelly was awarded a Presidential Certificate of Merit and in 1945 was elected to the National Academy of Sciences. Member of Franklin Institute; Fellow of American Physical Society, I.R.E., Acoustical Society of America, A.I.E.E., and the American Association for the Advancement of Science.

L. A. MACCOLL, A.B., University of Colorado, 1919; M A., Columbia, 1925; Ph.D., 1934. Bell Telephone Laboratories, 1919–. He has been concerned chiefly with mathematical research and consultation. Visiting lecturer, Princeton, 1948–49; author of "Fundamental Theory of Servomecha-

nisms" (1945); member of American Mathematical Society, the Mathematical Association of America, the Edinburgh Mathematical Society, the American Physical Society, Phi Beta Kappa and Sigma Xi; and Fellow of New York Academy of Sciences.

W. P. MASON, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals and in the study of ultrasonics.

H. J MCSKIMIN, B.S., University of Illinois, 1937; M.S., New York University, 1940. Bell Telephone Laboratories, 1937–. Here he has worked chiefly on crystal filters, piezoelectric elements, ADP crystals, studies of the acoustic properties of liquids and solids. Member of Acoustical Society of America, Eta Kappa Nu, and Sigma Xi.

J. R. PIERCE, B.S., in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Dr. Pierce has been engaged in the study of vacuum tubes.

W. SCHOCKLEY, B.Sc., California Institute of Technology, 1932; Ph.D., Massachusetts Institute of Technology, 1936. Bell Telephone Laboratories, 1936–. Dr. Shockley's work in the Laboratories has been concerned with problems in solid state physics.

ELIZABETH A. WOOD (Mrs. Ira E.), A.B., Barnard, 1933; M.A., Bryn Mawr, 1934; Ph.D., Bryn Mawr, 1939. Research Assistant, Columbia, 1941. The next year she was awarded a National Research Fellowship and spent two years studying quartz deposits in the United States. Bell Telephone Laboratories, 1943–. Her work has been chiefly in X-ray diffraction, and the X-ray and optical investigation of crystals. Delegate to the Second International Congress of Crystallography in Stockholm in 1951. Member of American Physical Society, American Crystallographic Association, the New York Mineralogical Club, Phi Beta Kappa, Sigma Xi; and a Fellow of the Mineralogical Society of America.

# The TD-2 Microwave Radio Relay System

By A. A. ROETKEN, K. D. SMITH and R. W. FRIIS

(*Manuscript Received July 5, 1951*)

The TD-2 microwave radio relay system is a recent addition to the telephone plant facilities for long distance communication. It is designed to supplement the coaxial system and to provide greatly expanded facilities for nationwide transmission of broad-band signals such as television pictures or large groups of message circuits. The system makes use of many microwave repeaters located 25 to 30 miles apart in line-of-sight steps. The great variety and number of components which make up such a system require the engineering of all components to close tolerances. This paper describes the system in some detail from the standpoints of overall objectives, component designs to meet such objectives and facilities for the maintenance of overall performance.

## I. Introduction

SUPER-HIGH or microwave frequencies began to attract the interest of communication research engineers during the late '30s. The practical application of microwaves to commercial communication circuits was delayed by the outbreak of World War II, but the microwave techniques which had already been developed were employed to advantage in the prosecution of the war. The concentrated development effort and mass production of microwave equipment for military applications greatly expanded the engineering knowledge and production skill in this relatively new communications field. After termination of the war, it was possible again to devote the necessary development effort toward application of microwave techniques to commercial purposes. In the Bell System this effort was applied to the development and construction of a long-haul radio relay system.

A broad-band multi-channel radio relay system now connecting some of the main communication centers of the United States, as shown in Fig. 1, represents the combined efforts of a Bell System team since 1945.[1] This chain of stations carrying hundreds of message circuits or a television picture on each broad-band channel, in giant 25 to 30-mile strides across the country, has opened up a new radio field. The first step was the development of an experimental system placed in service in November 1947 between New York and Boston.[2] Upon the successful completion of this project objectives were established for a system, which is called the TD-2 Radio System, capable of extension to at least 4000 miles with upwards of 125 repeaters.

The TD-2 Radio System provides no new types of service but will supplement existing facilities such as the coaxial system. Therefore, TD-2 must provide comparable reliability, economy and quality of service. It is

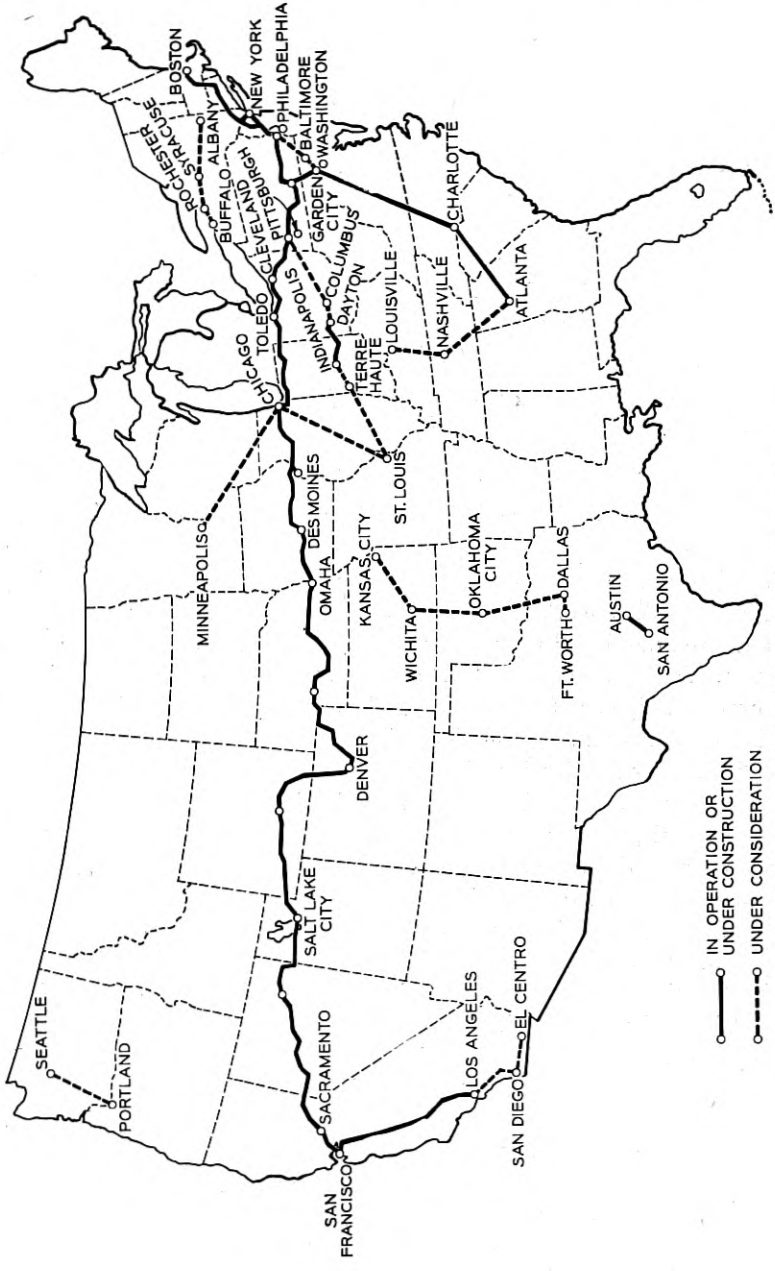



Fig. 1—TD microwave radio relay routes.

contemplated that by the end of 1951 there will be over 20,000 broad-band channel miles of radio relay in operation in the Bell System. Of this, about two-thirds will be used for television service and one-third to provide over 600,000 circuit miles of telephone circuits.

## II. TD-2 System—General

A radio relay system designed for long distances involves many problems new to radio but not new to long distance wire circuits. These problems are chiefly those of systems engineering to close transmission tolerances be-

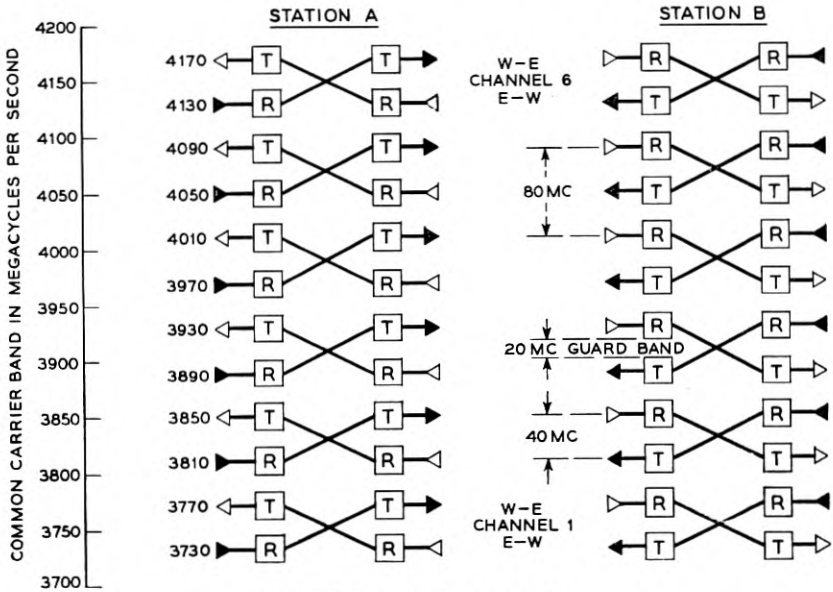



Fig. 2—TD-2 radio frequency plan.

cause of the many repeaters in tandem. To insure satisfactory systems operation, the transmission characteristics must remain stable over long periods of time to permit unattended operation. A reliable power plant and an alarm system are essential parts of the radio system.

### A. *Description*

The TD-2 Radio System utilizes frequency modulation and provides twelve broad-band channels, six in each direction, spaced 40 megacycles apart in the 3700–4200 megacycle common carrier band. A frequency assignment chart is shown in Fig. 2 and a systems block diagram in Fig. 3. Two broad-band channels in opposite directions provide a two-way message

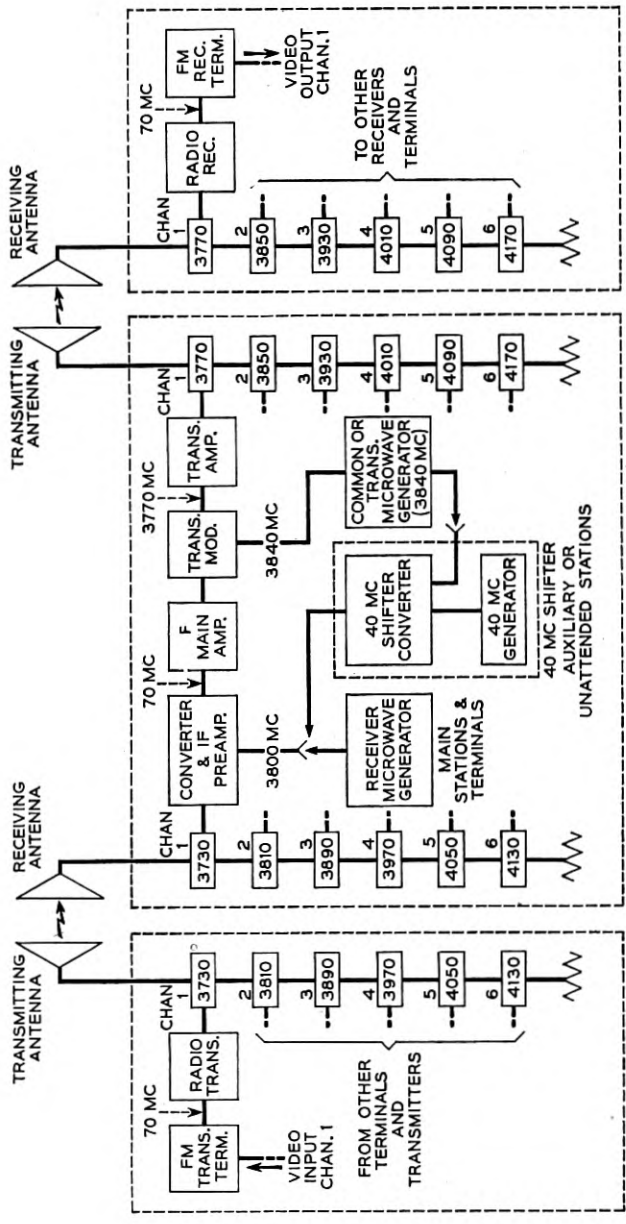



Fig. 3—Radio system block diagram.

system having a capacity of hundreds of 4 kc message circuits. Alternately each of these broad-band channels can be used to provide a 4-megacycle video circuit of the kind required for present day black and white television or they may be used to provide broader band television circuits if the need for such circuits develops. The video or message input to a channel is frequency modulated on a 70-megacycle carrier, translated up to the microwave band, amplified and combined with the microwave output of other channels in the same direction. The combined output is carried through a single waveguide to a directive transmitting antenna[3] beamed toward the next station. At a repeater point the six-channel signal is received on a single antenna, separated by means of channel separation networks, and each channel converted to a 70-megacycle IF band for amplification. At a through repeater point this 70-megacycle IF signal again modulates the 4000-mega-

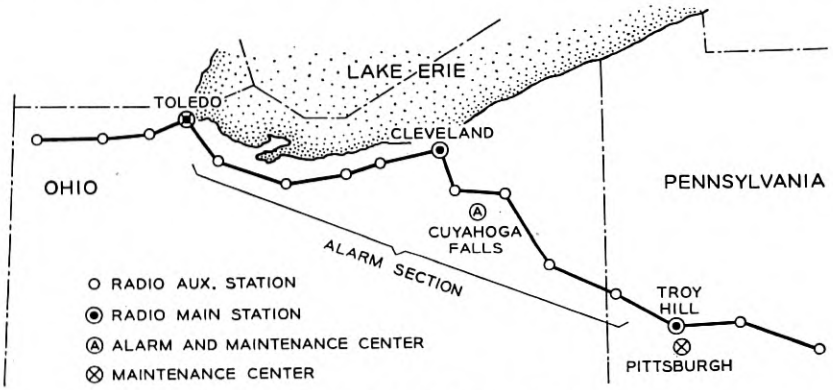



Fig. 4—Typical TD-2 route section.

cycle carrier, is amplified, added to the other channels in the same direction and delivered to the transmitting antenna. These are the simplified functions performed by the one-way radio repeater of the TD-2 System.

In order to feed a standard video or a multi-channel carrier signal into a TD-2 Radio System, an intermediate transmitter is required to frequency modulate these signals around a center frequency of 70 megacycles. This unit is known as the FM terminal transmitter. Likewise, at a receiving point an intermediate receiver is required to convert the 70-megacycle signal back to a video or carrier system signal. This unit is known as the FM terminal receiver.

A perspective of the system may be obtained from a typical route section as shown in Fig. 4. A long system is broken into sections by means of main repeater stations every few hundred miles. Auxiliary stations interconnect

the main stations. From an operating standpoint, main stations differ from auxiliary stations primarily in that each channel is terminated in IF switching circuits. This permits the removal of a channel for maintenance, or the replacement of a section which has failed by a spare circuit, by patching or remote control of the IF switching circuits. An alarm center may also be identified in Fig. 4 as an attended office to which a maximum of twelve repeater stations are connected by wire or radio for the purpose of reporting abnormal conditions that exist. Maintenance personnel are dispatched to unattended stations from this point. Not all TD-2 units are repaired and maintained at the radio stations. Maintenance centers are established along the route to service these units which require more elaborate test facilities than are provided at the stations. This requires the furnishing of certain spare units at the individual repeater stations.

### B. *Route Selection and Towers*

The interconnecting of two or more communication centers by a radio relay system presents many new problems in plant engineering. The selection of hundreds of mountain top sites to obtain line-of-sight transmission between stations, sites which are accessible to roads and power lines, sites which permit reasonable tower heights and which are an economic balance of these and other factors was a new challenge to the plant engineering force.[4] In brief, these were accomplished first by a detailed study of topographical and road maps, inspection of sites selected and finally the measurement of the transmission loss of the path.

The construction of towers several hundred feet high also involved new thinking by the building engineers.[5] The type of structure used on the New York-Chicago section of the TD-2 System was somewhat influenced by the availability of materials during 1948 and 1949. Concrete structures were used for this section of the system as shown in Fig. 5 with steel towers appearing on the Omaha to San Francisco section. Where steel towers are used, conventional type single-story buildings house the radio and associated equipment as shown in Fig. 6. Double antenna decks are provided on towers where branching radio routes are required.

## III. TD-2 Radio Equipment

### A. *Repeater—General*

The design of the TD-2 microwave repeater follows in principle that of its predecessor for the NewYork-Boston system.[2] Rapid advancement in the development of microwave vacuum tubes and other repeater components during the period from 1945 to 1947 led to a general improvement of repeater components for the TD-2 System. The realization in late 1947 of a

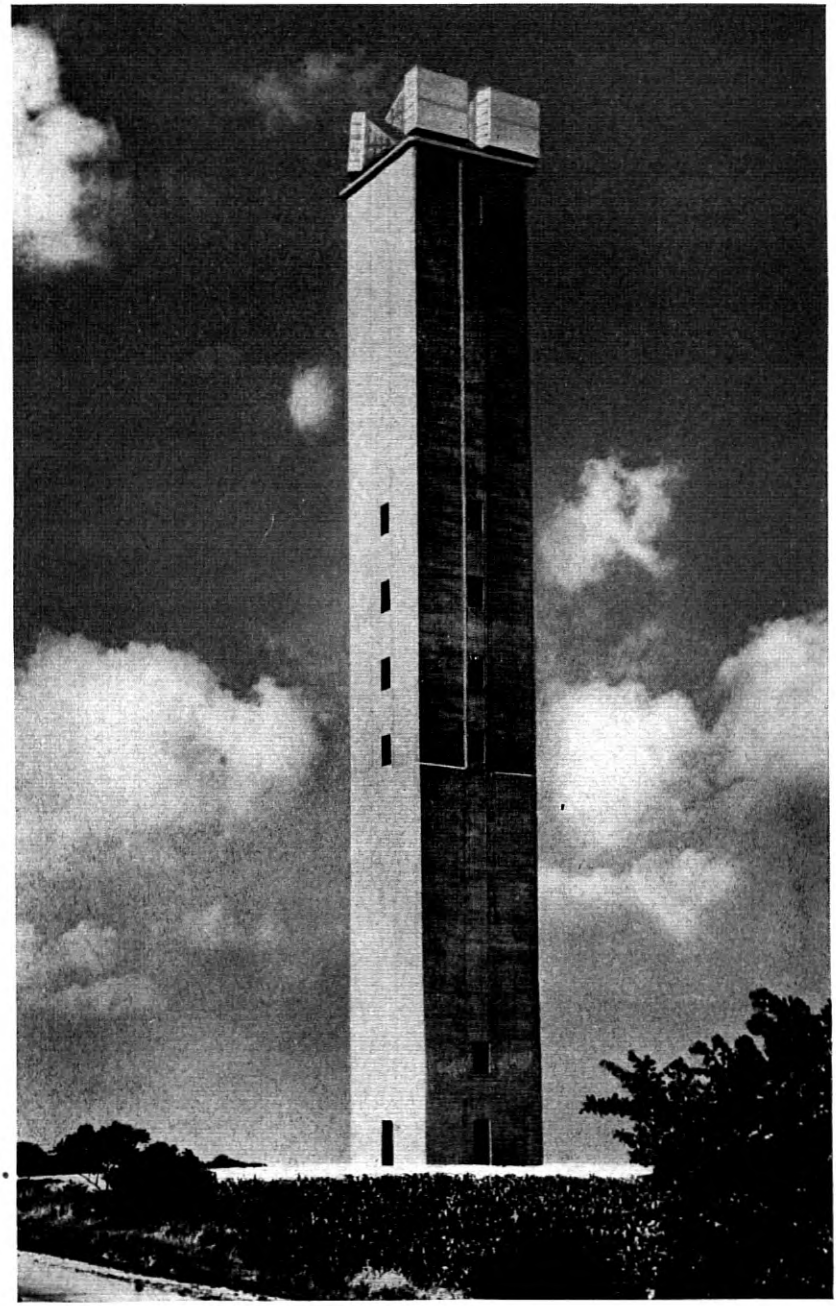


Fig. 5—190 foot concrete tower.

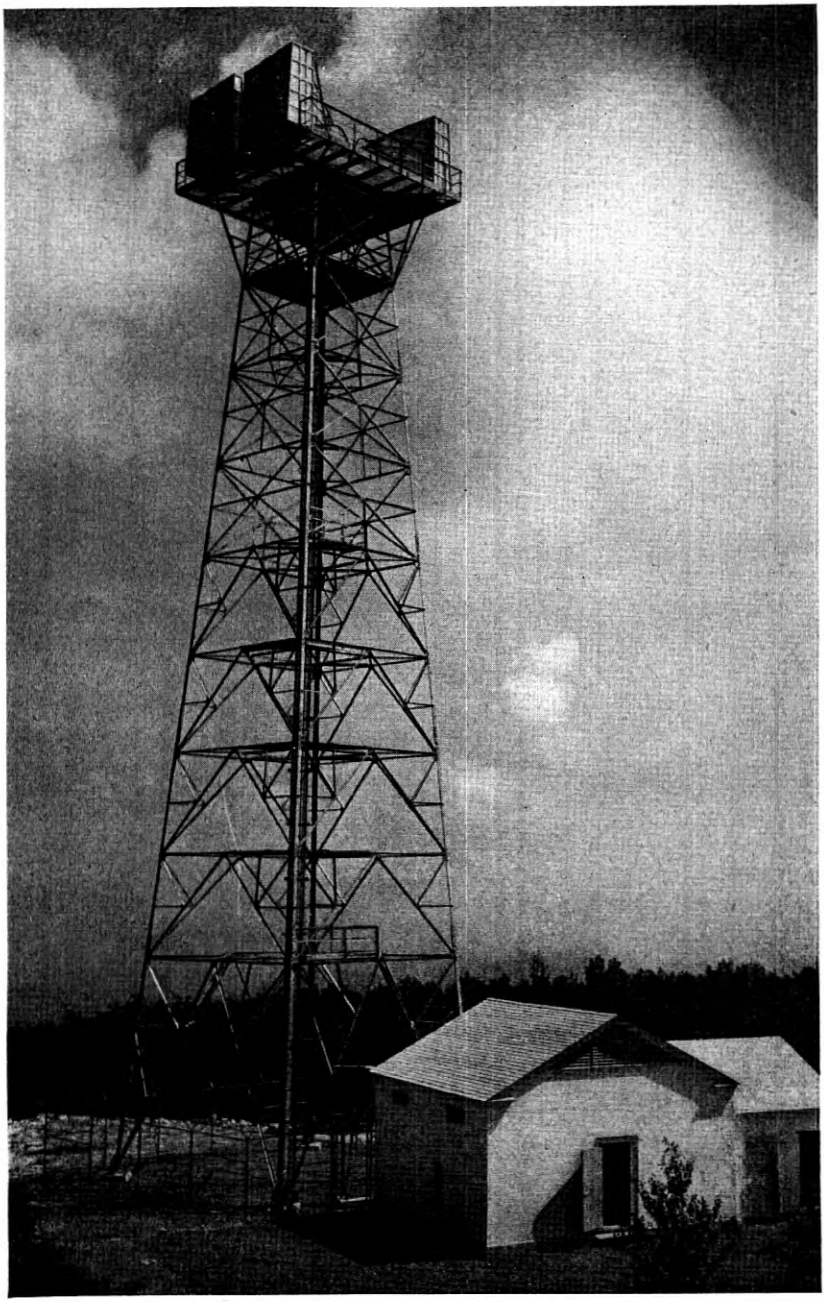


Fig. 6—125 foot steel tower.

practical triode amplifier for microwave application[6,7] was instrumental in determining a pattern for redesign, for it suggested the possibility of greatly simplifying the repeater while at the same time providing wider transmission bandwidths at greatly improved efficiency. By replacing the high voltage klystron (velocity variation type) amplifiers of the early repeaters with the new low voltage triodes, it became practical to design the system for battery operation—an important step toward increasing the system's reliability as it removed regulated rectifier tubes from the vulnerable portion of the system and eliminated the problem of hits during switchover from commerical to standby primary power. In the TD-2 System, vacuum tube heaters are generally operated from a 12-volt battery through dropping resistors,

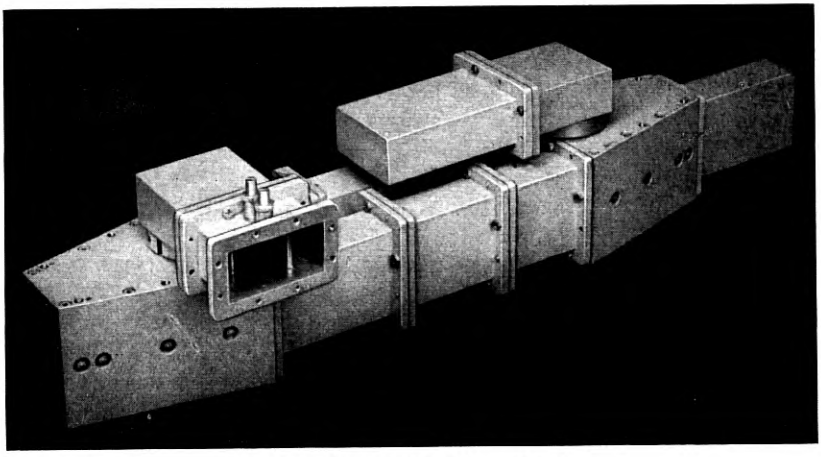


Fig. 7—Channel separation network.

thereby approximating constant heater power operation to increase tube life and reliability.

B. *Repeater Description*

A block diagram of a TD-2 repeater is shown in Fig. 3. An incoming microwave signal from a receiving antenna is selected by a channel separation network shown in Fig. 7. The signal is then combined in the receiver converter with energy from a beating oscillator source to provide an intermediate frequency signal band centered at 70 megacycles. Amplification, delay equalization and automatic gain control take place at the intermediate frequency of the radio receiver. In the transmitter this signal is combined in the transmitter modulator with a microwave source to provide a band offset 40 megacycles from the received frequency band. This signal is amplified and combined through transmitter channel separation networks with signals

from other channels for transmission through a common antenna. The 40-megacycle shift from receiving to transmitting frequencies is introduced in order to reduce the effect of crosstalk between transmitting and receiving antennas.

Main and auxiliary station repeaters differ in the following respects: An auxiliary repeater simply receives a particular channel signal, amplifies and transmits it to the next station. Here a common beating oscillator source for the transmitter and receiver, together with a stable 40-megacycle shifter, results in a systems frequency stability, for auxiliary stations alone, which is dependent only upon the stability of the 40-megacycle oscillator.[2] This feature cannot be used in the repeaters of a main station since each radio section between main stations must be independent of other sections for switching, branching, maintenance and terminating purposes. Here it is necessary to provide an independent oscillator source for each modulation process. In such an arrangement, errors in frequency add throughout the system and, therefore, the individual stability requirements for the oscillators are severe. In the TD-2 System this frequency stability is obtained by the use of a crystal controlled oscillator and harmonic generators. Two such microwave generators with temperature controlled crystals are used in each repeater bay at main stations, while one microwave generator and a 40-megacycle oscillator and shifter unit are used in each auxiliary station repeater.

A repeater bay using a 9-foot cable duct type framework is shown in Fig. 8. The top half of the bay contains the components which comprise the signal path through the repeater. These are the channel separation filters, image suppression filter,

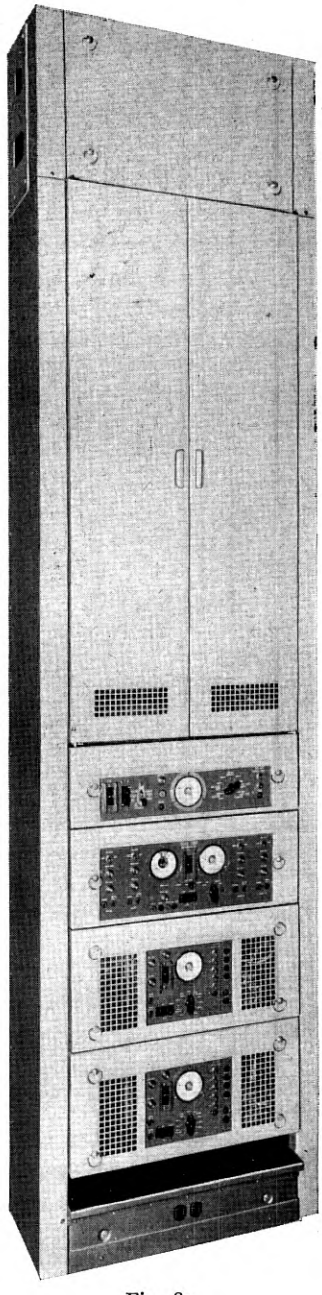


Fig. 8a

Fig. 8—Repeater bay.

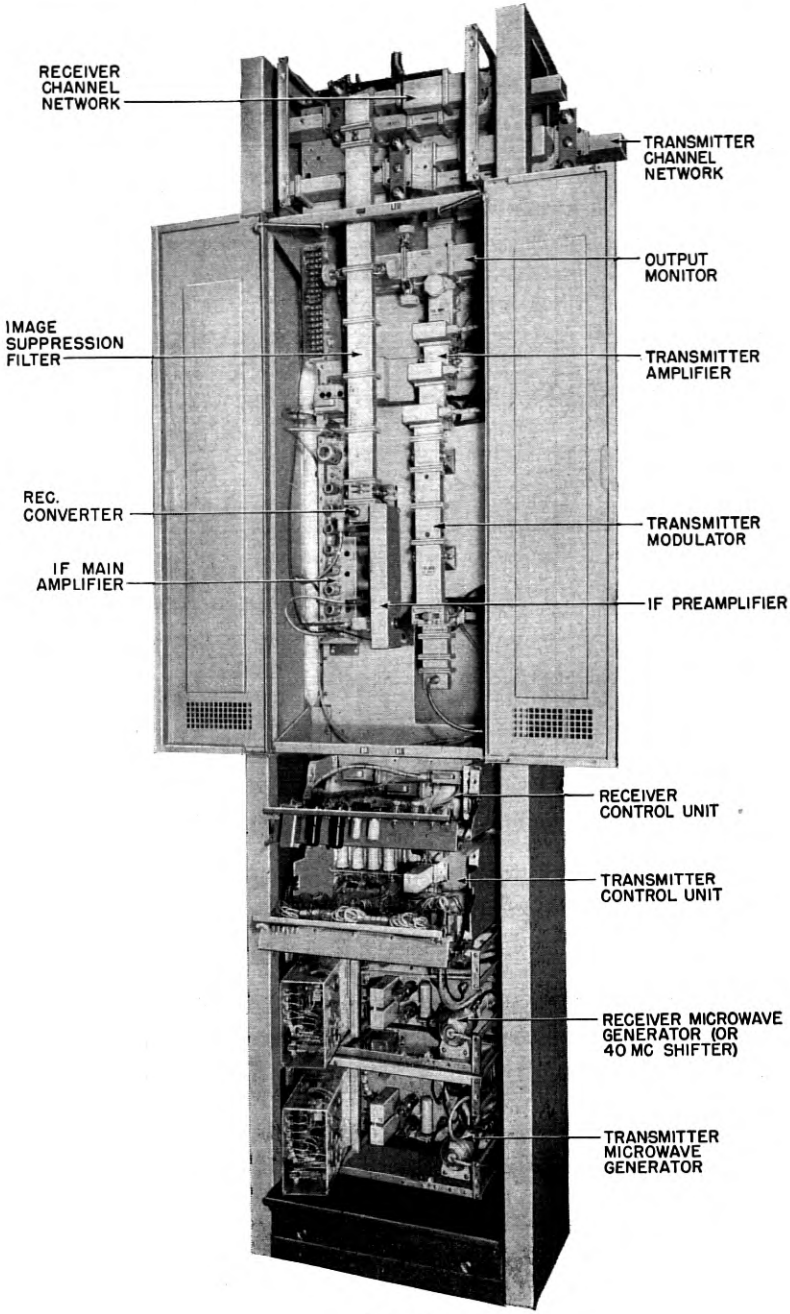

RECEIVER CHANNEL NETWORK
TRANSMITTER CHANNEL NETWORK
OUTPUT MONITOR
IMAGE SUPPRESSION FILTER
TRANSMITTER AMPLIFIER
REC. CONVERTER
TRANSMITTER MODULATOR
IF MAIN AMPLIFIER
IF PREAMPLIFIER
RECEIVER CONTROL UNIT
TRANSMITTER CONTROL UNIT
RECEIVER MICROWAVE GENERATOR (OR 40 MC SHIFTER)
TRANSMITTER MICROWAVE GENERATOR

Fig. 8b

receiving converter, IF preamplifier, delay equalizer, IF main amplifier, transmitting modulator and transmitting amplifier. The lower half of the bay contains four 19-inch wide oscillator and control units. These units, in the case of a main station repeater, are two microwave generators, a receiver control unit and a transmitter control unit. In the case of an auxiliary repeater, one of the microwave generators is replaced by a panel containing a 40-megacycle oscillator and shifter unit. All connections to the units of the bay are made by means of plugs and jacks for easy servicing.

A repeater receives a frequency modulated microwave signal at a normal level of about −38 dbm and transmits it at +27 dbm. Upward fades of 5 db and downward fades of 25 db are compensated to within about 1 db by automatic volume control action within the repeater. The amplitude characteristic is maintained flat to within 0.2 db over a 20-megacycle band.

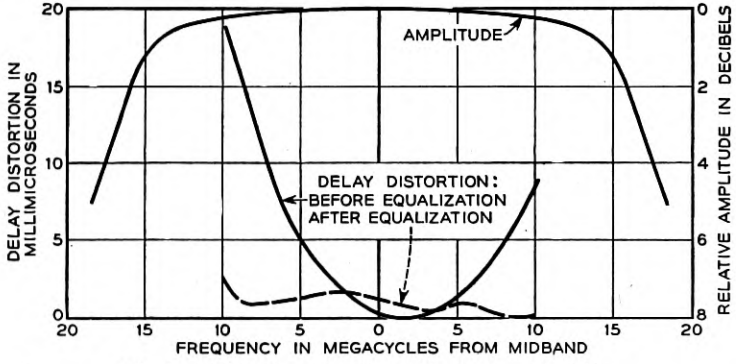



Fig. 9—Delay distortion & amplitude characteristics.

The amplitude and delay distortion characteristics of a repeater bay with and without delay equalization are shown in Fig. 9.

C. *Radio Receiver*

A channel separation network, as shown in Fig. 7, is required for each receiving channel. The network separates a particular channel from the six incoming 20-megacycle bandwidth channels for individual amplification and equalization in the repeater. It consists of two hybrid junctions and two band reflection filters which are tuned to the frequency band to be separated. An incoming signal is split into two parallel paths in passing through the first hybrid. A reflection filter in each of these paths returns the energy of the channel to be dropped to the first hybrid. By making the electrical path lengths from hybrid to filters differ by $\frac{1}{4}$ wavelength at the frequency of the channel to be dropped, the reflected signals are in phase op-

position at the hybrid and this results in transmission of the total signal energy through the fourth arm of the hybrid and into the receiving converter. The reflection filters are transparent to frequencies outside the desired 20-megacycle band. These frequencies are recombined in the second hybrid for transmission to the following channel separation networks.

An image suppression filter is located in the waveguide just ahead of the receiving converter. Its purpose is twofold: first, to provide discrimination against interfering signals which are the intermediate frequency image of the desired signal; and, second, to provide a critically spaced reflection of a beating oscillator component from the converter for control of the converter intermediate frequency output impedance.

The receiving converter is of the balanced crystal type in which the two crystals are mounted in a hybrid junction assembly. The signal input connection is by waveguide and the oscillator input connection is by coaxial line. An unbalanced output at intermediate frequency without the use of a balanced to unbalanced transformer is obtained by reversal of the polarity of one crystal relative to the other, permitting a parallel output connection. The 405A varistor unit, having symmetrical terminal design, was developed for this application. The preamplifier utilizes two 417A grounded grid triodes and has a gain of approximately 12 db. Its transmission band is centered at 70 megacycles and is flat to within 0.1 db over a 22-megacycle bandwidth. The converter-preamplifier has a net gain of approximately 6 db. The output of the preamplifier is coaxially connected through a delay equalizer to the input of the main IF amplifier.

The main IF amplifier shown in Fig. 10 has input and output impedances of 75 ohms and approximately 65 db gain. It consists of eight stages of amplification, the first being a 417A grounded grid triode followed by six stages of 404A pentodes and a 418A tetrode output stage. The input, output and interstage networks are all of the double-tuned impedance-matched type except the network between the sixth 404A pentode and the 418A output tetrode which is triple tuned and mismatched. Tuning of the triple-tuned network provides for adjustment of the over-all transmission band shape. Automatic gain control operating upon the grids of the first five pentode stages maintains the output power to within 1 db of a selected value between +4 dbm and +10 dbm for a 30 db range in input power. A low level bridging tap is available at the output of the main IF amplifier.

### D. *Radio Transmitter*

The transmitter modulator consists of a 416A microwave triode mounted in a structure which provides a resonant cavity between cathode and grid and another between plate and grid. This cavity structure is used in both

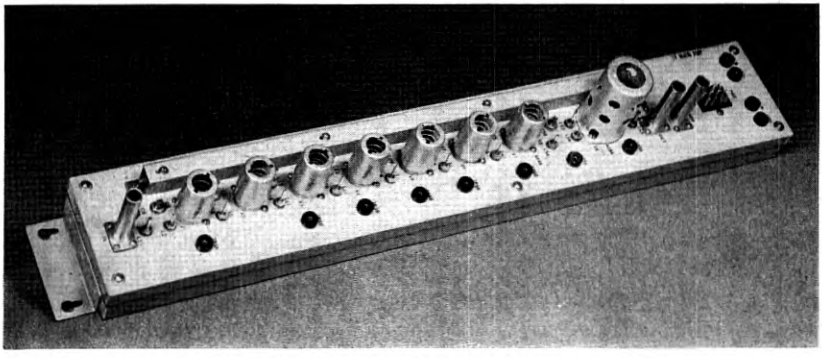


Fig. 10—Main IF amplifier.

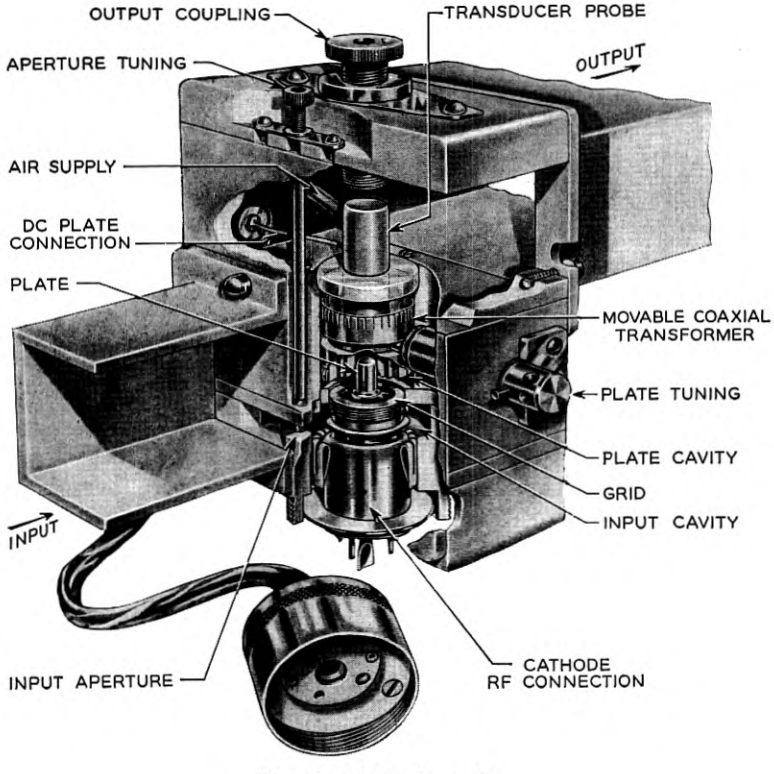



Fig. 11—416A tube cavity.

the modulator and the transmitter amplifiers and can best be described with the aid of a sectional view as shown in Fig. 11. The tube screws into the

cavity with the grid grounded directly to the body of the structure. The cathode of the tube is connected through an internal by-pass condenser to another part of the structure such that a cavity is formed around the tube between grid and cathode. An iris or aperture which is capacity tuned by a screw provides a means for coupling to the input waveguide. A coaxial cavity is formed around the plate which is tuned to the desired frequency by the movable coaxial transformer. The transformer couples the plate cavity to the transducer probe where the signal energy is transferred to the output waveguide.

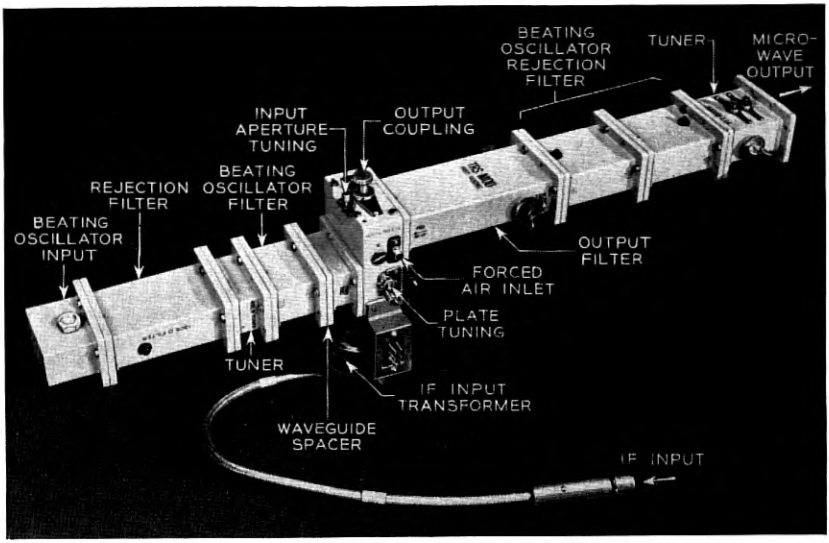



Fig. 12—Transmitter modulator.

The transmitter modulator is shown in Fig. 12. The oscillator power is applied to the cathode grid cavity through a tuner, a bandpass filter and a waveguide spacer. The IF power is applied between cathode and grid through a network which is mounted within a cylindrical compartment around the tube socket. The desired output sideband of the modulator is selected by a bandpass filter. Following this filter is a tuner unit which provides a means for adjusting the output impedance for a match with the following amplifier. A conversion gain of 9 db is realized in the process of shifting the IF frequency to the microwave band.

The modulator assembly is directly connected to the input of the transmitter amplifier, as may be seen in Fig. 8. An amplifier shown in Fig. 13 consists of three stages of 416A triodes mounted in cavity structures as

described above. The three stages are connected together in cascade through waveguide spacers and reactance tuners of such dimensions that the joining of each output cavity with the following input cavity (or filter section in the case of the output stage) forms a double-tuned critically coupled transformer. A flat over-all transmission characteristic is thereby obtained which is about 20 megacycles wide between points 0.1 db down. While capable of greater gain, the amplifier is adjusted to a gain of 18 db at an output power level of 0.5 watt. A double directional coupler in the waveguide between the transmitting amplifier and the transmitter channel separation filter provides monitoring and output alarm signals.

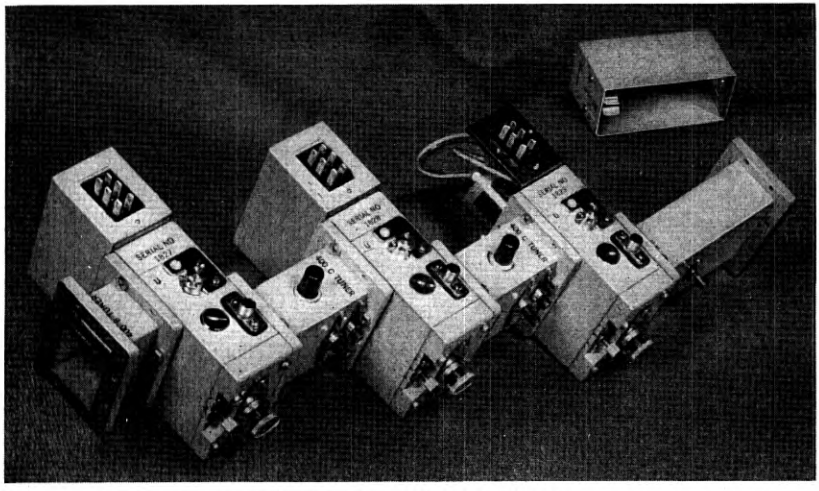


Fig. 13—Transmitter amplifier.

E. *Microwave Sources and Control Circuits*

The receiver control unit may be seen in Fig. 8. It contains a stabilized d-c amplifier for IF automatic gain control and level adjustments, and testing facilities for checking the performance of the receiver. The control unit also contains power controls and protection devices for the plate and filament circuits. The transmitter control unit contains controls for the application of power and bias to the transmitter and a means for metering various circuits.

The microwave generator, which furnishes about 200 milliwatts of beating oscillator power for the transmitter and receiver, is a stable microwave frequency source developed by harmonic generation from a quartz crystal in the vicinity of 18 megacycles. The multiplication takes place in six har-

monic generator stages, three doublers and three triplers. Only a few milliwatts of output power is required where the generator is used for the receiver beating oscillator alone, as in main stations or terminals. Here the final multiplier is operated as a sextupler, thereby permitting the elimination of the penultimate stage. At an auxiliary repeater, the receiving beating oscillator source is obtained from a 40-megacycle shifter converter, one input of which is from a part of the microwave generator output, and the other is from a crystal controlled 40-megacycle generator.

F. *Transmitter-Receiver Interconnections*

At auxiliary stations the IF output of the receiver is connected by a short coaxial line and 5 db resistance pad directly to the transmitting modulator in the same bay. This resistance pad is used as an impedance matching aid.

At main repeater stations the IF receiver output and the transmitter input are carried in coaxial lines to IF patching and switching equipment. With 30 to 60 feet of coaxial line between the receiver and transmitter, impedance match requirements are more severe than for short coaxial line connections. Here, a 6 db resistance pad is connected in the output line of the receiver and a 3 db resistance pad and buffer amplifier are connected in the input line of the transmitter modulator. The buffer amplifier consists of a single stage using a 418A tetrode and its gain may be set manually to provide $-1$ dbm to $+5$ dbm of signal power into the transmitter modulator as required. The bandwidth of the amplifier is approximately 20 megacycles and is sloped in such a manner as to approximately compensate for the small variation of loss over the band in the patching coaxial lines.

G. *IF Switching**

IF switching circuits are provided at terminals and main repeater points to facilitate maintenance operations as well as to provide flexibility for the changing requirements of network distribution. These switching and distributing operations are obtained by the use of unity gain amplifiers which are designated IF switching amplifiers and IF distributing amplifiers.

An IF switching amplifier functions as a single-pole double-throw switch for connection between intermediate frequency circuits of 75-ohm impedance. It has two input networks, each connected to a grid of a 404A pentode. The plates of the two tubes are connected in parallel to the output. Transmission through one or the other of the tubes is prevented by the application of a high negative grid bias to that tube. Switching the bias from one tube to the other thus permits the selection of either input signal. In most

* Prepared by T. R. D. Collins.

applications of the switching amplifier, signaling facilities are provided so that the switching operation can be controlled remotely.

An IF distributing amplifier provides three outputs from a single input, all at 75-ohm impedance. It consists of four 404A pentodes, the plate of one tube being connected to the grids of the other three tubes through an interstage network. Individual networks from the three output stages provide the desired distributing branches which are well isolated from each other electrically.

Switching and distributing amplifiers and a mounting framework are shown in Fig. 14. The two amplifiers have the same physical size and as many as five such units may be mounted in a frame on a plug-in basis. A number of such mounting frames are grouped and mounted on duct type bays to meet the needs of each switching and distributing location. Jack

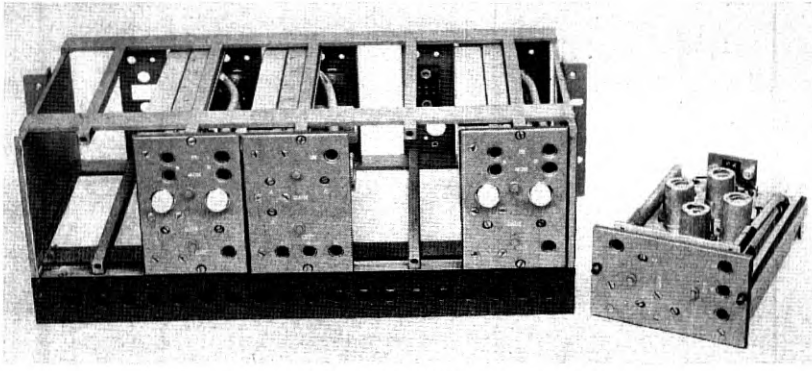


Fig. 14—Switching and distribution amplifiers.

fields associated with the mounting frames terminate the interbay coaxial trunks through which the switching and distributing connections are made.

Various combinations of switching and distributing amplifiers perform a large variety of interconnection functions within the system. Figure 15 indicates how these amplifiers may be used to replace a circuit which has failed by a spare circuit. At a transmitting terminal, the regular and spare channels may be paralleled. If a transmission failure occurs in channel 1 at one of the auxiliary repeater stations east of the main station, the failure of this channel is noted at the end of the system and service is switched to the spare channel 2. Since channel 1 is good except for the break east of the main station, the remote control for the switching amplifier in channel 1 is operated to switch output A of the channel 2 distributing amplifier to channel 1 radio transmitter. Thus channel 1 is connected in parallel with

channel 2 at this station and both a regular and a spare circuit are now available at succeeding stations.

H. *Automatic IF Switching**

At present IF switching is handled on a manual basis by attendants at the main stations or on a remote control basis over the order wire facilities. This type of switching is satisfactory for maintenance purposes but obviously is not fast enough to avoid a circuit interruption in replacing a circuit which has failed. The reliability of wire circuits will be difficult to meet in a long radio relay system without standby facilities because of vacuum tube failures and fading. Work is now under way to develop automatic IF switching facilities which will detect instantaneously a circuit

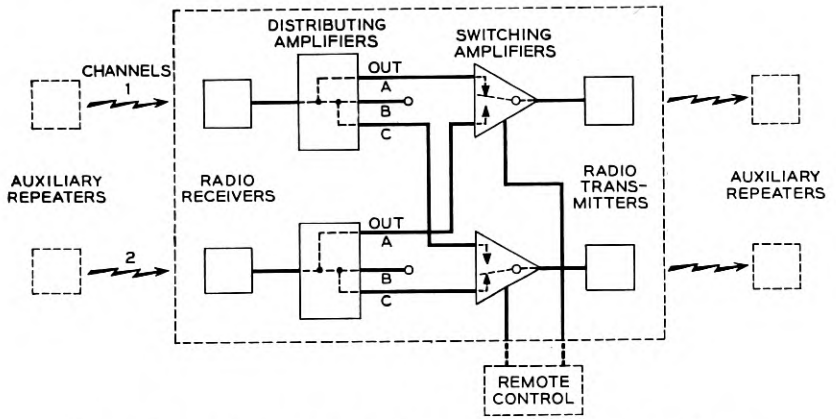



Fig. 15—IF switching and distributing amplifier. Interconnection diagram.

failure or an increase in noise on a radio channel and switch in a spare circuit for the poor section without circuit interruption. Fading data indicate that most deep fades which go beyond the range of the AGC circuit are of the selective type. Thus switching to a spare channel will provide frequency diversity advantages. With automatic switching it is believed the TD-2 System circuit outage time will not exceed that of wire circuits.

I. *Television Monitoring*

Visual monitoring facilities are provided at terminal and main repeater stations for observing circuit performance. Auxiliary repeater stations may also be so equipped in special cases. At transmitting or receiving terminals, monitoring connections are bridged to the video cables which run to operat-

* Prepared by T. R. D. Collins.

ing centers. The equipment units which make up the monitoring facilities are an auxiliary IF amplifier, an FM receiver, a video amplifier and a video monitor. A combination of these units is assembled in a bay to fit the needs of each monitoring location.

## IV. FM Terminal Equipment

### A. *General*

The TD-2 System will transmit a standard RMA black and white television signal or a band of message channels built up on a frequency division basis as provided by the coaxial cable message terminals. The FM terminal transmitter converts either of these signals to a frequency-modulated signal centered at 70 megacycles for application to the radio transmitter. The FM terminal receiver recovers the television or carrier signal from a frequency-modulated 70-megacycle signal. Thus the FM terminal equipment provides the connecting links between the TD-2 radio equipment and other facilities.

In a long system it may be necessary to bring the radio signal down to voice and back up to radio frequency many times in order to add and drop message groups. Each such process will require FM receiving and transmitting terminal equipment which consequently establishes severe linearity requirements for this equipment. An objective in the development of the terminals was to meet long haul systems performance requirements with sixteen pairs of terminals in tandem.

### B. *FM Transmitter*

A functional diagram of the FM terminal transmitter is shown in Fig. 16. It accepts a signal from an unbalanced 75-ohm line and delivers an FM signal centered at 70 megacycles to the radio transmitter. The input level may be adjusted from 0.2 volt to 2.5 volts peak-to-peak with an output level of 13 dbm at an impedance of 75 ohms. For television transmission with a $\pm 4$ megacycle swing the tips of the synchronizing pulses are at 74 megacycles and the picture white at 66 megacycles. For message service the nominal deviation is centered about 70 megacycles. For television transmission the output is automatically clamped to a predetermined frequency during each synchronizing pulse. These differences in operation are described in more detail below.

#### 1. *Description*

The input signal to the FM transmitter is applied through an adjustable attenuator to a video amplifier consisting of two similar three-stage feedback amplifiers in tandem which have a combined gain of 42 db. The video

amplifier output is applied to the repeller of a deviation oscillator described below.

A microwave heterodyne method of generating a 70-megacycle FM signal was selected because it was found possible to design a highly linear deviator in the microwave region. It also allows separate tests to be made of the transmitter and receiver linearity and thus facilitates maintenance.

A reflex klystron oscillator may be frequency-modulated by superimposing a modulating signal on the repeller d-c voltage. The rate of frequency change with change of repeller voltage passes through a minimum near the

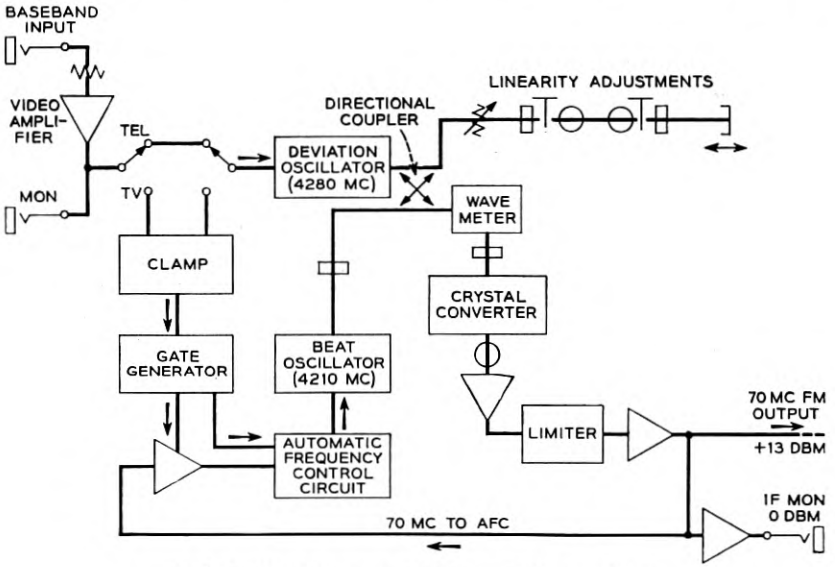



Fig. 16—Block diagram of FM terminal transmitter.

point of maximum power output. At a deviation of ±4 megacycles, the difference in FM sensitivity over the 8-megacycle swing would normally be sufficient to produce intolerable distortion. However, the operating frequency of a reflex oscillator is subject to modification by the load impedance seen by the oscillator. This effect is commonly called "pulling." In the deviation oscillator, this effect is made use of to provide deviation linearity over a range of more than 10 megacycles. The load circuit for the 4280-megacycle deviation oscillator consists of a variable attentuator, a short length of line, and a variable position short circuit. Adjustments of these two variables allows complete control of the reactance seen at the output

of the deviation oscillator. The length of circuit to the movable short is so chosen (about 35 inches) to provide the optimum rate of change of reactance with frequency. At optimum adjustment, the reactive component of the load pulls the frequency of the generator by just the amount necessary to straighten out the deviation curve. The deviation sensitivity is at the same time increased about 25%, which reduces the required video driving voltage.

A portion of the output signal of the deviation oscillator is fed through a directional coupler to a crystal microwave converter where it is mixed with a 4210-megacycle signal from another klystron to produce a 70-megacycle FM signal. The microwave output from the deviation oscillator is about 50 milliwatts, and after losses in the directional coupler and converter about one milliwatt of 70-megacycle FM output is available. This signal is amplified in a broad-band limiter-amplifier for application directly to the radio transmitter or indirectly through appropriate switching circuits.

2. *Clamper and AFC Circuit*

For television transmission the voltage supplied to the repeller of the deviation oscillator is clamped to a predetermined negative value during each synchronizing pulse in a conventional manner. This clamping action enables the transmission of video signal components down to direct current. For message telephone transmission the clamping circuit is disabled.

The automatic frequency control circuit used to control the frequency of the beat oscillator provides a high gain and stable AFC without a d-c amplifier. As shown in Fig. 16, a portion of the 70-megacycle output signal is diverted and after passing through a gated amplifier is applied to a discriminator. The discriminator network is of conventional design and the detector elements are germanium diodes. The direct-current output voltages are applied to the grids of two triodes acting as a pulse modulator. The anodes of these triodes are supplied with a high level positive pulse used for gating from a blocking oscillator associated with the clamper circuit. This oscillator is free running for message signals but is triggered by the synchronizing pulses when video signals are being transmitted. The unbalance voltage on the triode grids controls the amplitude and polarity of the pulse produced by this modulator. After two stages of a-c. amplification this error signal is combined with a second high level pulse from the same blocking oscillator source in a phase detecting circuit, and, after integration, the d-c. output of this detector is used for AFC. With television operation the gated amplifier operates only during synchronizing pulses, and the discriminator is adjusted for an output frequency of 74 megacycles. With multi-channel message operation, the gated amplifier is operated as a straight-through amplifier, and the discriminator is adjusted to hold an average output frequency of 70 megacycles.

C. *FM Receiver*

The FM receiver contains an IF amplifier, limiter, discriminator, and video amplifier, as indicated in Fig. 17. The input amplifier consists of two stages, each using a 404A pentode, with broad-band interstage networks. The two-stage instantaneous amplitude limiter has biased silicon varistors shunting the single-tuned plate loads of each of the 418A tubes. The bias voltages are so adjusted that the load impedance is high for signal voltages less than about one volt, and very low for any larger signal.

1. *Discriminator*

The discriminator circuit follows early conventional practice, in that two separately driven antiresonant circuits are used. The signal at the limiter output is fed to two 404A amplifier stages, one tuned above the signal band,

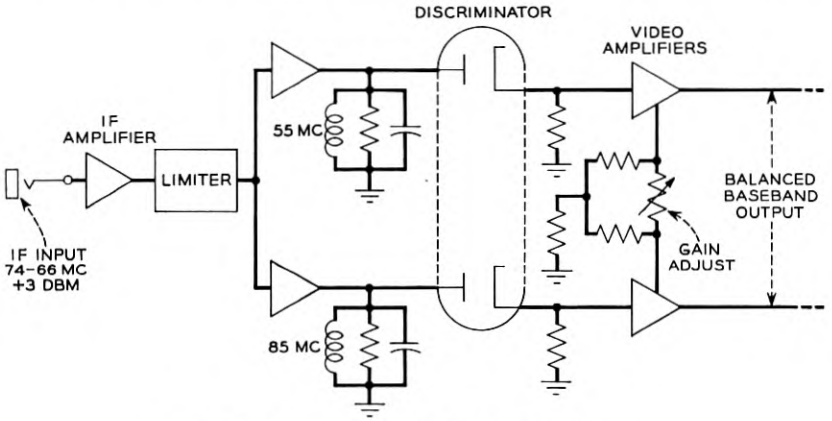



Fig. 17—Block diagram of FM terminal receiver.

the other below. The frequency-modulated signals produce amplitude variations of the voltage across these tuned circuits which are detected by diode rectifiers and applied to the video amplifier. A potentiometer in the cathode interconnection of the amplifier tubes provides a balance adjustment for the discriminator.

2. *Video Amplifier*

The video amplifier is a three-stage resistance-capacity coupled unit having negative feedback in each symmetrical half, and negative feedback to longitudinal voltages through a common cathode resistor. The gain is adjustable over a range of several db by means of a dual potentiometer which varies the common cathode resistance in each half of the amplifier. Whenever such an adjustment is made, a constant loop gain is maintained in the feedback system by varying simultaneously the local cathode de-

generation in the middle stages of the amplifier. A peak-to-peak voltmeter connected across one side of the balanced output is used to monitor the transmission level of television signals.

## V. System Maintenance and Test Equipment

### A. *General*

Most TD-2 stations are operated on an unattended basis. Test equipment is provided at each terminal, auxiliary and main station to perform the necessary maintenance functions. This consists of a radio test bay as shown in Fig. 18 for each auxiliary, main and terminal station, and an FM test console as shown in Fig. 19 at terminals and main stations where FM terminal equipment is provided. The philosophy is to provide sufficient test equipment at each station to isolate the trouble. When the unit in trouble requires extensive tests or repair, a station spare is substituted and the faulty unit is returned to a maintenance center. Maintenance centers are usually located in existing telephone offices along the route.

In maintaining the radio equipment each repeater bay is adjusted to provide a transmission band 20 megacycles wide, flat to within two-tenths of a db and centered about the assigned channel frequency. Trimming adjustments are provided on the receiver and transmitter to obtain this characteristic. This test involves the use of a swept signal source which is divided into a reference path and a path through the equipment under test, each of which is terminated in an identical detector. The outputs of these detectors are alternately applied to the vertical deflection amplifier of an oscilloscope at a 30-cycle rate, while a voltage proportional to the frequency excursion is applied to the horizontal amplifier. Generally, the vertical gain of the oscilloscope is adjusted so that a separation of one inch between the test and reference traces corresponds to a level difference of 1 db and the horizontal gain is adjusted so that one inch corresponds to a 10-megacycle frequency excursion. The reference trace is then matched to the test trace by adjustments of the equipment under test. The waveguide attenuators and directional couplers shown in Fig. 18 provide for testing over a wide range of levels.

### B. *Radio Test Bay*

The radio test bay contains a microwave swept frequency oscillator, a combined microwave and IF power meter, a cathode ray oscilloscope, RF and IF wave meters, detectors and attenuators and associated power supplies.

The microwave sweep oscillator is adjustable in sweep range up to 70 megacycles over the 3700 to 4200 megacycle band. The frequency is swept

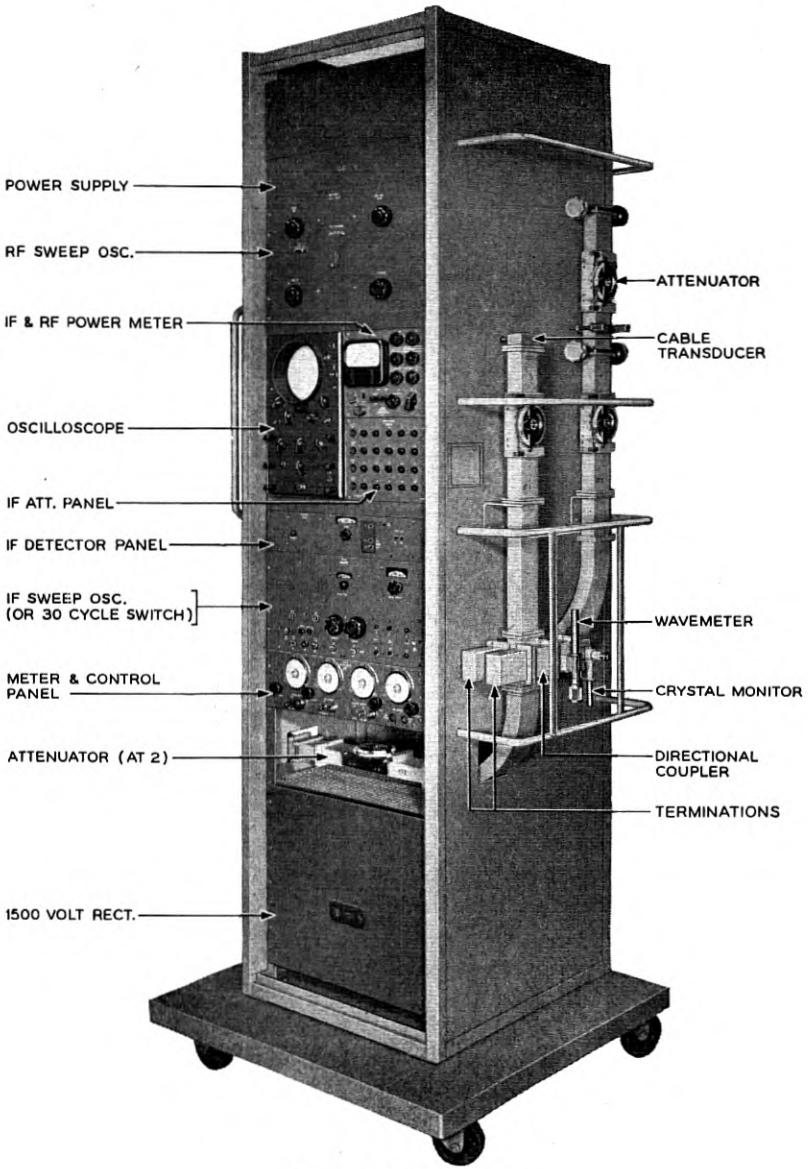



Fig. 18—Radio test bay.

by a motor driven reactive element in one of two cavities associated with the 402A velocity variation oscillator tube.

The RF and IF power meter consists of a temperature compensated thermistor bridge unit. It has separate input arrangements for the 3700 to 4200-megacycle and 50 to 90-megacycle bands. Accurate measurements of power may be made in the range from $-10$ dbm to $+6$ dbm.

The test bay used at maintenance centers has, in addition to the above equipment, a 50- to 90-megacycle swept frequency oscillator and associated detectors for the testing of intermediate frequency components. The opera-

Fig. 19—FM terminal test console.

tions carried out at the maintenance center include the repair and realignment of defective equipment returned from the radio stations. The maintenance center test equipment includes facilities for accurate impedance match measurement in the microwave and IF range, for varistor matching tests, vacuum tube transconductance tests and general component tests which cannot be made at the radio station. Usually the maintenance centers are operated by the same staff that maintains the radio stations in the section.

C. *FM Terminal Test Console*

The terminal test console shown in Fig. 19 is used to measure FM deviation, linearity of the FM transmitter and receiver and for routine monitoring of wave forms at video frequencies. The equipment includes a conventional CW signal generator covering the range of 50 to 90 megacycles, a video "A" scope, an electronic switch and patching and terminating facilities. A rather unique linearity test set described below and an FM terminal receiver are also included.

1. *Deviation Measurement*

For deviation measurements, the IF signal being monitored is patched into one input of the IF electronic switch which switches between inputs at a 1200-cycle rate, and the CW signal generator into the other input. After detection by the FM receiver, the signals are applied to the oscilloscope and a straight line corresponding to the CW generator frequency is displayed superimposed on the video signal. By adjustment of the CW reference frequency, the instantaneous frequency of any signal component may be determined.

2. *FM Receiver Linearity*

For a measurement of linearity of the receiver discriminator, the linearity test set is connected to an FM transmitter which is patched to the receiver under test. The linearity test set supplies a low level 100 kc modulating voltage to the deviation oscillator of the transmitter and a high level 60-cycle voltage to the transmitter beat oscillator. For this test the transmitter AFC circuit is disabled. Under these conditions the signal applied to the receiver discriminator swings over approximately 10 megacycles at a 60-cycle rate and over a small range of less than one megacycle at a 100 kc rate. The 100 kc video component in the receiver output is then proportional to the slope of the discriminator response curve. The envelope of this 100 kc amplitude is recovered in the linearity test set and the a-c. component is applied to the oscilloscope vertical amplifier. The horizontal deflection is synchronized with the 60-cycle deviation. A 30-cycle switch changes the amplitude of the 100 kc signal by a calibrated amount to provide two separated traces on the screen and make the device self-calibrating.

3. *FM Transmitter Linearity*

For a measurement of transmitter linearity, the same setup used in the receiver test is made use of except that both the 100 kc small signal and 60-cycle large signal are applied to the deviation oscillator of the transmitter under test. The beat oscillator AFC circuit is allowed to operate with a time constant sufficiently rapid to follow the 60-cycle fluctuation of the deviation oscillator, but not the 100 kc component. Thus the 100 kc modulation component is applied over a 10-megacycle range of the deviation oscil-

lator characteristic, but is applied to the receiver at a fixed (70-megacycle) frequency, so that the receiver discriminator does not enter the measurement except as a fixed gain detector. While the transmitter is being tested as above, the magnitude and phase adjustments of the deviation oscillator load impedances are made as required to meet the desired linearity of deviation which is normally 1% over the 10-megacycle range.

## VI. C1 Alarm and Control System*

The operation and maintenance of unattended repeater stations require a flexible and reliable alarm system whose performance is commensurate with the importance of the toll and television program services handled by the TD-2 System. The Cl alarm and control system has been developed for this purpose and, as its name implies, it serves two functions. The first is that of transmitting detailed alarm information from unattended repeater stations to the responsible alarm centers. The second function is that of transmitting orders, or remote control signals, from alarm centers to unattended stations.

The salient features of the C1 system may be summarized as follows:

1. It is a voice-frequency system, thus permitting its use with equal facility on cable pairs, open wire lines, or radio channels (or combinations thereof) capable of transmitting a 3000-cycle voice band.
2. It transmits a maximum of 42 separate alarms or indications from each unattended station to its associated alarm center.
3. It transmits a maximum of ten remote control orders in the opposite direction, that is, from an alarm center to each unattended station for whose operation it is responsible.
4. A maximum of twelve unattended stations may be associated with one alarm center.

A typical section of the TD-2 Radio Relay System is shown in Fig. 4. The alarm center for the section indicated is at Cuyahoga Falls, which in this case is also a maintenance center. Alarm centers and maintenance centers may be located at any attended central office or repeater station on existing cable and open wire routes.

Alarm signals are transmitted to the alarm center from the unattended station over a one-way, two-wire circuit as shown in Fig. 20. A four-wire local order circuit is used for voice communication between adjacent main radio stations and the intermediate unattended auxiliary repeater stations. The alarm centers and maintenance centers in that alarm section are also bridged on it. Remote control order signals from the alarm center are of such short duration that they can be transmitted without objectionable interference over one side of this four-wire local order circuit. An express

* Prepared by C. E. Clutts and G. A. Pullis.

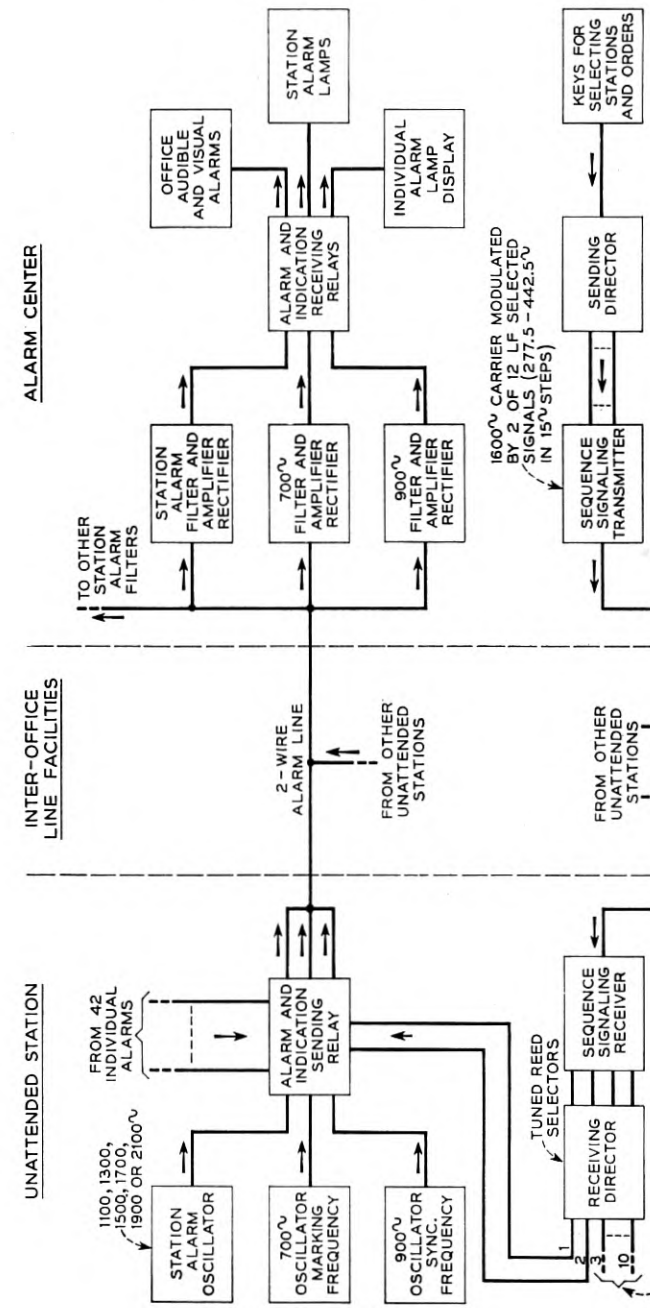



Fig. 20 – Block diagram of C1 alarm and control system.

order circuit is used to link the terminal stations of a system with the main stations, alarm centers and maintenance centers for system-wise radio maintenance and traffic control.

A. *Station Alarms*

Each unattended station transmits a continuous and distinctive tone to its associated alarm center. Interruption of this tone for approximately ten seconds registers an audible and visual alarm at the alarm center, and, because each station is assigned a different frequency, the station whose tone is interrupted is easily identified. As many as six stations can report to an alarm center over a one-way, two-wire alarm line. Six more alarm sending stations can report to the same alarm center by bridging them on a second alarm line (usually in the opposite direction from the first), and providing a second set of receiving filters with associated amplifiers and detectors at the alarm center. In this manner an alarm center can identify the alarms from a maximum of twelve unattended repeater stations. The six station frequencies that can be used on one alarm line are 1100, 1300, 1500, 1700, 1900, and 2100 cycles.

Each station alarm tone is selected at the alarm center by its associated receiving filter and individual amplifier-rectifier circuit. Automatic gain control action in the amplifier circuit permits a tone from the alarm line to vary $\pm 6$ db from its normal value without interfering with the proper operation of the system.

B. *Individual Alarm Indications*

The station alarm reports that a particular unattended point is in trouble, but it does not tell what the specific trouble is. Supplementing the station alarm circuit is an individual alarm indication circuit that reports which, if any, of 42 possible alarm conditions exist at an unattended station. The alarm indication sending circuit does not start automatically but only in response to an order sent out from the alarm center. Thus, after receiving a station alarm, an attendant at the alarm center sends an order over the control system described later to that particular station directing it to scan the individual alarms and report those that have operated.

The report is transmitted over the alarm pair and received on a miniature lamp bank located in the key shelf of the alarm receiving bay at the alarm center. Of a total of 60 lamps in the key shelf, 42 are used for alarm indications, 8 for identifying the six east or west reporting stations, and 10 for checking synchronization of the indication sending and receiving circuits. Figure 21 is a copy of the form which is placed over the lamp display to

FORM E-3/94 (1-50)

## C1 ALARM RECORD

SERIAL NO. _____

DATE _____ ACTION TAKEN _____

TIME RECEIVED A P BY _____

SENDING OFFICE _____ TROUBLE FOUND _____

RECEIVING OFFICE _____ DATE OK _____ TIME A P BY _____

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| | STATION IDENTIFICATION | | | | | |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| | SYNCHRONIZATION-START | | | | | GROUP A |
| 2 | ON | ON | OFF | ON | ON | |
| | SYNCHRONIZATION-STOP | | | | | GROUP B |
| 3 | ON | ON | OFF | ON | ON | |
| | LOW MICROWAVE OUTPUT E-W OR N-S CHANNELS | | | | | |
| 4 | 1 | 2 | 3 | 4 | 5 | 6 |
| | LOW MICROWAVE OUTPUT W-E OR S-N CHANNELS | | | | | |
| 5 | 1 | 2 | 3 | 4 | 5 | 6 |
| | LOW MW OUTPUT BRANCH CHANNELS | | | | | |
| 6 | A | B | C | D | | |
| | DISCH. FUSES | DISTRIBUTION FUSES | | | | OBSTR. LIGHTS OFF |
| 7 | 12V, 24V 130V 250V | RADIO 12V 130V 250V | ABS 24V 130V | MISC. 24V 130V 115 AC | BOTH TOP | SIDE OR ONE TOP |
| | COM'L. AC PWR. | | HIGH-LOW VOLTAGE | | | |
| 8 | FAIL | RESTORE | 12V | 24V | 130V | 250V |
| | GAS ENGINE | | | RECT. FAIL | H-L FLOAT | |
| 9 | FAIL | OPER. | LOW GAS | 12V 130V 250V | 12V, 24V 130V 250V | OPEN DOOR |
| | HIGH-LOW TEMP. | | | TUBE COOLING FAIL. | | |
| 10 | CRYSTAL OVEN | ROOM | WG LOW GAS PRESSURE | ONE BLOWER FAIL. | AIR FAILURE | |
| | A | B | C | D | E | F |

Fig. 21—C1 alarm record.

designate the lamps and provide a record of a specific alarm condition at an unattended station.

The alarm indication sending and receiving circuits utilize relay counting chains which scan over the 60 possible indications at a 5-cycle rate and

cause a 900-cycle pulse to be sent back to the alarm center for each indication scanned. Whenever an alarm condition or other indication is encountered, a 700-cycle pulse is transmitted simultaneously with the 900-cycle pulse. At the alarm center the pulses are selected by 700- and 900-cycle filters, and amplified and detected in the same manner as the station tones. The resultant d-c. pulses operate relays which in turn light particular lamps in the key shelf lamp display panel in the alarm center receiving bay whenever the two pulses are received simultaneously.

### C. *Sequence Signaling Remote Controls*

As mentioned earlier the C1 alarm and control system is capable of transmitting as many as ten orders from the alarm center to a particular station in trouble. Typical orders to a repeater station may be an order to scan all alarm indications or an order to start the gas engine alternator. Sequence signaling transmitters and receivers are employed for the transmission of orders to the unattended repeater stations. Sequence signaling is an arrangement in which two separate signals sent in a predetermined sequence are translated by the receiver into an order. One hundred and thirty-two different orders can be transmitted from an alarm center through sequence combinations of two out of twelve modulating frequencies available in 15-cycle steps from 277.5 to 442.5 cycles. The C1 system makes use of 120 of these orders at those alarm centers which remotely control as many as twelve unattended repeater stations.

An attendant initiates an order by operating the key of the station to be called, the proper order key and a start key. This operation selects the proper two low frequencies which modulate a 1600-cycle carrier oscillator and the sequence in which they are sent. The incoming signal to the sequence signaling receiver at an unattended station is amplified and demodulated. The two low-frequency tones recovered from the 1600-cycle carrier are applied sequentially to the receiving director. The director identifies the tones by means of four accurately tuned reed selectors, recognizes their sequence and translates them into one of ten orders for that particular repeater station.

## VIII. Power Equipment*

The TD-2 System is supplied by battery voltages of $-12$, $+130$ and $+250$ volts maintained by charging rectifiers which float the batteries within limits of $\pm$ 1%. A 24 volt battery to supply power to the C1 alarm and order wire circuits is also included in the power plan where necessary. The block diagram illustrated in Fig. 22 shows the inherent simplicity of the plant. During power failures the batteries carry the load until an automatic gas

* Prepared by J. M. Duguid.

engine alternator or diesel alternator is warmed up and assumes the load, at which time floated operation is resumed. An important characteristic is the absence of any direct switching in the load leads during power failures. The control equipment in all three battery plants is arranged for full automatic operation and additional charging rectifiers are switched in and out as required. After a power failure the rectifiers operate at full capacity until the battery is recharged, after which normal floating operation is resumed. Sufficient capacity is normally installed to give at least an eight-hour reserve

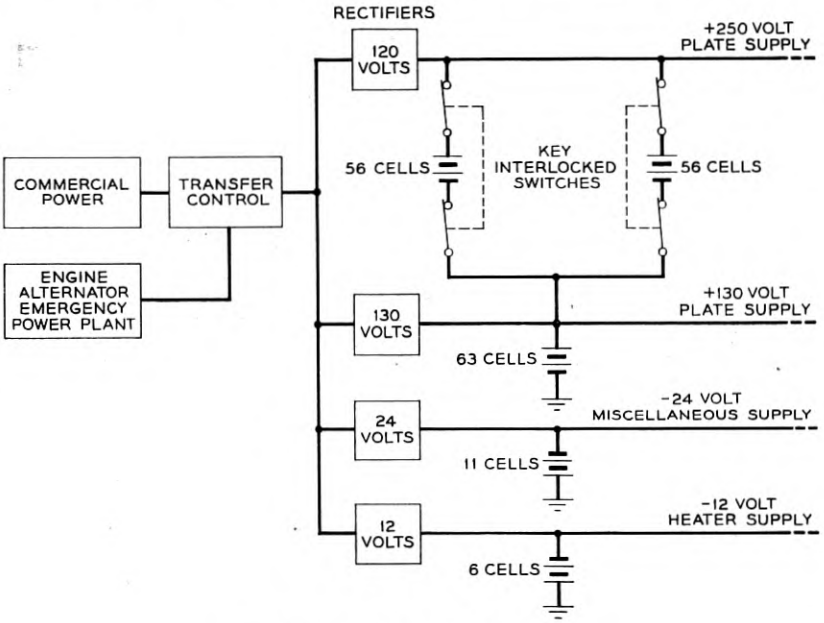



Fig. 22—Power plant block diagram.

to allow an attendant to get to the station in the event that the engine alternator fails to start.

A. *—12 Volt Supply*

The —12 volt heater supply consists of six battery cells floated by two or more parallel-connected 200-ampere full wave selenium rectifiers. The output voltage of the rectifier is controlled by a saturable reactor and regulating autotransformer in series with the primary of the stepdown power transformer which supplies the selenium bridge rectifier. The output voltage is automatically adjusted by the amount of d-c. current supplied to the saturable reactor by the electronic feedback control circuit in the rectifier. The

battery is floated at 13 volts and a discharge resistor in each fused discharge lead is adjusted during installation to drop the voltage to the normal limits of 11.0 ± 0.1 volts at the radio bays. Under 60-cycle a-c. power failure conditions before the gas engine or diesel alternator accepts the a-c. load, the radio bays may operate between their emergency limits of 9.9 to 11.5 volts.

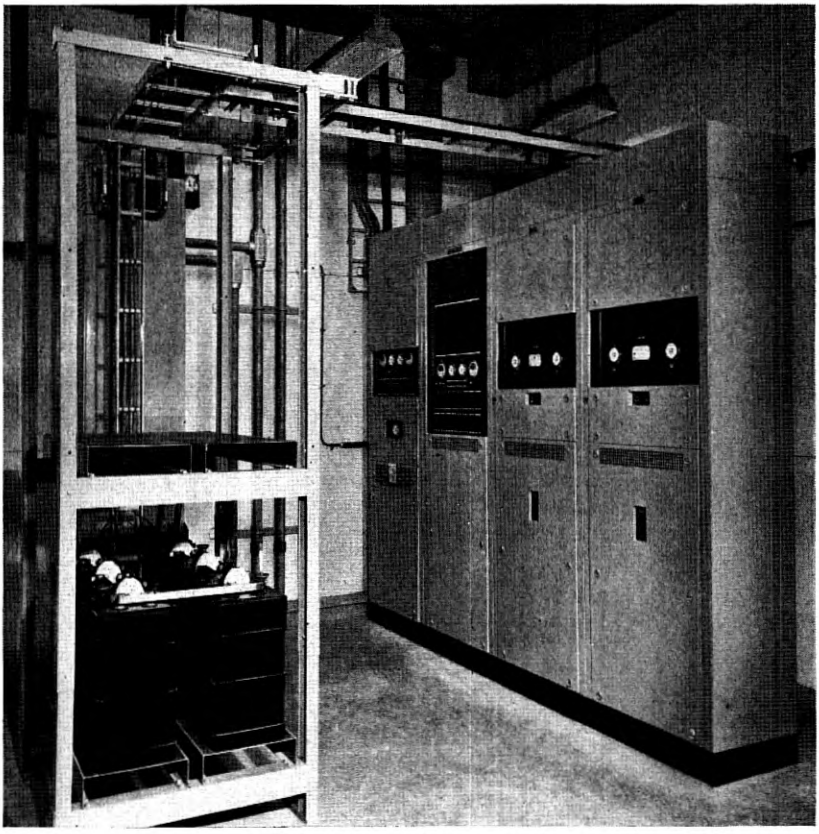


Fig. 23—12V and 24V power plant.

Figure 23 illustrates the installation in a typical tower of the −12 volt supply required for a main route of six radio channels in each direction.

B. *+130 Volt Supply*

The 130-volt plate supply consists of a 63-cell storage battery which is charged and floated by two to eight 8-ampere regulated tube rectifiers. As

shown on Fig. 22, this battery serves as the lower section of the 250-volt plate supply. Its capacity of 20 amperes is sufficient to supply the combined 130 and 250 volt loads. The regulated rectifiers normally float the plate battery at a voltage of 136 ± 1%. Under a-c. power failure conditions emergency limits of 116 to 140 volts are permissible. Due to the relatively high voltage involved and in order to insure maximum service and personnel protection, the rectifiers and their associated control and distribution equipment are mounted in sheet metal enclosures as shown in Fig. 24.

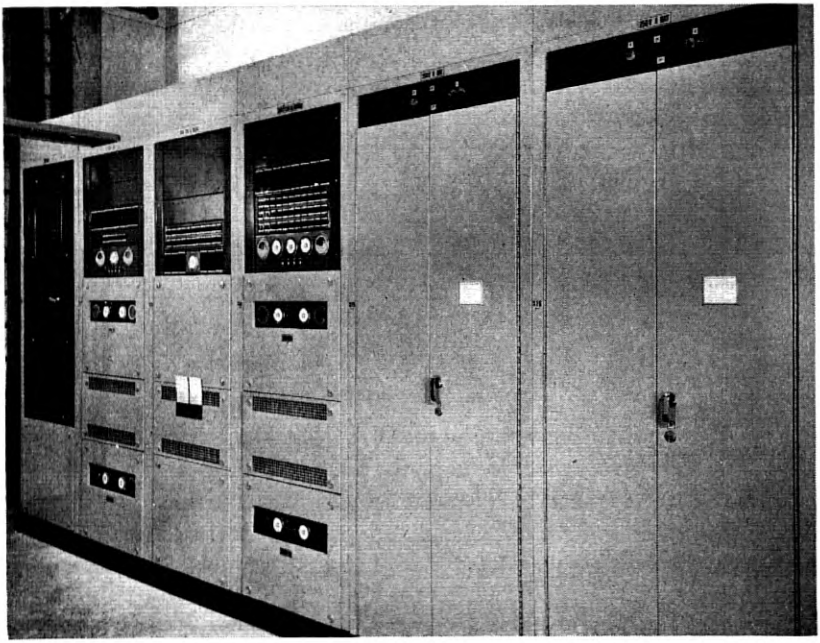


Fig. 24—130V power plant.

C. *+250 Volt Supply*

The 250-volt supply consists of duplicate 56 cell batteries in parallel which are in turn connected in series with the 63-cell 130-volt battery. Regulated thyratron rectifiers similar to those used in the 130-volt plate are connected across the 56 cells to float the load. The normal limits are 254 to 259 volts and the emergency limits are 224 to 266 volts. Each section of the 250-volt battery is housed in its own cabinet and key interlocked to protect maintenance personnel.

D. *−24 Volt Supply*

A −24 volt battery plant utilizing a regulated selenium charging rectifier capable of 6-ampere constant load supplies power for the alarm and order wire circuits. Voltage regulation is obtained by saturable reactors in a magnetic type of regulating circuit. This plant is shown as the extreme left bay in Fig. 23.

E. *Engine Alternator Reserve Plants*

The main route of the TD-2 system normally requires reserve engine alternators of 20 or 30 kw capacity. The initial sets used were of the automatic gasoline engine alternator type available in 20 to 60 kw capacity. The engines are fully automatic in operation. They accept the load after a predetermined period of commercial a-c. service failure and restore the load to the commercial service when it returns to normal. They are capable of long hours of operation under emergency conditions. Numerous alarms are available in the engine plant to indicate its status under all conditions. Recent development has made plants available similar to those mentioned above which are powered by automatic diesel engine driven alternators. It is expected that this latter type of engine will be used in the future where capacities of 20 kw or more are required.

## VIII. Conclusion

The New York-Chicago section of the TD-2 transcontinental radio relay system was opened for service with the transmission of television network programs on September 1, 1950. The system was extended to Omaha on September 30, 1950. Similar systems were put into service during September between New York and Washington and between Los Angeles and San Francisco.

By the fall of 1951 a transcontinental microwave radio relay system will be in service between New York and San Francisco carrying television programs and hundreds of telephone messages. This system will augment present intercity toll facilities and, in conjunction with coaxial cable, will provide a nationwide network of broad-band channels capable of handling television transmission or large groups of telephone circuits.

The growth of broad-band channels during the next few years can be handled by the addition of channels to partially loaded TD-2 Systems and by new routes. Further expansion of radio relay systems into higher frequencies, 6,000 and 10,000 megacycle bands now set aside by FCC for common carrier use, appear to offer room for further expansion of systems comparable to TD-2.

## References

1. H. T. Friis, "Microwave Repeater Research," *B. S. T. J.*, Vol. 27, No. 2, April 1948.
2. G. N. Thayer, A. A. Roetken, R. W. Friis and A. L. Durkee, "The New York-Boston Microwave Radio Relay System," *Proc. I.R.E.*, Vol. 37, pp. 183–188, February 1949.
3. W. E. Kock, "Metallic Delay Lenses," *B.S.T.J.*, Vol. 27, No. 1, January 1948.
4. C. E. Schooley and R. D. Campbell, "Spanning the Continent by Radio Relay," *Bell Telephone Magazine*, Vol. 29, No. 4, Winter 1950–51.
5. W. M. Marsters, "Radio Relay and Other Special Buildings," *Bell Telephone Magazine*, Vol. 29, No. 1, Spring 1950.
6. J. A. Morton and R. M. Ryder, "Design Factors of the B.T.L. 1553 Triode," *B.S.T.J.*, Vol. 29, No. 4, October 1950. (W.E.416A is the production version of the B.T.L. 1553 triode.)
7. A. E. Bowen and W. W. Mumford, "A New Microwave Triode: Its Performance as an Amplifier," *B.S.T.J.*, Vol. 29, No. 4, October 1950.

# Deterioration of Organic Polymers

By B. S. BIGGS

(*Manuscript Received July 9, 1951*)

This paper is a general review of deterioration processes in polymers. It is pointed out that changes in properties with aging are usually the result of chemical reaction with components of the atmosphere. The mechanisms of these reactions and some methods of preventing or retarding them are discussed.

ORGANIC compounds which have enough inherent strength to be used as structural materials—e.g. rubbers, plastics, textiles, and surface coatings—belong to a class called polymers. The deterioration of these materials in service is a serious problem, probably equal in dollar value to corrosion of metals, and one or another aspect of it has been under study in the Laboratories for years.[1] Everyone is familiar with the tendering of cotton cloth and with the loss of strength of rubber with time, but except among people who work with them there is not a wide recognition of the fact that plastics also suffer extensive damage from the weather. This is probably because organic corrosion is usually not visible in its early stages even though deep-seated changes may be taking place throughout the body of the material. In its advanced state, however, such deterioration is easily observable, manifesting itself in loss of strength, erosion, warpage, development of cracks, loss of transparency, or in other ways depending on the material and the application. These changes are of obvious importance in most engineering uses, particularly in the Bell System where apparatus frequently is expected to last thirty or forty years, and it is therefore desirable that they be understood. It is the purpose of this article to review in a rather general way the causes and mechanisms of deterioration.

Even casual consideration reveals that both chemical and physical changes may occur. The loss of plasticizer from a plastic, for example, can induce warping and embrittlement without a change having occurred in the chemical nature of any of the component molecules. Alternate periods of high and low humidity can cause swelling and shrinking in such hydrophilic materials as nylon and cellulose acetate and if stresses are present this can result in permanent distortion[2] (Fig. 1). The swelling of rubber in contact with oils is another example of physical change (Fig. 2). These phenomena are generally well understood and are taken into account in careful engineering. The effect of chemical changes can be even more striking as illustrated in Figs. 3, 4 and 5, but their mechanisms are more obscure and require more detailed discussion.

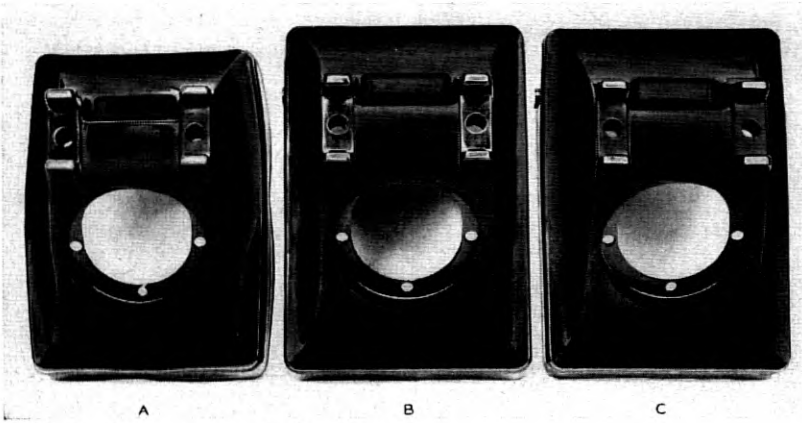



Fig. 1—Cellulose ester telephone housings.
A—Acetate after 7 cycles of high and low humidity.
B—Original.
C—Butyrate after 7 cycles of high and low humidity.

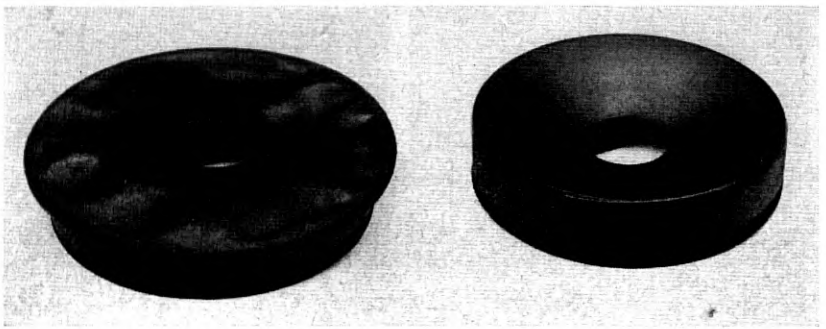


Fig. 2—Neoprene ear pad after one year's use, at left, and original at right.

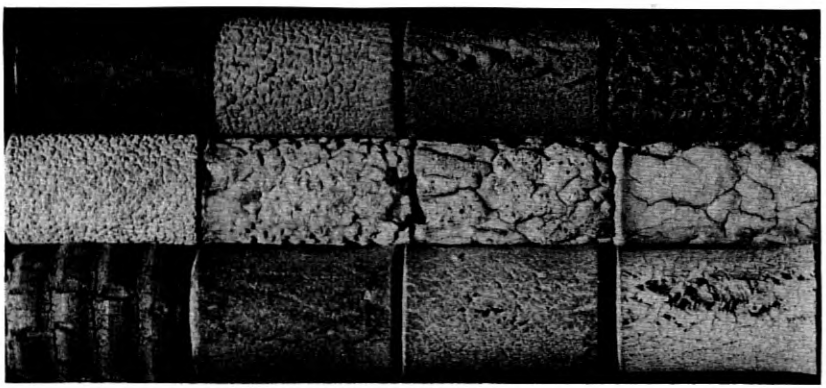


Fig. 3—Samples of rubber in various stages of weathering.

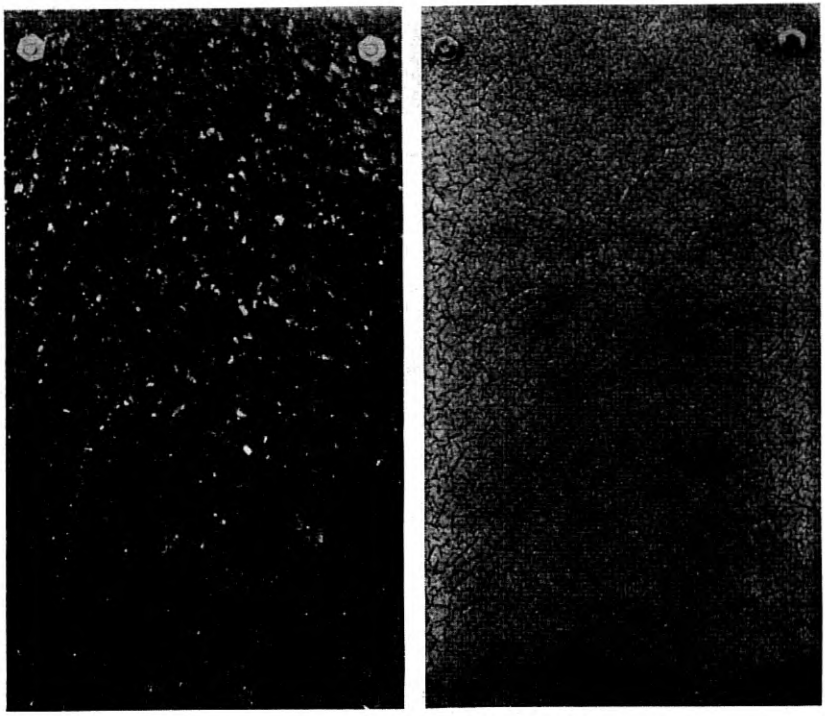
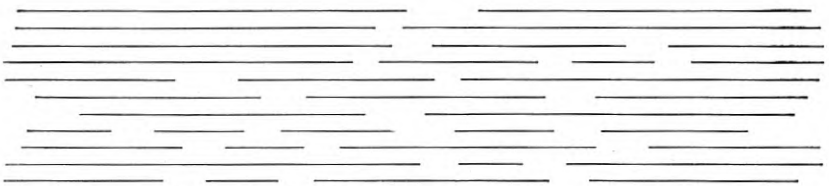


Fig. 4—Cellulose acetate panels exposed in Florida for six months.

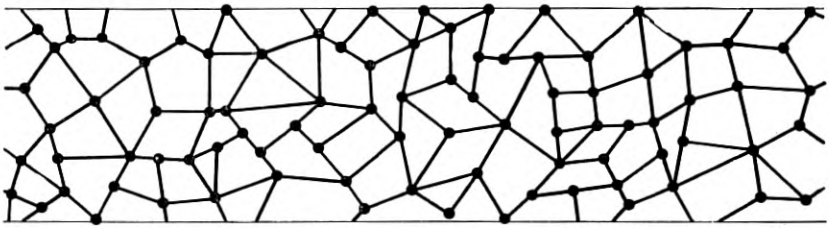
Fig. 5—Samples of rubber garden hose cracked by ozone.

The value of polymers as structural materials is derived entirely from the fact that they are composed of very large molecules. They are generally classified in two broad groups, the essentially linear or chain-like polymers comprising the thermoplastics and rubbers, and the very highly branched three-dimensional networks which are called thermoset materials. The fundamental difference between these groups is shown diagrammatically in Fig. 6 in which, for convenience, the linear polymers are shown as straight lines instead of in their usual randomly kinked shape.

The linear polymers are made up of molecules of finite average size, from a hundred to a thousand or more times as long as they are wide[3] and the



SCHEMATIC REPRESENTATION OF A LINEAR POLYMER



SCHEMATIC REPRESENTATION OF A THERMOSET POLYMER

Fig. 6—Schematic representation of a linear polymer, above, and of a thermoset polymer, below.

strength of the material is dependent on the size of these molecules much as the strength of a cotton thread is dependent on the length of the individual fibers of which it is composed. The forces holding the aggregate together are the cumulative interchain forces. In thermoplastics these forces may be quite strong. In rubbers they are weak until the rubber is vulcanized. Vulcanization connects the chain-like molecules into a loose three-dimensional network, but the number of cross-links is very low compared to typical thermoset polymers being only about one or two for every hundred chain atoms.[4] Vulcanized rubbers are therefore still largely linear polymers and their deterioration follows the pattern of the thermoplastics. The

thermoset materials, of which the phenolic resins are typical examples, are so highly interconnected that the molecular weight can be considered to be infinite. Each molding, for example, may consist of a single molecule. Because of their extensive internal cross-bracing their deterioration is usually a surface phenomenon.[5] It will be discussed later in this memorandum. The paragraphs which follow immediately will refer to linear polymers.

Any material chosen for an engineering application obviously must possess desirable characteristics and "corrosion" or deterioration changes these characteristics in some undesirable way. There are three ways in which a system of chain-like molecules can change: 1) the chains may be cut into smaller pieces, 2) the chains may be tied together by cross-links, and 3) the nature of any side groups along the chain may be modified. All of these changes have been found to occur during normal weathering of polymers and the properties of the product are determined by the extent of each change.[6]

The first type, chain scission, is usually the most serious because it cuts at the very essence of polymeric nature which is high molecular weight. As molecular weight is lowered, strength is lowered and ultimately is lost completely. To continue the analogy to a cotton thread, the individual fibers become so short that they cease to overlap each other adequately. Tough horny polyethylene, for example, deteriorates to something akin to paraffin wax. If chain scission occurs extensively in rubbers, portions of chains are cut loose from the relatively few cross-links and the product will appear to have become unvulcanized. This phenomenon is well known with natural rubber and is called "reversion".[7] (Fig. 7)

The second type of change caused by aging, the introduction of ties or cross-links, is not usually of great importance in plastics unless carried to an extreme when the rigidity and brittleness of thermoset polymers might result. As a matter of fact, the introduction of a few cross-links in a thermoplastic, without accompanying chain scission, probably serves to toughen the material. In rubbers, however, where high elongation is a desired property and is derived from the uncoiling of the molecules under stress, introduction of cross-links beyond those necessary for vulcanization tends to "shorten" the material and can eventually stiffen it to the point that it loses serviceability. The introduction of cross-links increases the density, and frequently when the surface of a plastic or rubber has been cross-linked extensively it develops an "alligator" or "mud crack" pattern resulting from excessive shrinkage.

The third type of change, the modification of side groups, normally has little effect on the strength of a polymer, but may have a pronounced effect on the dielectric properties, solubility, moisture absorption, etc., depending

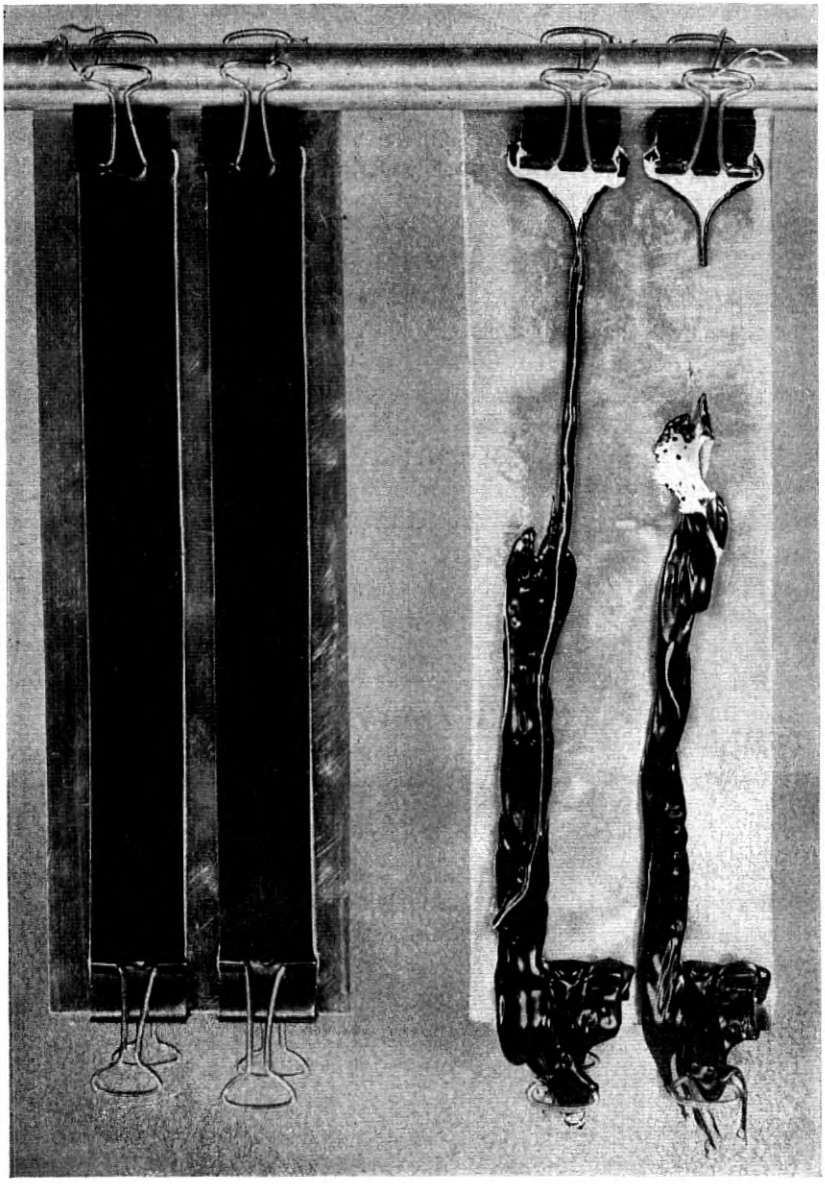


Fig. 7—Natural rubber tapes before and after oxygen bomb treatment.

on the nature of the groups introduced or modified. As indicated above, during normal deterioration all of these types of change are proceeding

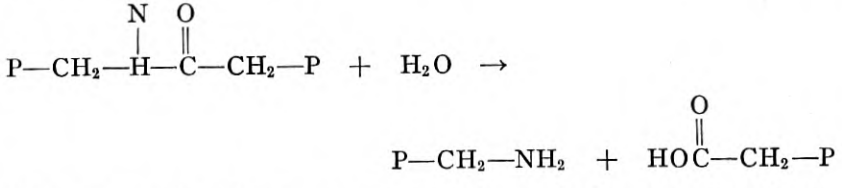
simultaneously to greater or less extent. The rates of the reactions vary from one material to another, and the same conditions which degrade natural rubber to a soft gum may cause neoprene or GR-S to become harder and stiffer.

Returning now to the thermoset polymers, one sees that neither occasional chain scission nor occasional cross-linking can have an important effect on the mechanical properties of a thermoset polymer since every part of the structure is tied to the rest of it by many bonds. For this reason the most conspicuous changes of thermoset materials on exposure to weather are on the surface and are of the third type discussed above.

From the viewpoint of physical structure all of the elements of deterioration are covered in the above paragraphs. However, nothing has been said about the agencies which cause the chemical changes or the mechanisms by which they are brought about. These agencies and mechanisms become the most important objects of study. One type of change—the cross-linking of molecules—in certain cases can occur by self-reaction under the influence of heat or light in complete absence of other chemicals. Self-reaction is not, however, an important effect in materials which are in engineering use. The changes which lead to the loss of utility of polymers during aging are caused by *chemical reaction with the environment.* Usually this environment is the atmosphere. There are normally three substances in the atmosphere which under various circumstances may be considered reactive toward organic compounds, namely water vapor, ozone, and oxygen. The next section will discuss the ways in which these chemicals bring about the destruction or organic polymers.

## Water

The chemical reaction of water with organic compounds is limited to materials which contain hydrolyzable groups either as part of their original composition or as a result of oxidation. Examples of such groups are esters, amides, nitriles, acetals, and certain types of ketones. The reaction is illustrated with an amide linkage, the unaffected portions of the molecule being represented by the letter P:



When these vulnerable groups are present as substituents on a polymer chain composed exclusively of carbon-to-carbon bonds their hydrolysis

may affect certain properties of the material (dielectric constant, power factor, insulation resistance, water absorption) but in general the molecular weight of the polymer is unaffected. When the vulnerable group is a link in the skeletal chain, however, the result of hydrolysis is much more serious because it constitutes scission of the primary chain and hence a lowering of molecular weight. Polymers which are subject to this kind of scission are polyesters, polyamides, cellulose and cellulose derivatives (ethers and esters). Hydrolysis is accelerated by high temperature and is catalyzed by acids and alkalies, and hence many polymers of the classes listed are stable only when kept neutral. Polyesters in particular are usually easily hydrolyzed and it is this fact which has been the main barrier to their greater commercial utilization. Hydrolysis as such is a well known reaction and is taken into account in current engineering with materials which are subject to it. For example, nylon molding powder is shipped dry in sealed containers to keep the moisture content low until after the molding operation which requires that the nylon be heated to a high temperature,[8] and cellulose esters undergo repeated careful neutralizations and washes after esterification to reduce acidity.[9] The extent to which water plays a role in the deterioration of hydrocarbon materials which are first attacked by oxidation is not yet known, but it is certainly secondary to the oxidation itself. An important effect of rain in outdoor weathering is the washing away of water soluble oxidation products with consequent exposure of new surface. Another effect is the removal of water soluble compounding ingredients. This may be distinctly beneficial as in the case of polyester rubbers vulcanized by acid-producing catalysts,[10] or harmful as in certain polyvinyl chloride formulations which contain water soluble protective agents.

## Ozone

Ozone is an extremely reactive chemical which is present in the air in extremely small amounts, ranging from 0 to 10 parts per hundred million. In this low concentration it has not been shown to have any effect on chemically saturated materials, but it is a very serious hazard for unsaturated compounds. Natural rubber and several synthetic rubbers fall in this class (Fig. 8). Ozone is a specific reagent for carbon-to-carbon double bonds, forming an ozonide which undergoes rearrangement resulting in chain scission.[12] When rubber is not being stretched the attack of ozone appears to be negligible, but when it is under stress the attack has very serious consequences resulting in transverse cuts which may sever the piece of rubber.[11, 13] Apparently the initial attack, starting in regions of highest local stress, cuts enough chains to cause a crack to open, and this exposes new surface and concentrates the stress so that the crack grows.

The practical significance of the reaction of ozone on rubber is very great since almost all rubber articles which undergo any appreciable stretching in service are in some degree subject to attack. Exceptions are articles composed of certain specialty rubbers such as silicones, Hypalon*, and some Thiokols. These are saturated materials and hence are not attacked. Neoprene and Butyl rubber are more resistant than natural rubber or GR-S, Butyl because it is only slightly unsaturated, and neoprene because its double bond is considerably deactivated by the adjacent chlorine atom.[14]

$$\begin{matrix} Cl & H \\ | & | \\ -C=C- \end{matrix}$$

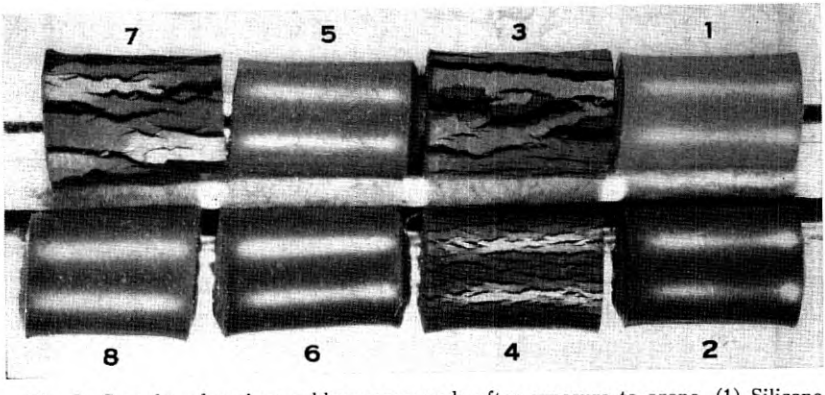



Fig. 8—Samples of various rubber compounds after exposure to ozone. (1) Silicone; (2) Hypalon; (3) Buna-N; (4) natural rubber; (5) & (6) Neoprene; (7) GR-S; (8) Butyl.

Large additions of pigments or plasticizers lower the ozone resistance of neoprene. The measure which has been found most effective for protecting rubber compounds from ozone is the inclusion of several percent of wax. The amount required varies with the type of wax, the polymer, and the other compounding ingredients, the absorptive power of any pigments present being an important factor. By proper compounding neoprene can be made extremely resistant to the attack of ozone, and the other unsaturated rubbers can be greatly improved. The chief effect of temperature changes on the cracking of rubber by ozone is in changing the solubility of wax in the rubber. At elevated temperature the wax film may redissolve and leave the rubber unprotected. This is illustrated in Fig. 9 which shows a tape wrapping which has been attacked on the sunny side, not by the light, but by ozone enabled to reach the rubber because the sun's heat had redissolved the wax in it.

* A chlorinated, sulphonated polyethylene manufactured by the Du Pont Company.

## Oxygen

The degradative agent of most general attack and of greatest economic importance is oxygen, which is capable sooner or later of bringing about change in almost any organic material. Even disregarding the oxidation of dead organic matter in nature, which is aided by bacteria and fungi, one finds many examples of oxidation familiar to the layman. The development of rancidity in foods is a common one. The production of sludge-forming acids in engine oils, and the spontaneous combustion of rags soaked with linseed oil are others. The loss of strength of cotton cloth after a few years of service is very largely due to oxidation although mildew or other fungus attack may have played a part depending on circumstances.[15, 16, 17] That changes

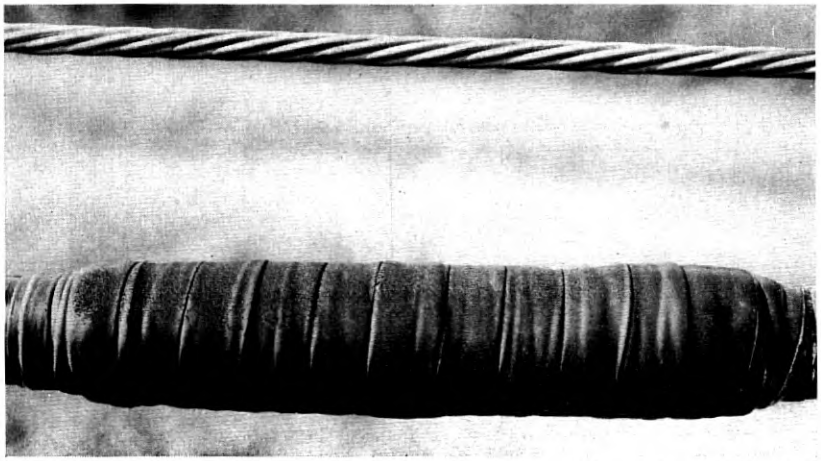


Fig. 9—A tape-wrapped splice after 6-weeks of exposure outdoors. An example of the acceleration of the ozone reaction by heat.

in polymers are indeed the result of oxidation is easily demonstrated in the laboratory by exposing samples to heat or to ultraviolet light in the presence and in the absence of oxygen. The results of such an experiment are shown in Table I, in which solution viscosity is used as a measure of molecular weight. It is seen that in nitrogen neither heat nor light brought about any serious loss of molecular weight.

Similar work has been reported with natural rubber with the conclusion that in an inert atmosphere rubber would retain its original properties "for at least thirty years".[18]

### *Gross Effects of Oxidation of Polymers*

Severe oxidation of organic polymers results in the drastic changes mentioned in the introduction and is easily detected. Photo-oxidation of poly-

ethylene,[19] nylon and cellulose esters,[20] for example, causes crazing, cracking, embrittlement, and in extreme cases granulation of the sample. (Fig. 10) In polyvinyl chloride it leads to hardening and discoloration.[21] In natural rubber, GR-S, and neoprene it causes the development of "mud-crack" patterns or "alligatoring" of the surface and loss of elongation. Thermal oxidation leads to embrittlement of thermoplastics, to "shortening" or loss of elongation in neoprene,[22, 23] nitrile rubbers, and GR-S, and to reversion or the development of tackiness in Butyl rubber and sometimes in natural rubber. As pointed out earlier, these varying effects result from the relative rates of cross-linking and chain-scission reactions. The mechanisms by which oxygen can attack polymers are discussed in the next paragraphs.

TABLE I
SOLUTION VISCOSITY OF CELLULOSE ACETATE BUTYRATE

| | |
|---|---|
| Original | 1.77 |
| After 4 Weeks Exposure to UV Light at Room Temperature | |
| In Nitrogen | 1.60 |
| In Oxygen | .15 |
| After 150 hrs. at 150°C | |
| In Nitrogen | 1.78 |
| In Oxygen | .52 |

*Mechanism of Oxidation Leading to Chain Scission*

The reaction of organic compounds with atmospheric oxygen, frequently called "auto-oxidation" or "autoxidation", has been of interest to chemists for a long time and a voluminous literature on the subject has accumulated.[24, 25, 26] While most of the work done has been on small molecules rather than on polymers it is becoming apparent that much of the mechanism of oxidation is the same and what has been learned on small molecules can be applied to large.[27, 28, 29] This is fortunate since polymers do not lend themselves readily to normal chemical manipulations. While it might be expected that different compounds would be attacked by oxygen in different ways a general mechanism has emerged which appears to be characteristic for aliphatic hydrocarbon structures and is probably applicable to many of the polymeric materials in current engineering use. It can be described as an autocatalytic free radical chain reaction.[30, 31, 32]

The sequence of events is believed to be as follows: Free radicals are produced in the substrate from the energy of heat or of light. They may arise from the decomposition of unstable groupings such as the —O—O—

bond in peroxides or by the dissociation of a relatively more stable bond such as —C—C— or —C—H. Needless to say, the ease with which such cracking occurs is influenced by chemical structure. These free radicals, which may be produced in very minute amount, react with oxygen to form

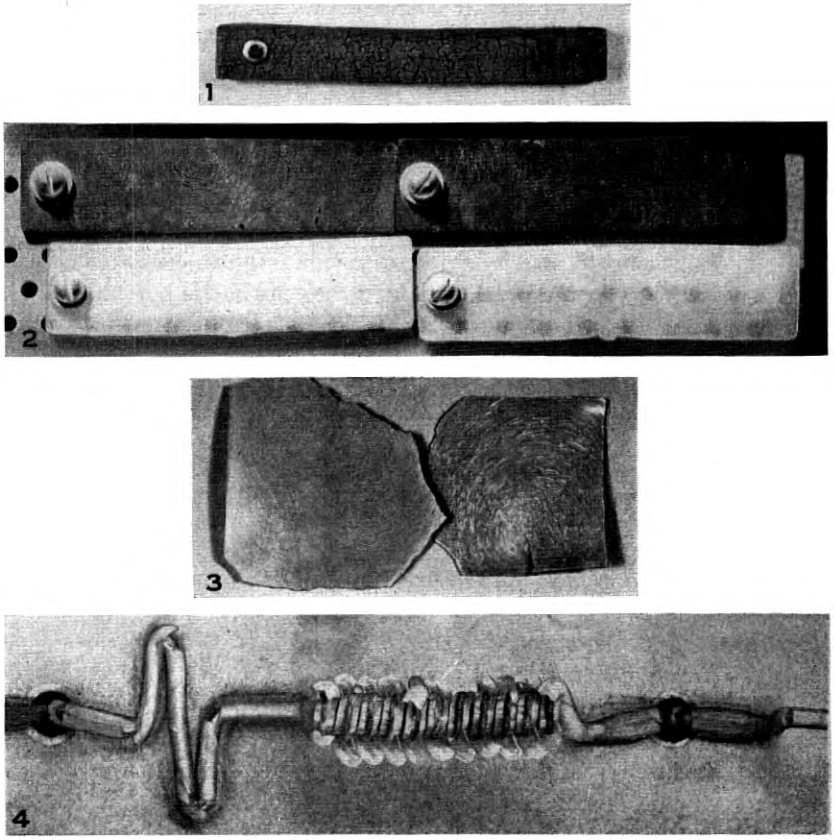



Fig. 10—(1) Cellulose acetate exposed six months at Murray Hill, N. J.
(2) Nylon test panels exposed 5 months at Yuma, Arizona.
(3) Clear Polyethylene sheet exposed 3 years at Murray Hill, N. J.
(4) Clear polyethylene coated wire exposed 3 years at Murray Hill, N. J.

peroxidic radicals. This is illustrated by the following chemical equation in which the radical is represented by the letter R and the fact that it is "free" or reactive is indicated by the dot.

$$R\cdot + O_2 \rightarrow R—O—O\cdot$$

The peroxidic radicals are also reactive entities, but their affinity is for hydrogen atoms and they tend to abstract hydrogen from some other molecule of substrate, thus:

$$ROO\cdot + RH \rightarrow ROOH + R\cdot$$

(These equations were written by Bäckström for the oxidation of benzaldehyde[30] and have been adopted by many others.)[31, 6, 33] The latter reaction results from molecular collision with formation of intermediate additive complexes which decompose into the indicated products and in general many ineffective collisions will occur before reaction takes place. The molecule of substrate which loses hydrogen in this way is now a free radical and it repeats the process, reacting first with oxygen and then with another molecule of substrate. This *linear* chain reaction continues until two radicals unite by collision with each other, thus terminating two chains. The word linear is italicized in the previous sentence to emphasize that this part of the reaction is not in itself autocatalytic. The autocatalytic nature of the oxidation stems from the fact that the product of the reaction as outlined is a hydroperoxide, ROOH. Such compounds are relatively unstable and slowly decompose into free radicals which initiate new chains. This might go as follows:

$$ROOH \rightarrow RO\cdot + \cdot OH$$

$$RO\cdot + RH \rightarrow ROH + R\cdot \text{ and } \cdot OH + RH \rightarrow HOH + R\cdot$$

Thus, though the original rate of generation of free radicals from cracking might have been very low, the combined rate increases quite rapidly since each molecule of peroxide produced in the chain reaction becomes a potential source of new radicals. Eventually the rate reaches what appears to be a steady state and finally levels off. A typical oxygen absorption curve for a liquid hydrocarbon is shown in Fig. 11. The region of fast reaction has received attention from those interested in the oxidation of small molecules but it is unimportant to people interested in polymers because it has been shown by various workers that only slight oxidation is required to destroy the useful properties of a polymer.[34] By the time oxidation has proceeded far enough to be getting into a rapid rate it has already resulted in enough chain scissions to have lowered the molecular weight below useful levels. (A simple calculation will illustrate this point. Suppose a polymer molecule whose molecular weight is 32,000 reacts with one molecule of oxygen (mol. wt. 32) and a chain scission results. The molecular weight of the polymer molecule will have been halved by reaction with .1% of its weight of oxygen. Not every reaction with oxygen results in chain scission of course;[27] but, even so, the amount of oxygen required to ruin the polymer is very small.)

The principal effect of the reactions described above is to introduce the hydroperoxide group into the polymer at various points. It is in the decomposition of these peroxides that chain scission occurs. Studies of the decomposition of the tertiary peroxides produced by oxidation of various dialkyl

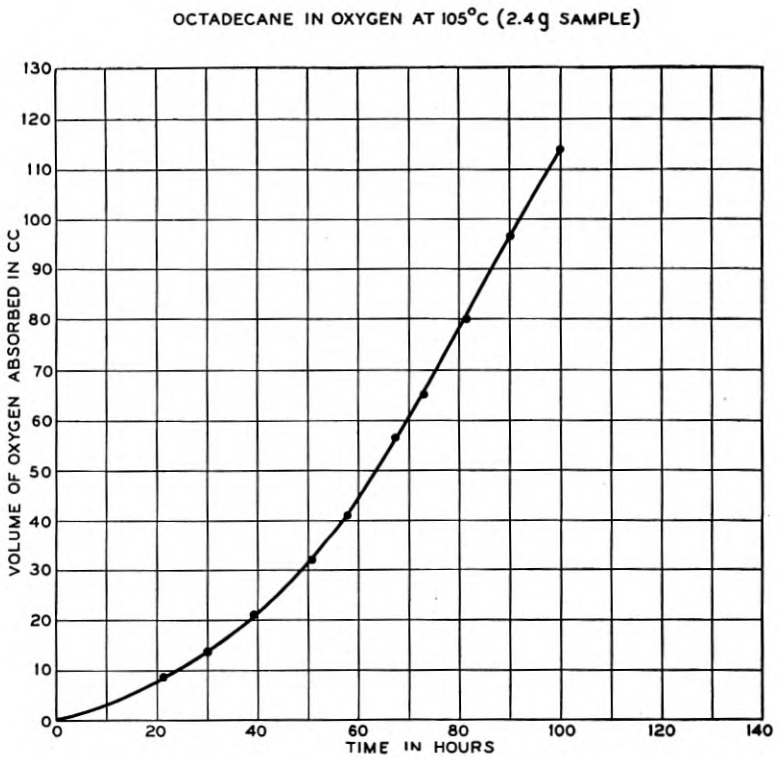
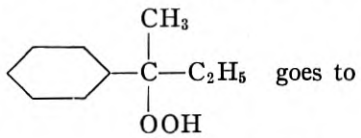
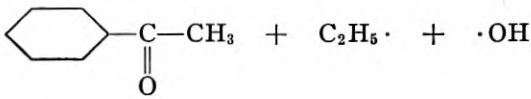


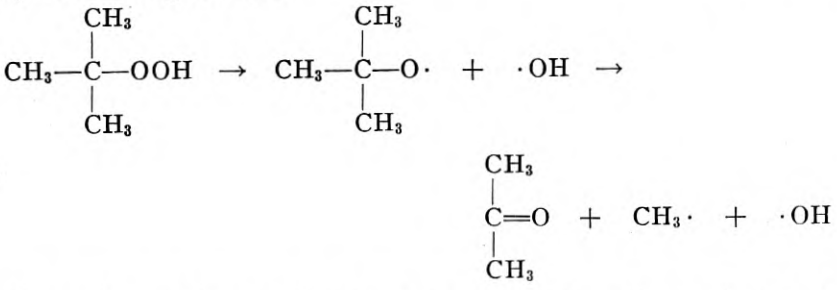

Fig. 11—Octadecane in oxygen at 105° C. (2.4 g sample)

phenyl methanes have shown that the product is invariably an alkyl phenyl ketone in which the alkyl group is the shorter of the two originally present.[35]

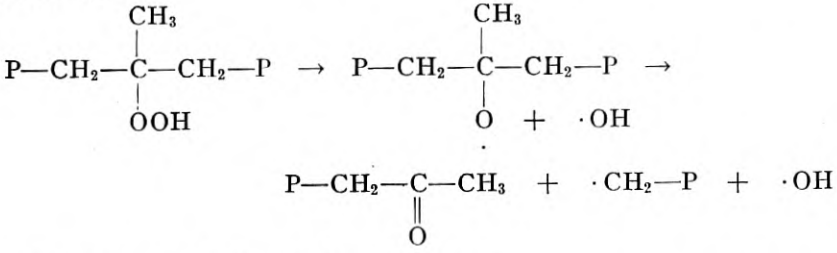
Thus:  goes to



the C—C bond attaching the ethyl group having been broken. The decomposition of tertiary butyl hydroperoxide has been shown similarly to proceed by the following reactions:[36, 37]

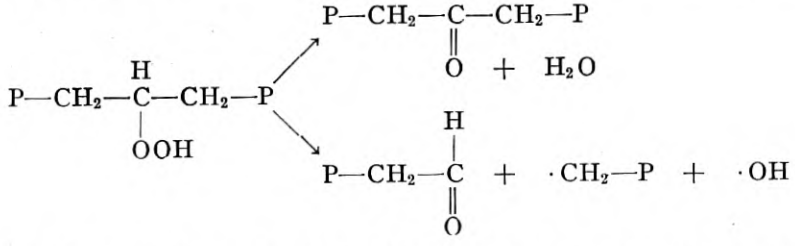


Here again a carbon-to-carbon bond has been broken. The implication of this work for a substituted polyethylene, for example, is quite clear. If oxidation occurred at a tertiary hydrogen as it most probably would,[38, 39, 40] the decomposition of the resulting tertiary peroxide would be as follows (the unaffected portions of the polyethylene chain being represented by the letter P):



Thus the polyethylene chain would be cut.

The course of events is less clear for secondary peroxides. Here at least two paths are possible:



If the decomposition proceeds according to the top arrow, no chain scission results, whereas if it takes the lower course the polymer is divided. The aldehyde is subsequently oxidized to acid. The fact that short fatty acids are produced in the oxidation of straight chain hydrocarbons such as octadecane

is proof that secondary peroxides or peroxidic radicals can decompose by the chain splitting process. That they do not decompose exclusively by that mechanism is shown by the high yield of tetralone obtained from the decomposition of tetralin hydroperoxide.

The mechanisms outlined, while certainly not complete, are adequate to account for the chain scission type of oxidative deterioration of many plastics and rubbers. The degradation of chlorine bearing plastics such as polyvinyl chloride and polyvinylidene chloride, while also being caused by oxygen and being energized by light and heat, is not believed to follow the patterns out-

TABLE II
FIELD RESULTS ON SAMPLES OF NATURALLY AGED NEOPRENE JACKETING (FROM DROP WIRE)*

| Original Months Exposure at | | Tensile Strength psi 2218 | Elongation, % 330 |
|---|---|---|---|
| Chester, N. J. | 15 | 2635 | 215 |
| | 31 | 2655 | 225 |
| | 57 | 2510 | 205 |
| Stone Harbor, N. J. | 21 | 1990 | 190 |
| | 64 | 2485 | 185 |
| | 78 | 2615 | 175 |
| Miami, Fla. | 14 | 2540 | 195 |
| | 48 | 2215 | 140 |
| | 60 | 2260 | 125 |
| | 74 | 2410 | 150 |
| | 87 | 2450 | 130 |
| | 109 | 2520 | 120 |
| San Antonio, Tex. | 11 | 2395 | 160 |
| | 22 | 2300 | 145 |
| | 34 | 2585 | 180 |
| | 45 | 2165 | 135 |
| Brawley, Cal. | 15 | 1980 | 165 |
| | 58 | 2405 | 165 |

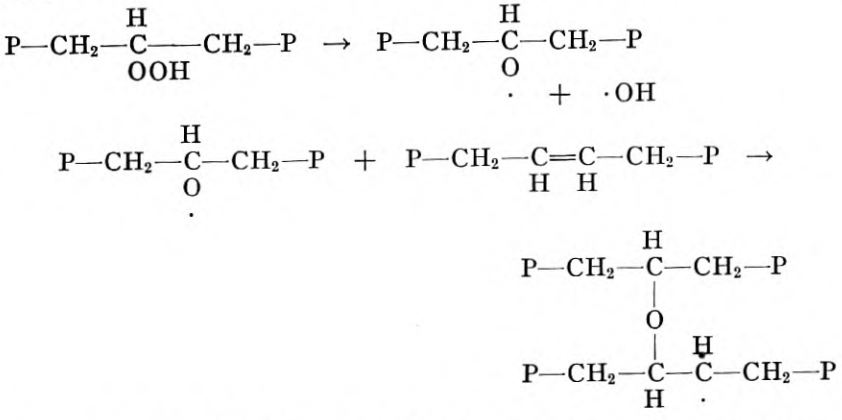
* From a paper by G. N. Vacca, R. H. Erickson and C. V. Lundberg[22]

lined above. The first step here is reported[21,41,42] to be the elimination of hydrogen chloride with introduction of a double bond, which makes the loss of more HCl easier and also increases the oxidizability.

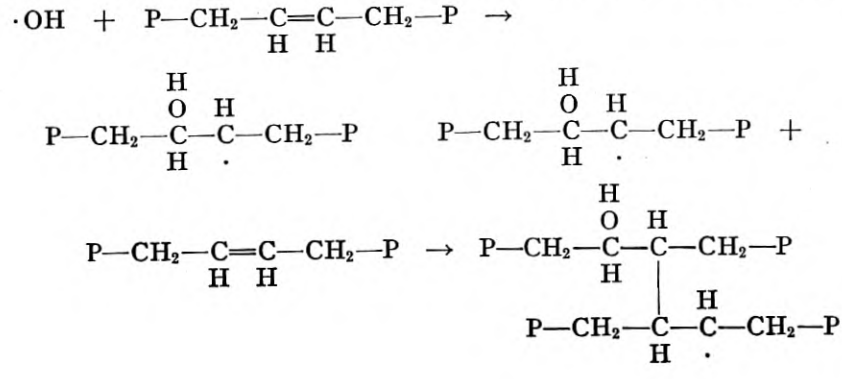
*Cross-Linking Resulting from Oxidation*

The second important effect of oxidation of polymers is cross-linking. This is of great consequence only with unsaturated compounds and these are principally the rubbers. If cross-linking is the dominant reaction (as it usually is on neoprene, GR-S, and the nitrile rubbers) the result is a decrease in elongation and an increase in hardness without a loss in tensile strength.

The strength may actually increase as shown in Table II. The rubber tends to "shorten" and ultimately ceases to be rubbery. Carried to an extreme condition, oxidized rubbers can resemble hard rubber or the phenolic resins. The detailed mechanisms of cross-linking are not worked out but certain deductions can be made about the reaction. That it is not just a polymerization of the double bonds can be shown by the fact that its rate in the absence of oxygen is extremely slow. That it is probably induced by free radicals can be inferred from the work on "vulcanization" of unsaturated polyesters with peroxides in which it was evident that a free radical attacking a double bond initiated the cross-linking reaction.[10, 43] The sequence might be as follows:[28]



Thus a link has been introduced. The new radical can react with another radical, it can react with oxygen to form another peroxide group, it can react with a double bond in another chain to form an additional cross-link, or when antioxidant is present it can react with antioxidant. The hydroxy radical resulting in the original decomposition of the peroxide could initiate a similar series of reactions resulting in one or more cross-links as follows:



The factor that determines whether or not cross-linking will be dominant in the aging of an unsaturated material must be the chemical structure of the polymer (and its peroxide.) The mode of decomposition of the peroxide which, of course, is a function of structure probably has the most important effect. While cross-linking can occur in saturated materials, as shown by the vulcanization of saturated polyester rubbers with peroxides, its rate is never high enough to result in a condition that could be called deterioration. Both polyethylene and cellulose acetate butyrate can undergo enough gelation on outdoor exposure to become insoluble, but if this were the only change occurring their toughness would be improved rather than degraded by it. Their deterioration in strength is due entirely to chain scission.

*Modification of Side Groups by Oxidation*

All the oxidation reactions discussed result in the introduction of oxygen into the polymer composition. If the polymer is one which already contains a high percentage of oxygen such as cellulose or even nylon, this may have little effect. If the polymer is a hydrocarbon, however, its power factor will be raised markedly. As a matter of fact the measurement of power factor is a very sensitive way of detecting the addition of oxygen to polyethylene. Except where the polymer is being used for its low power factor, however, the change in side groups will be secondary to the change resulting from chain scission and cross-linking.

*Acceleration of Oxidation*

The foregoing description of the mechanisms of auto-oxidation makes apparent several ways in which oxidation may be accelerated beyond what might be called the natural rate for a pure material. Since oxidation is a free radical process an obvious way to accelerate it is to add free radicals or materials which produce free radicals. Addition of peroxides to organic compounds generally accelerates the rate of oxidation.[33] Similarly the oxidation of a relatively stable material is accelerated if there is left in it a small amount of a chemical which itself is easily oxidized to peroxides. For example, an addition of turpentine greatly accelerates the air-oxidation of paraffin wax.[44] The addition to polyethylene of an unsaturated polymer such as natural rubber would probably have a similar effect.

It is apparent that the amount by which the rate of oxidation of a substrate is accelerated by peroxides, whether the latter are added as such or are self-generated, is dependent on the rate of decomposition of the peroxide. The latter rate can be accelerated by the presence of certain metallic ions and hence they act as catalysts for oxidation reactions. Copper is particularly active in this regard in natural rubber, and the rubber industry long ago learned to avoid it[45] (Fig. 12). Other metals which have been found to

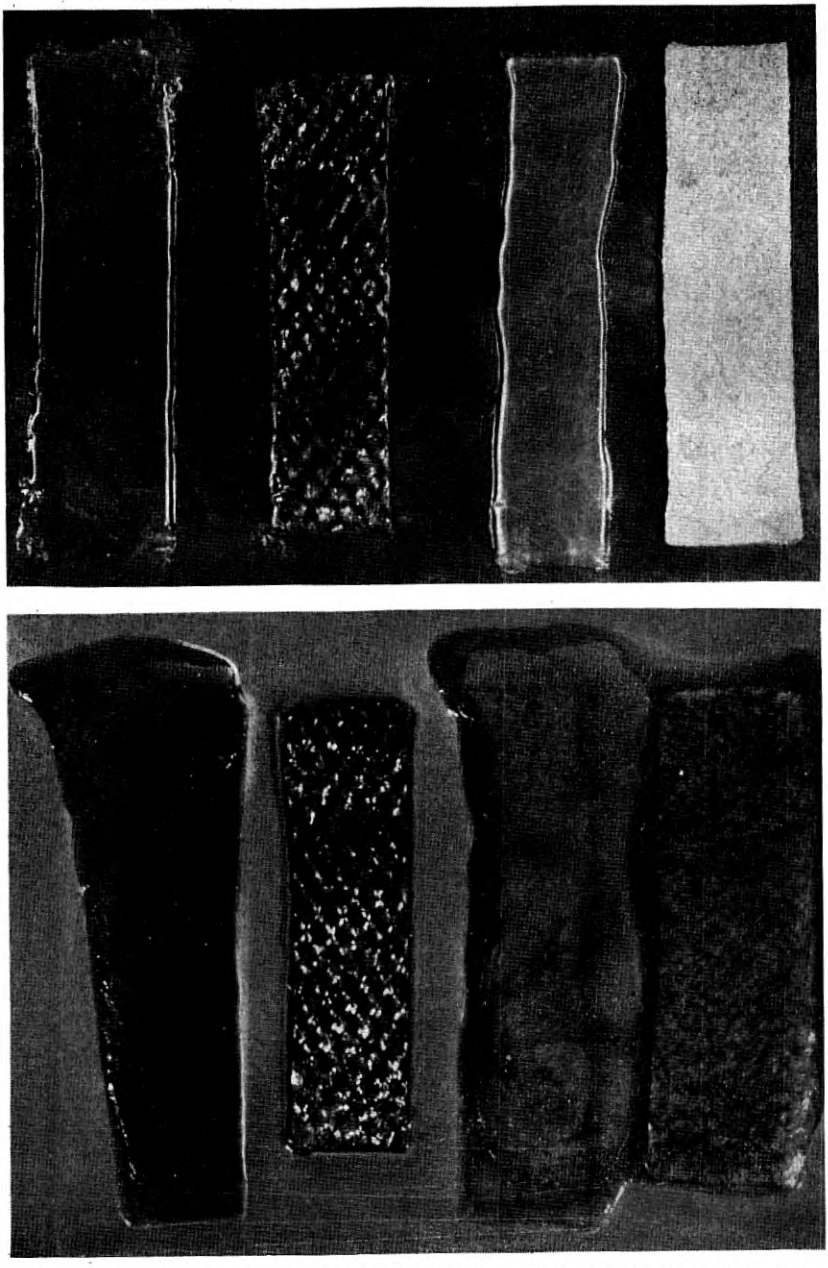


Fig. 12—Various samples of natural rubber oven aged on tin, above; and on copper, below.

cause poor aging at various times are cobalt, manganese, and iron.[46] Since the "drying" of paint is an oxidative reaction, and since rapid drying is a desirable feature, the paint industry has found it advantageous, to use certain metallic salts as "paint dryers".[47]

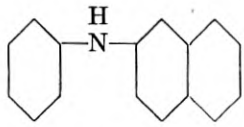
*Retardation of Oxidation*

*Antioxidants*

It was discovered by Moureu and Dufraisse about 1918[48] that the oxidation of many organic compounds could be very greatly retarded by the addition of small quantities of certain other chemicals, which they called "antioxygens." Although the mechanisms which they postulated for the action of these materials were later found to be incorrect, their discovery led to the wide use of such protective agents in industry, particularly in rubber which needs this protection badly. The action of what are now called "antioxidants" becomes clear when one understands the free radical chain mechanism of oxidation outlined above. Antioxidants are chain stoppers.[4] By interposing themselves in the chain reaction they terminate it by giving rise to relatively inert free radicals[50, 51] (stabilized by resonance). For example, the antioxidant, designated HA, could act in the following way:

$$ROO\cdot + HA \rightarrow ROOH + \cdot A$$

In this case the antioxidant satisfies the peroxidic radical by giving it the hydrogen atom it needs, but the residual radical $\cdot A$ is not sufficiently reactive toward oxygen to continue the chain. A typical antioxidant is $\beta$-phenyl naphthylamine,




It was pointed out earlier that many ineffective collisions of the radical $ROO\cdot$ with substrate molecules occur before reaction takes place. If the reactivity of $ROO\cdot$ toward HA is sufficient that few ineffective collisions take place, then small concentrations of HA in the substrate will be adequate to stop each chain at a very early stage. This not only saves all those substrate molecules which would otherwise have become links in these chains but, by so doing, it limits the number of molecules of peroxide produced and thus keeps the rate of initiation of new chains at a low level. The degree of protection by antioxidants varies with the length of the "natural" chain reaction (which is a function of the ratio of effective to

ineffective collisions in the absence of an inhibitor and depends on chemical structure) and on the efficiency of the antioxidant but, in some cases, particularly with liquids, very remarkable protection is obtainable as shown in Fig. 13. (The oxidation of the control sample is so fast at this high temperature that the autocatalytic period is not evident.) The effect is less in solids but is still of great value. Antioxidants are of the greatest benefit where the rate of initiation is low, a condition usually true of thermal oxidation. The

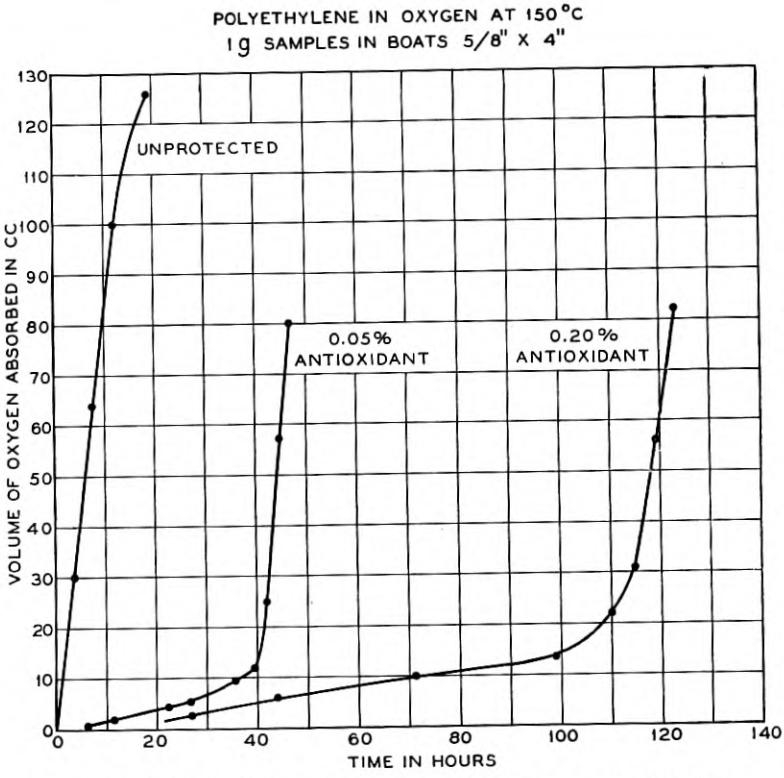



Fig. 13—Polethylene in oxygen at 150° C, 1 g samples in boats $\frac{5}{8}'' \times 4''$.

reason is that since antioxidant is *consumed*[52, 27] in doing its job the rather limited amounts which can be added from a practical point of view (usually not over 1 or 2%) do not last long if the rate of chain formation is very high. This explains the oft stated fact that an antioxidant is far more effective if added before oxidation starts, than if added after oxidation has proceeded for a while.[53] In the latter case enough peroxide will have been produced to overwhelm the antioxidant relatively quickly. This is also the explanation of

the fact that antioxidants are limited in their effectiveness against photooxidation. The rate of initiation of chains in a material exposed to sunlight is so great that any antioxidant present is used up relatively fast. Furthermore, the oxidation of the antioxidant itself is rapid in sunlight and hence if it were not removed in the one way it would be in the other.

There are many chemicals in current use as antioxidants and more are being created all the time. Most of them are either phenols or aromatic amines. It is frequently asked why one antioxidant is more effective than another, if indeed such differences do exist. The answer is not altogether clear but certain statements can be made about it. In the first place, for any given substrate there are usually several antioxidants which are equally good. However, gradations of effectiveness of many commercial antioxidants can be demonstrated. Many factors can exert an influence on this. Some are solubility in the substrate, volatility, inertness toward the substrate. Beyond these are the reactivity of the antioxidant toward free radicals, both hydrocarbon and peroxidic, and the relative stability of the free radical left when the antioxidant reacts. Undoubtedly, some intermediate level of reactivity is desirable in an antioxidant[54, 55] and this desired level probably varies from one substrate to another.

*Light Screens*

It was mentioned above that antioxidants are of little effect against relatively strong photooxidation because of the overwhelming rate of generation of chains. The most serious problems of deterioration in the Bell System are, of course, in outdoor applications, and it is quite clear that this is because of exposure to short wave light. The extensive commerical use of unprotected material outdoors came about because of a lack of appreciation of this fact. Once this vulnerability or organic materials to light is appreciated the remedy is obvious, at least in principle, and that is to shut off the light. For this purpose there are many pigments available as well as many light-absorbing organic compounds. A great deal of work has been done with various substrates in testing the effects of the absorbers, and this can be summarized as follows:

In the class of light colored pigments, none offers complete protection. Most of them have a slight effect; a few are fairly helpful; and a few are actually harmful, acting as photosensitizers. Of the darker pigments several are quite effective but the outstanding ones are lead chromates, iron oxides and carbon black, the last being the best. A study of the effect of various types of carbon black in various concentrations in polyethylene has been reported[19] wherein it is shown that under the most favorable conditions the useful life of polyethylene, as judged by accelerated tests, can be extended

at least 30 fold. It was shown in this work that for best results the carbon black should be finely divided and well dispersed. The use of polyethylene as a sheath material on outdoor cable would not have been practical without the protective effect of carbon black. The efficacy of carbon black as a light screen is apparently quite general although detailed studies have been made only with polyethylene, rubber,[56] cellulose esters,[57] and polyvinyl chloride,[58] in all of which it is effective.

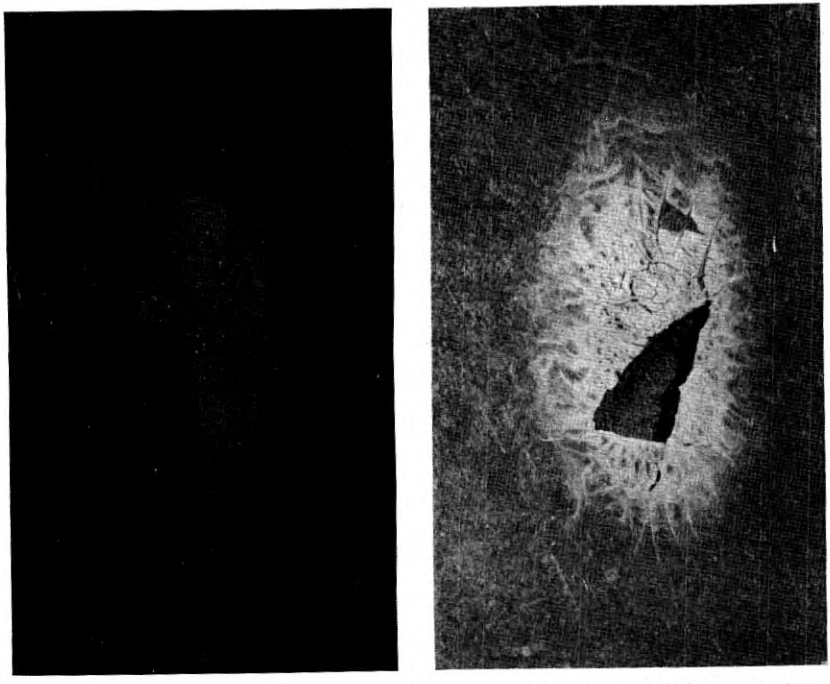


Fig. 14—Cellulose Acetate Butyrate panels exposed to concentrated beams of UV light filtered through pyrex bottles filled with water. Sample at left contains 1% carbon black. Sample at right contains 1% Salol.

In many applications of plastics and rubbers, colors are desirable and for these the use of carbon black is, of course, precluded. Some success has resulted from the use of organic materials which are transparent to visible light but which absorb in the ultraviolet. For example, phenyl salicylate, known as Salol, at a concentration of 1% is a fairly effective protective agent for transparent cellulose esters.[20] Such effects appear to be quite specific, since Salol is not nearly so effective in most other polymers (it is reported to be effective in Saran[59]), and many other compounds which are even better

absorbers of ultraviolet light are much less effective in cellulose esters. Some of them actually are sensitizers. Even in cellulose esters Salol at a concentration of 1% is a poor second to carbon black, giving in accelerated tests less than half the life imparted by 1% of a well dispersed, finely divided carbon black.[57] (Fig. 14)

## References


1. G. T. Kohman, *J. Phys. Chem. 33*, 226 (1929).
2. R. Burns, *A. S. T. M. Bulletin* #134, pg. 27 (May 1950).
3. T. Alfrey, "Mechanical Behavior of High Polymers," pp. 464–465, Interscience Publishers, (1948).
4. P. J. Flory, *Chem. Reviews, 35*, 51 (1944).
5. L. H. Campbell, A. H. Falk, R. Burns, *Proc. A. S. T. M. 46*, 1465 (1946).
6. R. B. Mesrobian and A. V. Tobolsky, *Jl. Polymer Science, 2*, 463 (1947).
7. C. C. Davis and J. T. Blake, "Chemistry & Technology of Rubber," pp. 538, Reinhold Publishing Co., N. Y. (1937).
8. Du Pont *Technical Service Bulletin* No. 8B, March 1950.
9. C. J. Malm and C. L. Crane, *U. S. Patent* #2,346,498.
10. B. S. Biggs, R. H. Erickson and C. S. Fuller, *Ind. and Eng. Chem., 39*, 1096 (1947).
11. J. Crabtree and A. R. Kemp, *Ind. and Eng. Chem. 38*, 278 (1946).
12. A. Rieche, R. Meister, H. Santhoff, H. Pfeiffer, *Liebig's Ann. Chem., 553*, 187 (1942).
13. R. G. Newton, *Jl. of Rubber Research, 14*, 27 (1945).
14. C. R. Noller, J. F. Carson, H. Martin, K. S. Hawkins, *Jl. Am. Chem. Soc. 58*, 24 (1936).
15. J. D. Dean et al, *Am. Dyestuff Reporter 36*, 705 (1947).
16. J. D. Dean and R. K. Worner, *Am. Dyestuff Reporter 36*, 405 (1947).
17. G. S. Egerton, *Am. Dyestuff Reporter 36*, 561 (1947).
18. Admiralty Engineering Lab., *Journal of Rubber Research 15*, 737 (1946).
19. V. T. Wallder, W. J. Clarke, J. B. DeCoste and J. B. Howard, *Ind. and Eng. Chem. 42*, 2320 (1950).
20. L. W. A. Meyer and W. M. Gearhart, *Ind. and Eng. Chem. 37*, 232 (1945).
21. V. W. Fox, J. G. Hendricks, H. F. Ratti, *Ind. and Eng. Chem. 41*, 1774 (1949).
22. G. N. Vacca, R. H. Erickson, and C. V. Lundberg, *Ind. and Eng. Chem. 43*, 443 (1951).
23. D. C. Thompson and N. L. Cotton, *Ind. and Eng. Chem. 42*, 892 (1950).
24. Symposium on Oxidation, *Trans. Faraday Soc. 42* (1946).
25. K. C. Bailey, Retardation of Chemical Reactions, Longmans, N. Y. (1937).
26. H. H. Zuidema, *Chem. Reviews 38*, 197 (1946).
27. L. Bateman, *Trans. of Inst. of Rubber Ind. 26*, 246 (1950).
28. A. V. Tobolsky, *India Rubber World 118*, 363 (1948).
29. J. L. Bolland and P. TenHave, *Trans Faraday Soc. 45*, 93 (1949).
30. H. L. J. Bäckström, *Zeit. für Physichalische Chem. B25*, 99 (1934).
31. L. Bateman and G. Gee, *Proc. Royal Soc. 195*, 376 (1949).
32. E. H. Farmer, G. F. Bloomfield, A. Sundralingham, and D. A. Sutton, *Trans. Faraday Soc. 38*, 348 (1942).
33. J. L. Boland, *Proc. Royal Soc. 186*, 218 (1946).
34. R. Houwink, *Kautschuk 17*, 67 (1941).
35. H. N. Stephens, *Jl. Am. Chem. Soc. 50*, 2523 (1928); *57*, 2380 (1935).
36. J. H. Raley, F. F. Rust and W. E. Vaughn, *Jl. Am. Chem. Soc. 70*, 1336 (1948).
37. N. A. Milas and D. M. Surgenor, *Jl. Am. Chem. Soc. 68*, 205 (1946).
38. H. S. Taylor and J. O. Smith, *Jl. Chem. Physics 8*, 543 (1940).
39. A. D. Walsh, *Trans. Faraday Soc. 42*, 269 (1946).
40. P. George and A. D. Walsh, *Trans. Faraday Soc. 42*, 272 (1946).
41. R. F. Boyer, *Jl. Phys. and Colloid Chem. 51*, 80 (1947).
42. P. I. Pavlovich, *Legkaya Prom.* (1945). 23 C.A. *40*, 7699 (1946).
43. W. O. Baker, *Jl. Am. Chem. Soc. 69*, 1125 (1947).
44. F. E. Francis, *Jl. Chem. Soc. 121*, 502 (1922).
45. C. O. Weber, "The Chemistry of India Rubber," pp. 220 and 299, Charles Griffin and Co., London, 1902.

46. A. Van Rosse and P. Dekker, *Ind. and Eng. Chem. 18*, 1152 (1926).
47. *Proc. of the Scientific Sec. Nat. Paint, Varnish and Lacquer Assoc.*, Circ. 546, pp. 307 (1938).
48. C. Moureu and C. Dufraisse, *Chem. Reviews 3*, 113 and ref. cited therein (1926).
49. J. A. Christianson, *Jl. Phys. Chem. 28*, 145 (1924).
50. J. L. Bolland and P. TenHave, *Trans. Faraday Soc. 43*, 201 (1947).
51. H. S. Taylor, *A. S. T. M. Proc. 32 Part II*, 9 (1932).
52. H. N. Alyea and H. L. J. Bäckström, *Jl. Am. Chem. Soc. 51*, 90 (1929).
53. A. M. Wagner and J. C. Brier, *Ind. and Eng. Chem. 23*, 46 (1931).
54. L. F. Fieser, *Jl. Am. Chem. Soc. 52*, 5204 (1930).
55. C. D. Lowry, C. G. Dryer, G. Egloff, and J. C. Morrell, *Ind. and Eng. Chem. 24* 1375 (1932).
56. W. N. Lister, *Trans. Inst. of Rubber Ind. 8*, 241 (1932).
57. R. H. Erickson, unpublished work.
58. V. T. Wallder and J. B. DeCoste, unpublished work.
59. R. F. Boyer, *U. S. Patent* 2,429,155.

# The Development of Electron Tubes for a New Coaxial Transmission System

By G. T. FORD and E. J. WALSH

(*Manuscript Received July 27, 1951*)

## 1. Introduction

AS THE demand for long distance telephone circuits has increased, new transmission systems capable of handling more channels per conductor have been developed. Also the advent of television has created a demand for broad band channels for network facilities. One of the latest developments now nearing completion is the L3 Coaxial System.

Three new tubes have been developed specifically to meet the exacting requirements of this system: two tetrodes, the W.E. 435A and W.E. 436A, and a triode, the W.E. 437A. All three types are used in the line and office amplifiers. The new tubes make possible a substantially higher level of broad band amplifier performance compared to their predecessors. They represent the result of improvements made by applying well known basic principles through new tube-making techniques. These techniques have been developed largely within the framework of existing conventional telephone tube manufacturing methods.

The development of special, small, low power vacuum tubes for high frequency application in the Bell System began in 1934. The tube program was instituted originally as part of a research project in the field of radio communications. When the development of the L1 Coaxial System began it was recognized that similar tubes would be needed. Part of the tube development effort was therefore directed toward the coaxial requirements. Work on the W.E. 384A and W.E. 386A tubes used in the L1 system was completed in 1939 as an outgrowth of this program.

The demand for amplification over wider frequency bands resulted in further development work along the same lines. During World War II this effort was applied to the development of the 6AK5 tube which became available early in 1943 and was used widely in IF amplifiers in radar equipment. Shortly after the war the W.E. 408A tube was developed for telephone repeater uses. This is a long life version of the 6AK5 tube having the same electrical characteristics except for the heater voltage and current. The W.E. 404A tube appeared in telephone circuits in 1949. This tube, having a higher figure of merit than the W.E. 408A, provided improved performance in the IF amplifiers used in the New York to Boston radio

relay system and in the TD2 radio relay system. The W.E. 435A, W.E. 436A and W.E. 437A tubes are the latest types to come out of this long range program.

It will be seen in what follows that the key to continued development along these lines has been improvements in the techniques of grid making to meet the basic objective of providing a grid which can be spaced very close to the cathode and which, in effect, acts as a uniform potential plane controlling the current drawn from the cathode without offering any physical obstruction. This objective is approached by using many turns of very small diameter wire for the grid winding. The reason for the close grid-cathode spacing is that the transconductance or sensitivity depends on this factor. Although the increase in input capacitance which results is a disadvantage because of its effects on the interstage circuits, this disadvantage is more than compensated for by the higher transconductance obtained.

## 2. Principles of Design

### 2.1 *Requirements*

The overall requirements for the L3 system, and the manner in which they are related to the tube parameters, are very complex. However, in its simplest terms, the objective for the L3 system is to provide on one coaxial pipe a facility suitable for the simultaneous transmission over a 4000-mile circuit of a television signal and 600 one-way telephone channels or, alternatively, 1800 one-way telephone channels when no television channel is required. The transmission band being provided is from approximately 0.3 MC to approximately 8 MC. The amplifier needed to compensate for the cable attenuation must meet very exacting requirements with respect to gain-frequency characteristics, stability, noise, and linearity.

The design features necessary to provide suitable electron tubes for use in the L3 amplifiers are closely related to the requirements mentioned above for the amplifiers. In general terms, the tube design objectives are: (1) high transconductance-capacitance ratio (figure of merit), (2) minimum excess phase shift or phase delay, (3) low noise, (4) well controlled modulation, (5) long life, (6) interchangeability, and (7) lowest cost consistent with the first six objectives. In the material which follows, each of these objectives will be discussed in detail and its relationship to the system objectives brought out.

### 2.2 *Figure of Merit*

Figure of merit is of particular importance. It is a direct measure of the bandwidth over which the required amplification can be obtained. In gen-

eral, a given factor of improvement in the figure of merit can be translated directly into a wider transmission band providing more communication channels.

For a two-terminal type of interstage such as that used in the L3 amplifier, the figure of merit is

$$F = GB = \frac{KG_m}{2\pi(C_1 + C_2)} \tag{1}$$

where $F$ is the figure of merit, $G$ is the voltage amplification, $B$ is the bandwidth between the frequencies where the gain is 3 db below that at the center frequency, $K$ is a constant whose value depends on the particular interstage design, $G_m$ is the transconductance of the tube, $C_1$ is the input

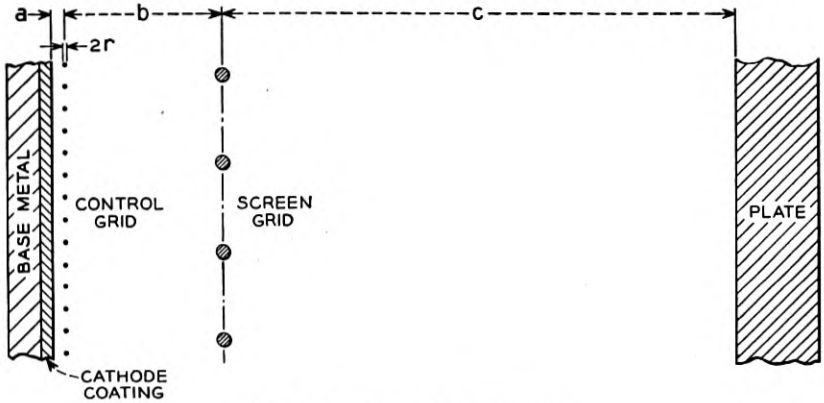



Fig. 1—Geometry of the 436A tube.

capacitance, and $C_2$ is the output capacitance. This figure of merit is directly applicable to a tetrode operated as a small-signal voltage amplifier and is a well known relationship.[1] It will be used to show how the tube design factors influence the figure of merit of the W.E. 435A and W.E. 436A tubes.

Using equation (1) and applying certain simplifying assumptions which can be made without materially affecting the results, expressions are derived in the appendix showing the relationship between the figure of merit and the tube parameters. Equations (2) and (4) in the appendix show how the figure of merit is affected by the grid-cathode spacing "$a$", the grid-screen spacing "$b$", the screen-plate spacing "$c$", and the grid wire radius "$r$". See Fig. 1. Equation 3 gives the required screen voltage for the as-

[1] "Characteristics of Vacuum Tubes for Radar Intermediate Frequency Amplifiers," G. T. Ford, *B.S.T.J.*, Vol. XXV, p. 389, July, 1946.

sumed current density and geometry. Since these expressions are rather involved, the manner in which the various factors influence the figure of merit can be brought out best by a series of curves. Figures 2, 3, and 4 show how $F$ is affected by changes in "$a$", "$b$", and "$c$". Figures 6 and 7 show how the screen voltage required to get the assumed current density with a given bias $E_{c1}$ varies with "$a$" and "$b$" (equation 3). The screen voltage is essentially independent of "$c$".

These relationships are also applicable to the W.E. 437A tube with minor modifications.

### 2.21 *Design Considerations*

How the various factors in equations (2), (3), and (4) affect the figure of merit will be discussed in detail. They are listed in Table I.

The factor $M$ is the ratio of the plate current to the cathode current. The figure of merit is directly proportional to this factor. $M$ can be increased

TABLE I

| Factor | Design Values | Practical Design Considerations |
|---|---|---|
| $M$ | 0.75 | Mechanical, plate-grid capacitance |
| $I_0$ | 50 MA/cm$^2$ | Stability of emission, life |
| $a$ | 0.00635 cm | Mechanical |
| $b$ | 0.0444 cm | Screen voltage, mechanical |
| $c$ | 0.150 cm | Formation of potential min. |
| $r$ | 0.00038 cm | Mechanical |
| $E_{c1}$ | $-1.5$ volts | Grid current |
| $E_{c2}$ | 150 volts | Dissipation, life |

by using smaller wire in the screen grid, the minimum practical wire size being determined by the mechanical rigidity and heat dissipation capability required. $M$ can also be increased by reducing the number of turns on the screen, but this is limited by the necessity for sufficient shielding effect to meet the requirement that the plate-grid capacitance be less than a specified value.

Since the figure of merit is directly proportional to the cube root of the cathode current density $I_0$, the improvement with increasing $I_0$ is not very rapid. The problems of obtaining uniform initial performance and long life are aggravated by increasing the current density, for several reasons. There is no direct evidence to show that high current density per se causes accelerated loss of available emission. In fact there is some evidence to the contrary.[2] However, there is ample evidence that phenomena

[2] "Influence of Density of Emission on the Life of Oxide Cathodes," S. Wagener, *Nature*, p. 357, Aug. 27, 1949.

usually associated with high current density tend to shorten the life. Higher electrode temperatures, higher potentials, and the production of more ions are the major items in this category. It is presumed that the shorter life found under these conditions is due to the greater rate of contamination of the cathode by material from the other parts of the tube. Great efforts have been made to find and to use processing techniques which will minimize this kind of limitation and to introduce constituents into the cathode which will counteract such deterioration. The situation at the time the L3

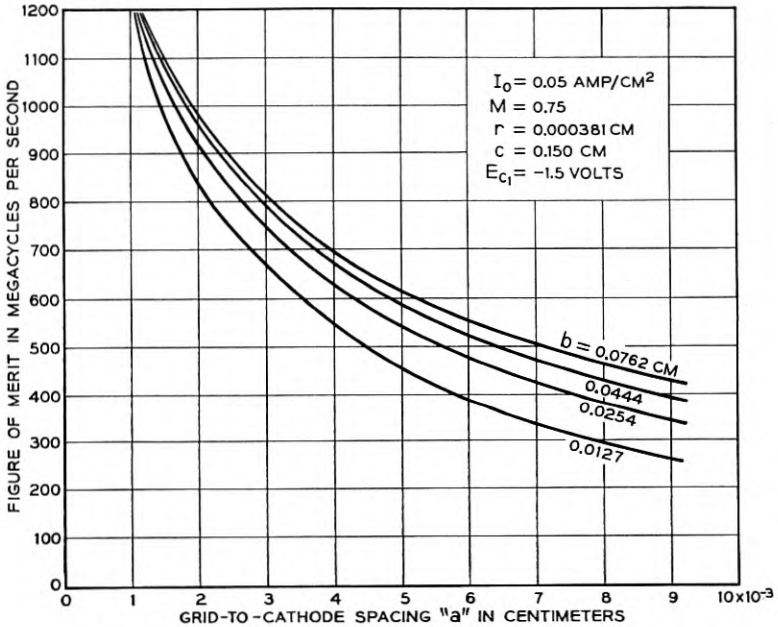



Fig. 2—Figure-of-merit vs. grid-to-cathode spacing.

tubes were being developed was that 50 MA/cm² was as high a current density as seemed to be consistent with the long life required.

It is apparent from the curves in Fig. 2 that the figure of merit increases rapidly as the grid-cathode spacing "*a*" is reduced. The limitation here is mechanical and manifests itself in two ways. One is the practical difficulty of spacing the parts so closely with sufficient accuracy. The other is the problems associated with fabricating grids wound with wire of small enough diameter to make effective use of the close grid-cathode spacing. This part of the subject will be discussed in detail later. It is one of the most important aspects of the design of the L3 tubes.

It would appear from Fig. 3 that the grid-screen spacing "$b$" should be as large as possible. However, the required screen voltage increases as "$b$" increases, and it is desirable to keep the screen voltage low. Therefore, "$b$" is made as low as possible without causing too much penalty on figure of merit. A good compromise value for "$b$" depends on the range of grid-cathode spacing "$a$" being considered, but it will usually be from 0.005 cm to 0.020 cm for close spaced tubes.

Figure 4 shows that the figure of merit increases as "$c$" increases, but there is very little advantage in making it more than 0.040–0.050 cm. Making it much larger also increases the outside dimensions of the struc-

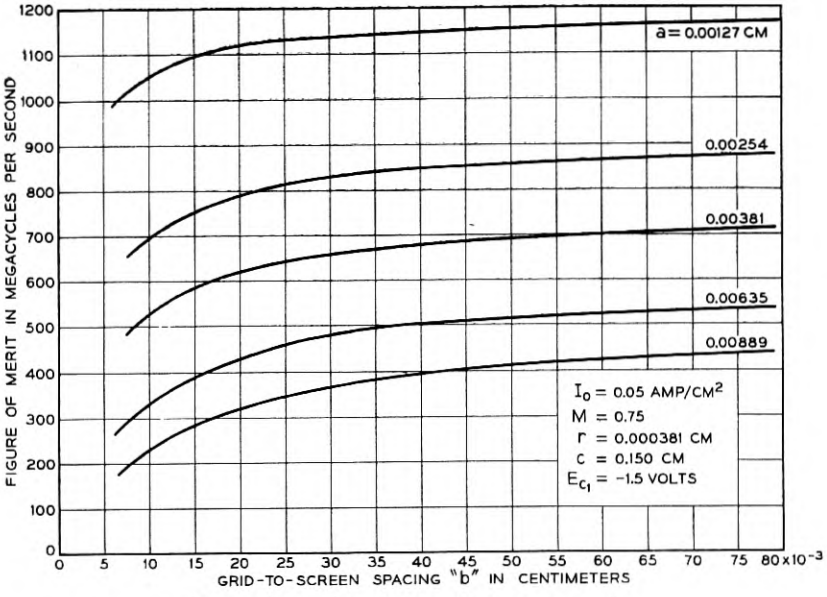



Fig. 3—Figure-of-merit vs. grid-to-screen spacing.

ture unnecessarily, and eventually leads to a spacing which will cause irregularities in the plate current-plate voltage characteristic due to space charge effects in the screen-plate space.

Although it is not apparent from the curves or from what has been said above, it is desirable to have "$r$" as small as possible. It is obvious that "$r$" must be less than $\frac{1}{2n}$, otherwise the grid is completely closed. Under the assumption that $na = 1$, this means that "$r$" must be less than $0.5a$ if there is to be open space between the grid wires. Actually, it is desirable to have not more than 30% of the projected area of the grid closed, which

means that "$r$" should be less than $0.15a$. In addition to this consideration, it is desirable to have "$r$" considerably less than $0.15a$ so that the required screen voltage will be as low as possible. This comes about because the amplification factor $\mu$ increases as "$r$" is increased, other quantities held constant, and equation (3) shows that $E_{c2}$ increases as $\mu$ increases. The diagram shown in Fig. 5 illustrates the trend in grid-cathode spacings and grid wire sizes. The W.E. 416A tube (formerly BTL 1553) represents the

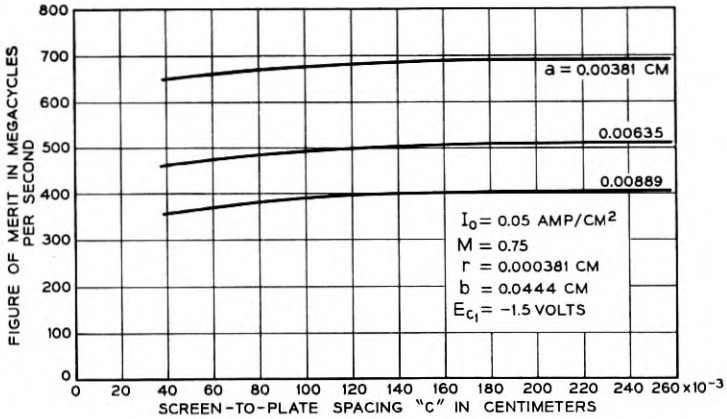



Fig. 4—Figure-of-merit vs. screen-to-plate spacing.

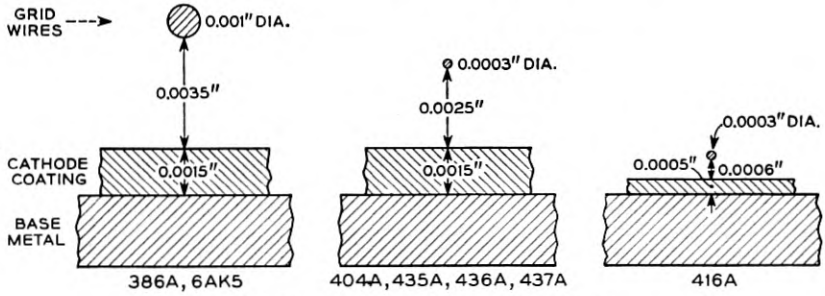



Fig. 5—Trend in spacing and grid wire size.

greatest extension of this trend reported as far as grid-cathode spacing is concerned.[3]

The figure of merit increases as the absolute value of the bias $E_{c1}$ is reduced, since $I_0$ increases. However, a minimum bias value of about $-1.5$ volts is usually necessary in order to avoid undesirable effects due to the

[3] "Design Factors of the Bell Telephone Laboratories 1553 Triode," J. A. Morton and R. M. Ryder, *B.S.T.J.*, Oct. 1950.

collection of electrons of high initial velocity by the control grid. Such grid currents contribute to the noise, cause input loading, and may also cause excessive signal distortion.

Since $I_0$ increases as the screen voltage $E_{c2}$ is increased, the figure of merit likewise increases. It is desirable to keep $E_{c2}$ as low as possible for at least three reasons. The most important is that high screen voltage will generally have an adverse effect on tube life. The second is that low power consumption is desirable for economic reasons and the third is that it helps from the standpoint of maintaining low temperatures of the components in an

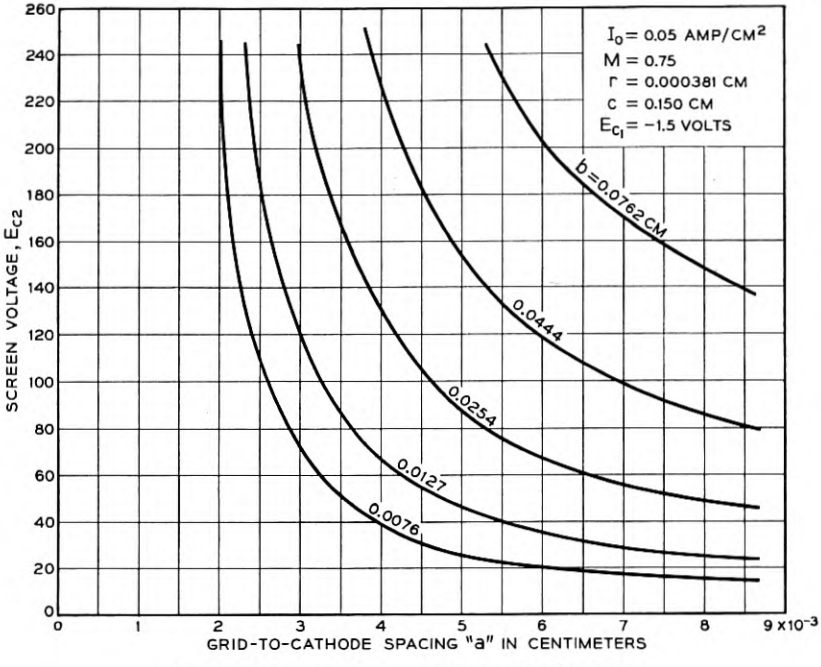



Fig. 6—Screen voltage vs. grid-to-cathode spacing.

amplifier. Figures 6 and 7 show how $E_{c2}$ depends on "$a$" and "$b$". The requirement that $E_{c2}$ be kept low means that the range of "$a$" and "$b$" which can be used is restricted.

### 2.3 *Phase Shift*

The effects of electron transit time and lead inductance in the tubes must be taken into account in order to meet the amplifier requirements with respect to phase margin. In order to maintain stable operation over

the desired transmission band, the gain and phase characteristics must be controlled up to about 200 MC. The amount of phase shift at the frequency where the gain becomes unity ("cross-over point") is of particular interest. In the L3 amplifier this frequency is about 40 MC. Phase shift introduced by electron transit time and by lead inductance is referred to as "excess phase." Ideally, of course, the excess phase would be zero.

The time required for an electron to travel from the cathode to the plate is of the order of $10^{-10}$ seconds. This corresponds to about 5° of excess phase at 40 MC. Close spacings and high electrode potentials tend to reduce the

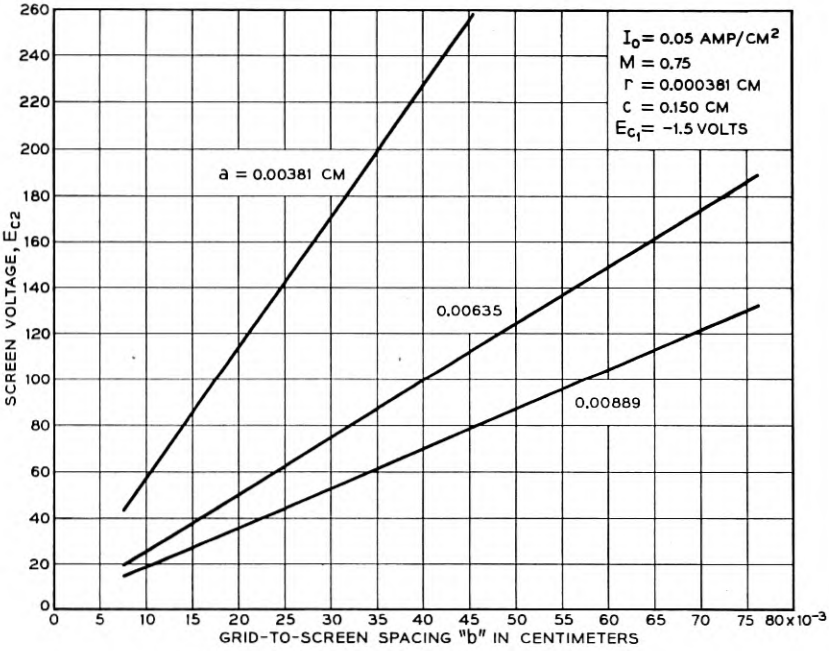



Fig. 7—Screen voltage vs. grid-to-screen spacing.

transit time. However, the considerations discussed in Section 2.2 have been the major factors in setting the spacings and potentials because the transit time, though important, is far less so than the figure of merit.

By using relatively heavy lead wires and mounting the tube structure in such a way as to make the lead wires as short as possible, the additional excess phase due to the lead wires has been minimized so that it amounts to about 5° at 40 MC.

In order to insure adequate margin against a singing condition, the amplifier has been designed to have about 20°–30° less phase shift at 40 MC

with these tubes than that which will cause singing. With this situation, it can be seen that any substantial factor of increase in the excess phase introduced by the tubes, or any other components, could begin to reduce the phase margin seriously.

### 2.4 *Noise*

Fluctuation noise is an important factor in the W.E. 435A used in the first stage of the input amplifier and in the W.E. 436A used in the first stage of the output amplifier. There is adequate margin against the effect of low frequency noise components such as microphonics, power frequency hum, and "sputter noise" if reasonable precautions in tube and circuit design are taken. From a design standpoint, the fluctuation noise is minimized by adopting a combination of cathode temperature and current density drawn such that, with a normally active cathode, the space current is substantially space charge limited, with ample margin for some loss of cathode activity in service before the temperature limited condition is approached. When the temperature limited region is reached, the noise is substantially higher than for the space charge limited condition. The temperature and the cathode current density ratings for these tubes have been set at values which take these considerations into account.

### 2.5 *Modulation*

Since a major purpose of using feedback is to reduce the modulation products arising in the amplifiers, the more nearly an ideal linear transfer characteristic can be approached in the tube design the better, because less feedback is required to obtain a given grade of system performance. Unfortunately, however, a conventional triode or tetrode type of vacuum tube operating under normal space charge limited conditions necessarily has a transfer characteristic which is non-linear. Several possible special structures which might give less modulation were explored, but none were found which would provide the required figure of merit and be sufficiently stable and reproducible.

Considerable emphasis was placed on the problem of controlling the variation in modulation from tube to tube. The most important factors are grid-cathode spacing, uniformity of grid pitch, and cathode activity. Although these factors must be well controlled for other reasons also, the special requirements on modulation necessitated a thorough investigation.

The effect of the grid-cathode spacing can be expressed in terms of the d-c. plate current and the signal level. For a triode having an idealized three-halves-power transfer characteristic with $\frac{d\mu}{dE_{c1}} = 0$ as in Section 2.2,

and for small signals, the ratio of the fundamental signal current to the second harmonic component for the case of a very small load impedance is

$$\frac{I_p}{I_{2p}} = 12\frac{I_b}{I_p} \quad \text{(see appendix for derivation)} \tag{11}$$

This means that, for a given signal current amplitude $I_p$ in the output, a tube having the assumed characteristics will give a ratio which depends only on the d-c. plate current, which in turn is very sensitive to changes in grid-cathode spacing.

A study of the variations in grid pitch and their effects on modulation in a particular experiment showed that reducing the standard deviation of the pitch distance from 16% to about 7% reduced the second-order modulation by 4 db. Since the second-order modulation must be reduced by feedback which is at a premium in the L3 system, this experiment showed that control of the grid pitch was important, and that periodic checks on this factor would be desirable in manufacture.

The effect of cathode activity on modulation was studied in diodes so as to eliminate the effects of grid variations. The variations in modulation from tube to tube were found to be about the same as when grids were present. The geometry of the diodes was so closely controlled that dimensional variations could not account for the differences in the modulation levels. This part of the investigation led to a recognition of the importance of obtaining the best possible uniformity of cathode activity. It also became apparent that the surface condition of the anode was a factor, and that it is therefore desirable to maintain a high degree of cleanliness of the electrodes to which positive potentials are applied.

### 2.6 *Life*

Long tube life is a very important requirement in the L3 system. The most important consideration is the effect of the life on the reliability of the system. There is also the obvious effect of the life on maintenance costs.

Short life tends to reduce the reliability of a system which contains a great number of tubes because the potential failures cannot be predicted so accurately as when the life is long, without a prohibitively costly amount of testing. Even with the most frequent and accurate testing procedure which might be considered, it would be amazing if more than 90% of the potential failures were replaced before causing transmission trouble. To illustrate the effect of short life, consider a 100-mile section of L3 line. There will be five tubes in each of 24 amplifiers. If the performance of any one of the total of 120 tubes becomes poor enough to make the circuit uncommercial, that section must be taken out of service until the defective

tube (or amplifier) is replaced. Now, for a going system, with 120 tubes, and assuming an abnormally short life of say 1200 hours, a tube will fail every ten hours on the average unless preventative testing is used. Even if very frequent testing be done in order to replace 90% of the potential failures before they occur, one circuit interruption every 100 hours may be expected.

It is evident that a life many times greater than that assumed in this illustration is imperative if reliable service is to be obtained and costly maintenance avoided. Laboratory life tests predict that a tube life of at least 15,000 hours may be expected in the L3 system. The actual results will depend on the extent to which the operating conditions are closely controlled, the severity of the field rejection limits, and the ability of the tube factory to control the processing.

### 2.7 *Interchangeability*

The objective is to make the characteristics of the tubes sufficiently uniform so that tubes may be replaced at will without circuit adjustments being needed. In the L3 amplifiers, the circuits have been designed so that a relatively wide range of characteristics can be accepted for individual tubes. However, it is essential that the average characteristics be held in close control from one manufacturing lot to another. This has been provided for by setting up distribution requirements which will be discussed further in a later section.

### 2.8 *Cost*

As will be seen from the description which follows, it has been possible to meet the L3 requirements with tube designs which do not require too great departures from the manufacturing methods employed for conventional telephone tubes. With a reasonable demand, it is accordingly expected that the tube costs should be moderate.

## 3. Design Description and Characteristics

### 3.1 *Mechanical Description and Mechanical Problems*

Figure 8 shows the three L3 tubes along with some of the earlier high figure of merit tubes. The W.E. 386A (left hand side) was designed to be soldered directly into the circuits and had its input lead at the stem end while the output lead came out through the top of the bulb. The flexible leads used for soldering purposes, and the double-ended lead construction,

add materially to the cost of tube construction and testing. Early in the L3 tube development the question of factory cost compared with circuit performance was weighed and it was decided that the advantage of lower tube cost plus the very large advantage of simple plug-in tubes would outweigh

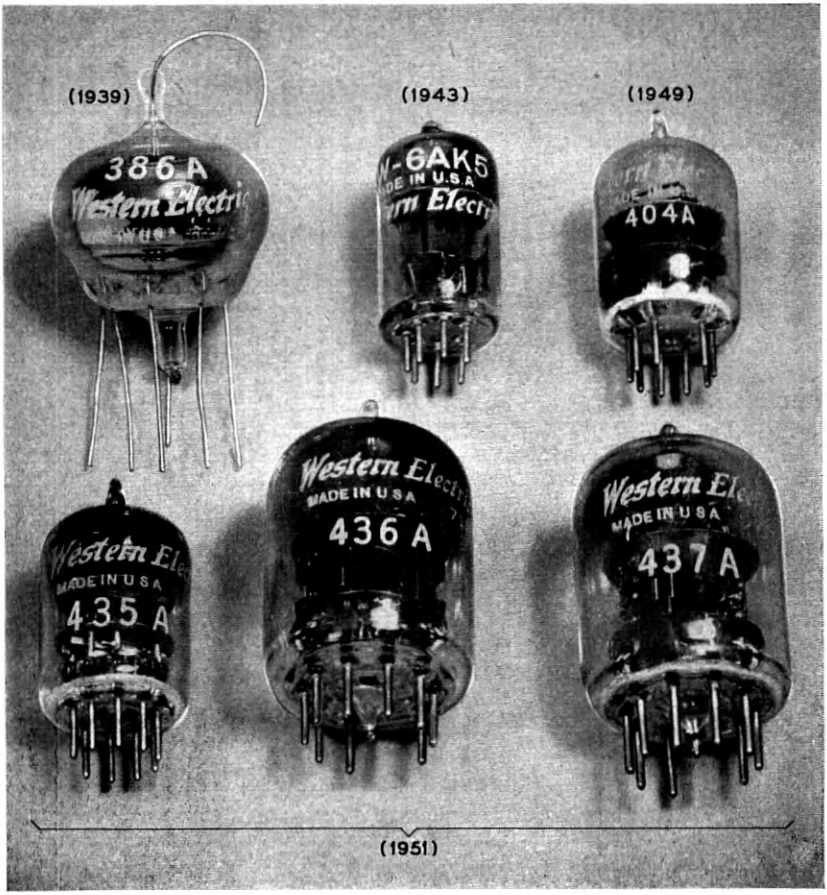



Fig. 8—The 386A, 408A, 404A, 435A, 436A and the 437A tubes approximately actual size.

the cost in performance. Accordingly, all three L3 tubes are of the stiff pin, plug-in type designed to fit existing sockets. The price paid for obtaining the advantages of lowered cost and interchangeability has been a loss in feedback of approximately 2 db per amplifier.

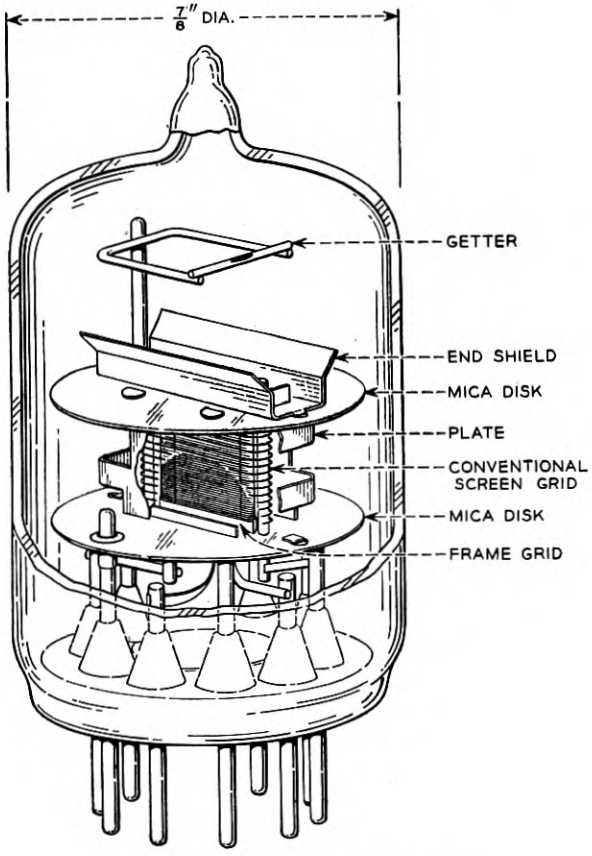



Fig. 9—Cutaway view of the 435A.

Figures 9, 10, and 11 are the cutaway views of the 435A, 436A, and 437A tubes. The overall dimensions of the L3 tubes are:

| | 435A | 436A | 437A |
|---|---|---|---|
| Max. seated height | $1\frac{1}{2}''$ | $1\frac{5}{8}''$ | $1\frac{5}{8}''$ |
| Max. diameter | $\frac{7}{8}''$ | $1\frac{3}{16}''$ | $1\frac{3}{16}''$ |
| Number of pins | 9 | 9 | 9 |
| Pin circle diameter | $\frac{15}{32}''$ | $\frac{11}{16}''$ | $\frac{11}{16}''$ |

Conventional construction for small repeater tubes may be thought of as two mica wafers between which are assembled the active elements of the tube. The mica wafers serve to support and space the tube elements. This

mica and element assembly is then mounted upon a glass stem or platform which is next sealed into a glass bulb after which the exhaust and activation procedures complete the tube. These cutaway views show that these L3 system tubes are very similar to conventional tubes. The reason for wanting to continue with the conventional type structures is simply that of tube

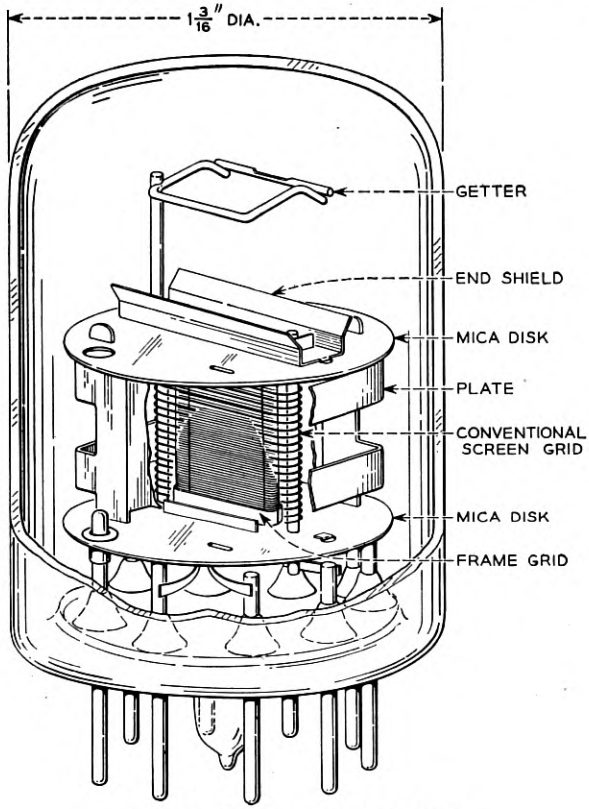



Fig. 10—Cutaway view of the 436A.

cost. A production line such as that for the W.E. 408A or 6AK5 tubes could very readily be changed over to any one of these tubes. The only change needed would be in the control grid supply and in the dimensional control procedures.

The principal distinctive design feature in these tubes, compared to earlier repeater tubes, is the "frame" type of control grid which was first introduced in a somewhat different form in the W.E. 404A[4] and W.E.

[4] "The 404A- A Broadband Amplifier Tube," G. T. Ford, *Bell Laboratories Record*, Vol. XXVII Feb. 1949.

418A tube. The L3 frame grids are illustrated in Fig. 12, together with a conventional control grid from the 6AK5 tube and the control grid from the 404A vacuum tube. The conventional grid consists of two large side rods, usually of nickel, around which is spirally wound the grid lateral wire. The lateral wire is joined to the side rod at each intersecting point by first knifing a groove into the side rod, laying the lateral wire into the groove,

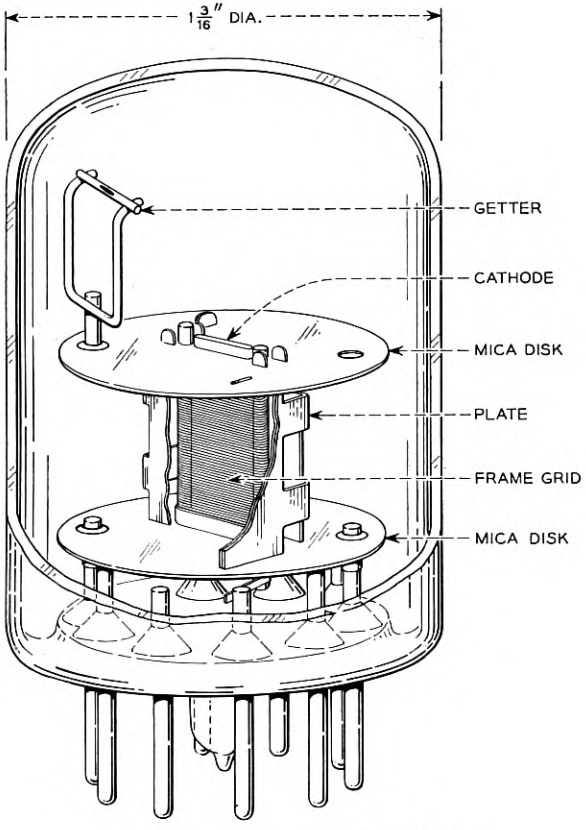



Fig. 11—Cutaway view of the 437A.

and then swaging the groove closed. Since, in these conventional grids, the lateral wire is usually larger than 0.0008″ diameter, the grid is self supporting and needs no strengthening members. For the high figure of merit tubes, control grid lateral wires of the order of 0.0003″ diameter are needed. Wire of this diameter is not self supporting in the necessary lengths and for that reason the two large side rods are first joined together by the cross straps

which are located at the ends of the grid proper. These then form a rigid frame around which the very fine lateral wire can be spiraled without any danger of having laterals out of place. It can be seen that this technique produces the extremely flat grid plane which is necessary for the desired

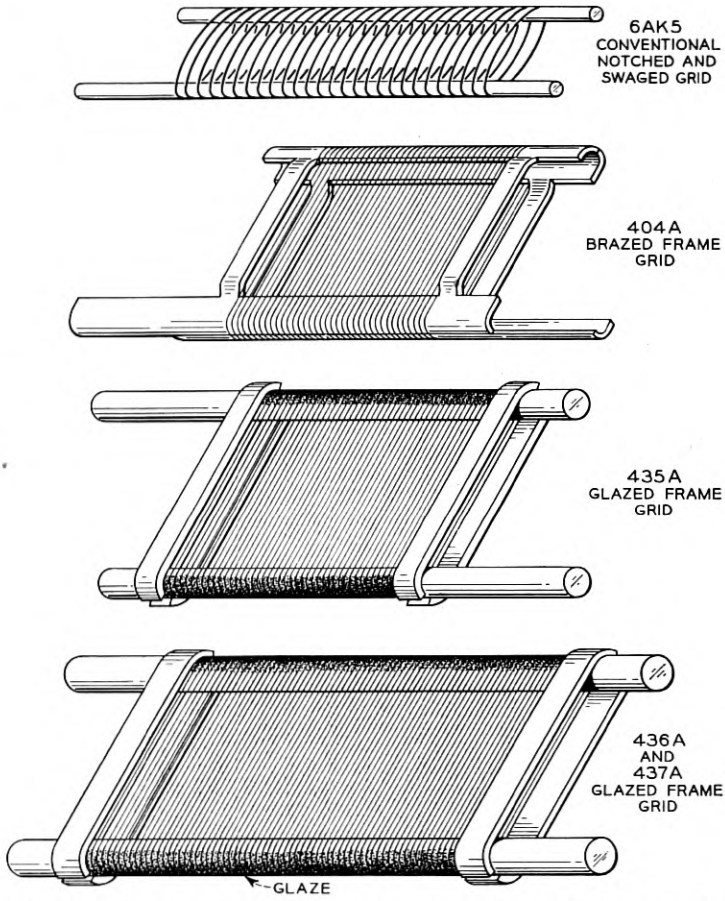



Fig. 12—Control grids for the 408A or 6AK5, the 404A and the L3 carrier tubes.

tube performance, and which is the real difference between these high figure of merit tubes and the more conventional tubes.

The fabrication of the 404A frame grid has been discussed in a previous article.[5] That article also mentioned the 418A grid which is a side rod type

[5] "Fine-Wire Type Vacuum Tube Grid," E. J. Walsh, *Bell Laboratories Record*, Vol. XXVIII April 1950.

frame grid. The L3 grids are a further development. The major difference between the earlier frame grids and these is in the method of bonding the 0.0003″ lateral wire to the side rods. In the earlier grids a gold braze was used to bond the laterals to the side rods. This necessitated heating the unit to approximately 1070°C to flow the gold. The newer grids have the lateral wires bonded by a glass glaze which allows the process to be carried out at approximately 700°C. There is a differential expansion between molybdenum and tungsten of about five to four. The net result of the reduction of temperature in this process is that the tungsten wires are stretched less at the lower temperature and thus when returned to room temperature have higher residual tensions. This is important because the higher the residual tension the higher the resonant frequency of the lateral wires. This in turn means that the noise level of the tube due to vibration or shock will be reduced since loose grid wires will give rise to microphonic noises. Tighter wires also decrease the possibility of grid to cathode shorts. Tests have shown that an increase of about 25% in the resonant frequencies of the lateral wires can be expected as a result of using the glass glazing technique as compared to the gold brazing technique.

It is interesting to note that the residual stress in the lateral wires of these grids is of the order of 200,000 pounds per square inch. This figure is roughly ten times as great as the allowable working stress for steel beams such as are used in the construction industry.

When the glazing technique is used, the grid is gold plated after the glazing operation has been completed. The gold is used to inhibit thermionic emission from the grid wires. This is a necessity for tubes of this type when used in the circuits for which they were designed. The need for the plating exists because of the proximity of the grid wires to the hot cathode and their unfortunately favorable position for receiving a deposition of active material from the cathode during its processing and operation. The desired amount of gold on the grid wires is that which will cause a diameter increase of about 0.00002″. This is an extremely difficult increase to measure because the measurement must be non-destructive, since it is made on the finished grid and is used as a production control. The method used to date has been an optical measurement at a magnification of about 500×.

A very high degree of precision, compared to that previously available, has been obtained for some of the parts whose dimensions are critical. The cathode sleeve is now obtainable with minor axis limits of ±0.0003″. The mica discs are now made with the critical holes to that same tolerance. The frame grid side rod is made to ±0.0001″. These are the basic elements of the tube and, after inspection has shown them to be acceptable, their assembly becomes close to that of a conventional tube. The inspection of

these parts is difficult when production numbers are considered. The micas in particular presented a serious problem. Mica sheet is composed of a large number of laminations many of which are of the order of 0.0001″ in thickness. When the mica discs are punched out, these laminations leave, not smooth edge holes as do metal stampings, but rather a large number of minute jagged edges. The method used to check these was an optical one in which the mica was projected at about 40 times size onto a glass screen on which engraved lines acted as go-no-go gages. This reduced tool and human error considerably. The cooperation of several industrial concerns which supply some of the critical parts and the measuring instruments was very helpful in obtaining the desired tolerances.

It was evident from the start of the development of the L3 tubes that the performance requirements for high gain conventional structure tubes would be pushing to the limit the available process controls and measuring techniques. A statistical quality control program was put into effect on the tubes after the final laboratory design had been crystallized. The statistical study covered the tube dimensions and the data collected from those tubes after they had been processed. The net result of the study was to indicate that better measuring methods and process controls are needed.

With the amount of d-c. feedback employed in the working circuits, the space current does not vary too rapidly with tube geometry. In the case of the most critical spacing, that between the grid and the cathode, a 10% change in the spacing would be expected to cause only about 2.5% change in space current. However, the transconductance is more sensitive to the grid-cathode spacing, with a 14% change in transconductance to be expected for a 10% change in the spacing. This comes about because the transconductance is a function of the spacing, even at a fixed space current.

Since a 10% change in spacing is only 0.00025 inch, the importance of close tolerances on the parts dimensions controlling it is evident. The test specification limits on transconductance permit a variation of about ±25%, so that the 0.00025 inch change in spacing would use up over half of the allowed deviation. Preproduction runs at the Laboratories have shown that the tubes are practical and that their performance in the amplifier circuits has justified their design.

### 3.2 *Electrical Characteristics*

The nominal electrical characteristics are shown in Table II. The corresponding characteristics for the earlier types 386A, 6AK5 and 404A are also given for comparison. The last row in the table shows figure of merit values which are a measure of the circuit performance. The tabulated values for the figure of merit were calculated, taking into account the effect of

space charge on the input capacitance, and include a total allowance of 3 mmf. for socket and wiring capacitance (input plus output). The L3 tubes are somewhat better than the 404A and substantially better than the 6AK5. The figures of merit tabulated will not check with the values shown in Figs. 2–4 since the curves were calculated for cold tube capacitances and zero socket and circuit capacitances.

TABLE II

| Classification | 386A Pentode | 6AK5 Pentodes | 408A Pentodes | 404A Pentode | 435A Tetrode | 436A Tetrode | 437A Triode |
|---|---|---|---|---|---|---|---|
| Heater voltage | 6.3 | 6.3 | 20.0 | 6.3 | 6.3 | 6.3 | 6.3 volts |
| Heater current | 0.150 | 0.175 | .05 | 0.30 | 0.30 | 0.45 | 0.45 amps |
| Plate current | 7.5 | 7.5 | | 13 | 13 | 25 | 40 ma |
| Screen current | 2.5 | 2.5 | | 4.5 | 3.5 | 8 | — ma |
| Transconductance | 4000 | 5000 | | 12500 | 15000 | 28000 | 45000 μmhos |
| Input capacitance* | 3.6 | 3.9 | | 7.0 | 7.8 | 15.2 | 11.5 mmf |
| Output capacitance* | 2.6 | 2.85 | | 2.5 | 2.5 | 3.3 | 0.9 mmf |
| Plate-grid capacitance* | 0.025 | 0.01 | | 0.03 | 0.025 | 0.05 | 3.5 mmf |
| Figure of merit** | 61 | 72 | | 123 | 146 | 165 | — mc |

* Cold capacitances.

** This is the frequency at which unity voltage amplification would occur with a simple parallel tuned circuit interstage. Allowances have been made for stray capacitances and for the increase in input capacitance when a tube is energized. No figure is given for the 437A tube because the relations derived for the earlier stages in the amplifier do not apply to the output stage.

### 3.3 *Performance in Repeater*

The positions of the tubes in an auxiliary repeater are indicated in the diagram of Fig. 13. The overall insertion gain is a little over 30 db at 4 mc. While the noise contributions of the first 435A tube and the 436A tube are important, they have been reduced to 48 db below one volt and 62 db below one volt, respectively, for 1200 repeaters. The second 435A tube and the "lower" 437A tube appearing in Fig. 13 are the major contributors to the modulation. The expected modulation levels from these tubes are those associated with single tone ratios of about 34 db for the fundamental to second harmonic ratio and 70 db for the fundamental to third harmonic ratio, with a grid signal level of 0.1 volt r.m.s.

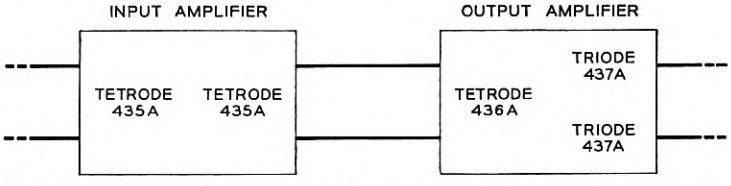



Fig. 13—Position of tubes in the input and output amplifiers for the L3 carrier system.

The gain-band performance in this repeater, or in any other circuit, will be less than that shown in the curves in Figs. 2–4 since the total shunt capacitances in a working circuit are always larger than the cold tube capacitances used in calculating the inherent figure of merit.

### 3.4 *Test Specifications*

The test specifications for the L3 tubes were written with the L3 system requirements as the prime consideration. In addition to the usual tests made on small tubes, a modulation test was included for the 435A and 437A tubes because of the importance of this characteristic in terms of system performance. In order to avoid penalties which can result from unwanted systematic deviations which pile up in a long system, requirements have been set up which will control the distribution of transconductance, modulation, and some of the most critical interelectrode capacitances. By the application of suitable quality control methods, it is expected that these requirements can be met economically and that such measures will prevent the manufacture of large numbers of tubes having average characteristics far off the design values. When simple go, no-go limits are used, it is not economical to set close enough limits to attain the desired control of the average characteristics.

## 4. Conclusions and Future Possibilities

The fundamental problem in the development of repeater tubes for broad band coaxial systems has been to devise means for utilizing closer and closer grid-cathode spacings without sacrificing life performance.

The closer spacings have been made possible by devising rigid control grid supporting structures which can be wound with very small diameter wire. The wire is held under tension by the supporting frame so that a flat winding is produced which can be spaced very close to a flat cathode.

The possibilities for the development of tubes which will provide still better broad band amplification depend to a great extent upon the kind of system design to be considered. If higher figure of merit, as defined in this article, can be utilized, considerable improvement can be realized with space charge controlled tubes such as the 435A, 436A and 437A by using mounting arrangements which provide more precise means of establishing and maintaining the critical dimensions.

## Acknowledgements

Several members of the technical staff and their assistants have contributed materially in solving the numerous technical problems which arose during the development of these tubes. In addition, those who fabricated

the grids and assembled the experimental tube models made important contributions in terms of skill and painstaking effort. It is not practical to name all of the persons involved.

The development of broad band tubes over the last fifteen years was under the direction of the late Dr. H. A. Pidgeon until 1943, and Mr. J. O. McNally from 1943 to date. The writers wish to acknowledge the importance of their helpful guidance and encouragement.

## APPENDIX

*Meanings of Symbols*

$F$ = Figure of merit
$G$ = Voltage amplification
$B$ = Bandwidth between 3 db points
$K$ = Interstage circuit constant
$G_m$ = Grid-plate transconductance
$C_1$ = Input capacitance
$C_2$ = Output capacitance
$n$ = Turns per unit distance on grid
$a$ = Grid-cathode spacing
$b$ = Grid-screen spacing
$c$ = Screen-plate spacing
$A$ = Area of active structure
$I_b$ = Plate current, d-c
$M$ = Ratio of plate current to cathode current
$E_{c1}$ = Grid-cathode voltage
$E_{c2}$ = Screen-cathode voltage
$\mu$ = Grid-screen amplification factor
$I_0$ = Cathode current density, d-c
$r$ = Grid wire radius
$I_p$ = Amplitude of fundamental component of a-c plate current
$I_{2p}$ = Amplitude of second harmonic component of a-c plate current
$i_p$ = Fundamental a-c plate current component
$i_{2p}$ = Second harmonic plate current component
$G_0$ = Perveance factor
$E_g$ = Amplitude of grid signal voltage
$p$ = Frequency $\times 2\pi$
$t$ = Time
$i$ = a-c plate current

Units: length in cms; practical electrical units; time in seconds.

*Assumptions*

I $$na = 1$$

The ratio of the pitch distance to the grid-cathode spacing is held constant. This is done so that the effect of the variation in field along the cathode surface resulting from the finite grid pitch distance will be small and the same throughout the discussion.

II $$C_1 = 0.0885 \times 10^{-12} A \left(\frac{1}{a} + \frac{1}{b}\right)$$

$$C_2 = \frac{0.0885 \times 10^{-12} A}{c}$$

The input and output spaces are treated as if they can be represented by ideal condensers. This amounts to assuming that the grids are perfect planes and that there are no end effects. The effects of space charge on the capacitances are neglected, and the socket and wiring capacitances are also neglected. This means that the resulting calculated figure of merit represents the limiting value inherent in the tube structure.

III $$I_B = \frac{2.33 \times 10^{-6} MA \left(\frac{E_{c2}}{\mu} + E_{c1}\right)^{3/2}}{a^2 \left(1 + \frac{1}{\mu} \frac{a+b}{a}\right)^{3/2}}$$

The expression for plate current assumed is an idealized one, but holds fairly well for these tubes.

IV $$\frac{d\mu}{dE_{c1}} = 0$$

It is assumed that the triode amplification factor is independent of the control-grid voltage. This holds fairly well under the conditions of assumption I.

V When the interstage consists of a single parallel tuned circuit, $K = 1$. This case is assumed.

*Derivations*

Beginning with the above assumptions, and substituting in equation (1), equations (2), (3) and (4) can be derived. The procedure will be outlined below:

$$F = \frac{KG_m}{2\pi(C_1 + C_2)} \tag{1}$$

$$F = \frac{4.74 \times 10^8 M I_0^{1/3}}{a^{4/3}\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right)\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{2}$$

$$E_{c2} = \mu\left[5.68 \times 10^3 a^{4/3} I_0^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right) - E_{c1}\right] \tag{3}$$

$$\mu = \frac{2.73\frac{b}{a} - \log_{10}\cosh\left(2\pi\frac{r}{a}\right)}{\log_{10}\coth\left(2\pi\frac{r}{a}\right)} \tag{4}$$

Substitutions for $C_1$, $C_2$, and K in (1) are made from assumptions II and V. $G_m$ is found by differentiating the plate current expression (assumption III) with respect to $E_{c1}$, remembering that $\mu$ is independent of $E_{c1}$ according to assumption IV.

$$G_m = \frac{3}{2}\frac{2.33 \times 10^{-6} MA\left(\frac{E_{c2}}{\mu}+E_{c1}\right)^{1/2}}{a^2\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}} \tag{5}$$

From assumption III we can write

$$\left(\frac{E_{c2}}{\mu}+E_{c1}\right)^{1/2} = \frac{I_B^{1/3} a^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{1/2}}{(2.33 \times 10^{-6})^{1/3} M^{1/3} A^{1/3}} \tag{6}$$

Substituting in (5)

$$G_m = \frac{3}{2}\frac{(2.33 \times 10^{-6}) MA I_B^{1/3} a^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{1/2}}{a^2\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}(2.33 \times 10^{-6})^{1/3} M^{1/3} A^{1/3}}$$

$$G_m = \frac{1.5(2.33 \times 10^{-6})^{2/3} M^{2/3} A^{2/3} I_B^{1/3}}{a^{4/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{7}$$

Since $I_B = MI_0A$ by definition,

$$G_m = \frac{2.64 \times 10^{-4} MA I_0^{1/3}}{a^{4/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{8}$$

Substituting in (1),

$$F = \frac{2.64 \times 10^{-4} MAI_0^{1/3}}{a^{4/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right) 2\pi \left[.0885 \times 10^{-12} A \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c}\right)\right]} \tag{9}$$

Collecting the constants and cancelling the A's,

$$F = \frac{4.74 \times 10^{8} MI_0^{1/3}}{a^{4/3}\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c}\right)\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)} \tag{2}$$

The expression for $E_{c2}$ can be found by substituting $I_B = MI_0A$ in assumption III and solving for $E_{c2}$.

$$MI_0A = \frac{2.33 \times 10^{-6} MA\left(\frac{E_{c2}}{\mu} + E_{c1}\right)^{3/2}}{a^2\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}}$$

$$\frac{E_{c2}}{\mu} + E_{c1} = \frac{a^{4/3} I_0^{2/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)}{1.76 \times 10^{-4}}$$

$$E_{c2} = \mu\left[5.68 \times 10^{3} a^{4/3} I_0^{2/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right) - E_{c1}\right] \tag{3}$$

The expression for $\mu$ can be found by applying assumption $I$ and substituting $n = \frac{1}{a}$ in the Vogdes-Elder formula* for a plane structure.

$$\mu = \frac{2\pi nb}{2.303 \log_{10} \coth (2\pi nr)} - \frac{\log_{10} \cosh (2\pi nr)}{\log_{10} \coth (2\pi nr)} \tag{10}$$

Substituting $n = \frac{1}{a}$

$$\mu = \frac{2.73\,\frac{b}{a}}{\log_{10} \coth\left(2\pi\,\frac{r}{a}\right)} - \frac{\log_{10} \cosh\left(2\pi\,\frac{r}{a}\right)}{\log_{10} \coth\left(2\pi\,\frac{r}{a}\right)} \tag{4}$$

Equation (11) can be derived by considering a particular structure and introducing a small sinusoidal signal $E_g \cos pt$ added to the d-c. voltage in the plate current expression, Assumption III. This can be written as

$$I_B + i = G_0\left(\frac{E_{c2}}{\mu} + E_{c1} + Eg \cos pt\right)^{3/2} \tag{12}$$

* F. B. Vogdes and Frank R. Elder, *Phys. Rev.*, 24, pp. 683–689, Dec., 1924.

For zero signal this becomes

$$I_B = G_0 \left( \frac{E_{c2}}{\mu} + E_{c1} \right)^{3/2} \tag{13}$$

For a pure resistance load which is small compared to the plate resistance, the fundamental component, neglecting contributions from third and higher order terms, is

$$i_p = G_m E_g \cos pt \tag{14}$$

Neglecting the contributions of fourth and higher order terms, the second harmonic component is

$$i_{2p} = I_B \frac{3}{16} \frac{E_g^2}{\left( \frac{E_{c2}}{\mu} + E_{c1} \right)^2} \cos 2pt \tag{15}$$

This is found by expanding (12) into the binomial series. From Assumption III and equation (5),

$$\left( \frac{E_{c2}}{\mu} + E_{c1} \right)^2 = \frac{9 I_B^2}{4 G_m^2} \tag{16}$$

The amplitude of the second harmonic component, from (15) is

$$I_{2p} = \frac{3}{16} \frac{I_B E_g^2}{\left( \frac{E_{c2}}{\mu} + E_{c1} \right)^2} \tag{17}$$

Substituting for $\left( \frac{E_{c2}}{\mu} + E_{c1} \right)^2$ from (16) in (17).

$$I_{2p} = \frac{G_m^2 E_g^2}{12 I_B} \tag{18}$$

From (14),

$$G_m^2 E_g^2 = I_p^2 \tag{19}$$

Substituting from (19) in (18),

$$I_{2p} = \frac{I_p^2}{12 I_B} \tag{20}$$

This can be written

$$\frac{I_p}{I_{2p}} = 12 \frac{I_B}{I_p} \tag{11}$$

# Telephone Traffic Time Averages

By JOHN RIORDAN

(*Manuscript Received April 25, 1951*)

This paper describes the determination of the first four semi-invariants of the distribution of the average, over an arbitrary time interval, of traffic carried by a telephone system with an infinite number of trunks, during a period of statistical equilibrium. Both finite and infinite numbers of independent call sources are considered, and the distribution function of call holding times is left general.

## 1. Introduction

FOR mathematical studies of telephone traffic, like those of call loss or delay which are used in trunking engineering, the traffic is considered as a flow of probability in time. In the period of most importance, the busy hour, this flow is usually regarded as stationary; that is to say, the probability of a given number of busy trunks, or the probability of delay of an incoming call (or any other probability of the system which comes in question) is taken as independent of the particular moment in the busy hour at which the system is examined. The system is said to be in statistical equilibrium.

For such theoretical studies, the statistical quantities which determine these probabilities, like the rate at which calls appear, are of course taken as given, but in the application they must be determined by observations, such as those being taken in the current extensive program of traffic measurements. Here a difficulty appears. To abridge the extensive amount of observational material, either measurements are made of traffic averages over periods small compared to the busy hour (but not small enough to be neglected) or the measurements of continuous recorders are averaged by hand. It may be noticed here that for application of the results given below the traffic averages obtained by measurements must be those of a continuous device which records all traffic changes and not, as in some measuring devices, those obtained from a number of "looks" at points within the averaging interval. But to use these measurements in determining the traffic parameters by standard sampling theory, a corresponding theoretical study of the averages is necessary.

Such a study, within limits to be described presently, is given here. No attempt is made to describe the sampling studies possible from the results reached. These seem to be of many kinds, not necessary to describe, but for

concreteness it may be mentioned that the most important, at the moment, seems to be that of setting confidence limits for the average traffic.

The most important of the limits to this study are those implied by the assumptions of statistical equilibrium with fixed average, and an infinite number of trunks. The former limits application to periods in which, roughly speaking, average traffic is neither rising nor falling; the latter is justified only by the extreme mathematical difficulties produced by assuming otherwise. The traffic variable is the number of busy trunks in a period of statistical equilibrium. For pure chance call input, the call holding time characteristic is left arbitrary throughout the development, but main interest lies in the two extreme cases of constant holding time and exponential holding time, which are examined in detail.* For calls from a limited number of sources, results are obtained only for exponential holding time.

More precisely, if $N(t)$ is the random variable for the number of busy trunks at time $t$, the variable studied, the average number of calls in an interval of length $T$, is

$$M(T) = \frac{1}{T}\int_0^T N(t)\,dt \tag{1}$$

The question is: What are the statistical properties of $M(T)$?

The results given are the first four cumulants (semi-invariants) of $M(T)$, which seem to have the simplest expressions. For the convenience of the reader it may be noticed that the first cumulant is the mean, the second the second moment about the mean which is the variance, the third the third moment about the mean, and the fourth the fourth moment about the mean less three times the square of the variance.

In all cases the mean of $M(T)$ is the mean of $N(t)$ and for pure chance call input is called $b$, the average number of calls in unit average holding time, $h$.

The other cumulants for pure chance call input, $k_n$, have the general expression

$$k_n = b\,\frac{n(n-1)}{T^n}\int_0^T dx g(x)(T-x)x^{n-2}; \qquad n = 2,3,4$$

with

$$g(x) = \frac{1}{h}\int_x^\infty f(t)\,dt$$

* F. W. Rabe [6] has reported results for these two cases for relatively long averaging intervals, which are verified below. I owe my interest in this problem to a report on Rabe's work made by Messrs. Gibson, Hayward and Seckler in a probability colloquium at Bell Telephone Laboratories initiated and directed by Roger Wilkinson.

and $f(t)$ the probability that a call lasts at least $t$, that is, the distribution function of holding times. The specializations of this, for constant holding time and exponential holding time, appear in section 4. The results for finite source input have a similar character.

The procedure in obtaining these is as follows. The cumulants are determined from the ordinary moments (about the origin) and the latter are determined by the integration of expectations. Thus the first moment, the mean is determined from

$$E[M(T)] = \frac{1}{T}\int_0^T E[N(t)]\,dt = E[N(t)] \tag{2}$$

where $E(x)$ is written for the expectation or mean of $x$.

Similarly the second moment is given by

$$E[M^2(T)] = \frac{1}{T^2}\int_0^T \int_0^T E[N(t)N(u)]\,dt\,du \tag{3}$$

and so on for higher moments. Correlation effects appear in (3) in $E[N(t)N(u)]$ and are included in the development by formulation of transition probabilities, that is, those probabilities determining the traffic flow in time. The transition probability $P_{jk}(t)$ is defined as the probability of transition in $t$ from $j$ calls in progress (busy trunks) to $k$ calls in progress, and fixes the inter-relatedness of call probabilities at different time epochs. Only for large values of $t$ are these probabilities independent.

Hence, the first task is to determine these simple transition probabilities, then those of double and triple transitions, then the expected values of pairs, triples and quadruples of numbers of busy trunks, and finally the moments.

## 2. Transition Probabilities

For exponential holding time, and infinite sources, infinite trunks, these probabilities have already been determined by Conny Palm [5]. Palm's work has been summarized both by Feller [1] and by Jensen [3], and describes the whole process, not merely the equilibrium condition. For the equilibrium condition, a different procedure,* similar to that used by Newland [4] for another purpose, allows the assumption of a more general holding time characteristic.

* Thanks are due S. O. Rice for suggesting this, as well as for many corrections and improvements. I also have had the advantage of a careful reading of the mss. by E. L. Kaplan.

For infinite sources, and calls arriving individually and collectively at random with average density $a$, the well-known formula for the probability that exactly $k$ calls arrive in time interval $t$ is the Poisson

$$\pi_k(t) = e^{-at}(at)^k/k! \tag{4}$$

Then, if $P_{ij}(t; k)$ is the conditional probability of transition from $i$ to $j$ when $k$ calls arrive in time $t$,

$$P_{ij}(t) = \sum_{k=0}^{\infty} P_{ij}(t; k)\pi_k(t) \tag{5}$$

Consider $P_{ij}(t; 0)$, that is the (conditional) transition probabilities when no calls arrive. Let the probability that a call lasts at least $t$ be $f(t)$, so that the average holding time $h$ is given by

$$h = \int_0^\infty u[-f'(u)]\, du = \int_0^\infty f(u)\, du \tag{6}$$

The $i$ calls initially in process are independent of each other. Select one of them and suppose the time from its arrival (its age) is $t_1$. Then the probability that it will also exist $t$ units later is the conditional probability $f(t + t_1)/f(t_1)$. Since in equilibrium conditions all moments of arrival have equal probability, the corresponding probability for an arbitrary call is

$$g(t) = \int_0^\infty f(t + t_1)\, dt_1 \div \int_0^\infty f(t_1)\, dt = \frac{1}{h}\int_t^\infty f(u)\, du \tag{7}$$

Hence the transitional probability $P_{ij}(t; 0)$ is the binomial expression

$$P_{ij}(t; 0) = \binom{i}{j} g^j(1 - g)^{i-j} \tag{8}$$

and its generating function is

$$P_i(t, x; 0) = \sum P_{ij}(t; 0)x^j = [1 + (x - 1)g]^i \tag{9}$$

In (8) and (9), for brevity, the argument of $g$ is omitted.

Now, suppose one call arrives in interval $t$. The moment of arrival is uniformly distributed in $t$; that is, if $u_1$ is the moment of arrival,

$$Pr(u < u_1 < u + du) = du/t$$

and the probability that a call arriving at an arbitrary moment will be in existence at time $t$ is, say,

$$Q(t) = \int_0^t f(t - u)\frac{du}{t} = \frac{1}{t}\int_0^t f(u)\, du = \frac{h}{t}(1 - g(t)) \tag{10}$$

The corresponding generating function is

$$1 - Q(t) + xQ(t) = 1 + (x - 1)Q(t)$$

and, since calls arriving are independent, the generating function for $k$ calls arriving is

$$[1 + (x - 1)Q]^k$$

and

$$P_i(t, x; k) = [1 + (x - 1)g]^i[1 + (x - 1)Q]^k \tag{11}$$

Hence, finally by (5),

$$\begin{aligned} P_i(t; x) &= \sum P_{ij}(t)x^j \\ &= [1 + (x - 1)g]^i \sum_{k=0}^{\infty} [1 + (x - 1)Q]^k \frac{e^{-at}(at)^k}{k!} \\ &= [1 + (x - 1)g]^i \exp(x - 1)\,at\,Q \\ &= [1 + (x - 1)g]^i \exp(x - 1)\,ah\,(1 - g) \end{aligned} \tag{12}$$

The last step uses (10).

This is the generating function for the simplest transition probabilities, and is quite like Palm's result; indeed, for exponential holding time $g = f = e^{-t/h}$. The probabilities themselves are obtained by expansion of the generating function in powers of $x$, or by substituting $g$ for $e^{-t/h}$ in Palm's result. But they are not needed here; the generating function is most apt for determining the averages of interest, as will appear.

Before going on to the other transition probabilities, it is interesting to notice certain checks of equation (12). In statistical equilibrium the traffic has Poisson density (Palm l.c.) that is, in the present notation

$$Pr(N(t) = k) = e^{-b}b^k/k!$$

where $b = ah$. This of course is independent of time. Then, if $N(0)$ has this density, so should $N(t)$ as determined from $N(0)$ and the transition probabilities implicit in (12). This is verified by

$$\begin{aligned} \sum P_i(t,x)e^{-b}b^i/i! &= \exp(x - 1)b(1 - g) \sum [1 + (x - 1)g]^i \frac{e^{-b}b^i}{i!} \\ &= \exp[(x - 1)b(1 - g) - b + b + (x - 1)bg] \\ &= \exp(x - 1)b. \end{aligned} \tag{13}$$

Also, $g(0) = 1$ and $g(\infty) = 0$ so that

$$P_i(0, x) = [1 + (x - 1)]^i = x^i \tag{14}$$

$$P_i(\infty, x) = \exp(x - 1)b \tag{15}$$

showing that in zero time no transit to another state is possible, and in infinite time the equilibrium probabilities are reached no matter what the initial state has been.

Finally, in a Markov process (cf. Feller [2], Chap. 15) the simple transition probabilities alone are needed since

$$P_{ijk}(t, u) = P_{ij}(t)P_{jk}(u)$$

A test for this is the Chapman-Kolomogorov equation, namely

$$P_{ik}(t + u) = \sum_j P_{ij}(t)P_{jk}(u)$$

Using (12), the corresponding relation of generating functions is

$$[1 + (x - 1)g(t + u)]^i \exp b(x - 1)[1 - g(t + u)]$$
$$= [1 + (x - 1)g(t)g(u)]^i \exp b(x - 1)[1 - g(t)g(u)];$$

so the process is Markovian only if

$$g(t + u) = g(t)g(u)$$

which is true only for exponential holding time.

For the next transition probability $P_{ijk}(t, u)$, consider first the condition in which no call arrives in the whole interval $t + u$. As before

$$P_{ij}(t) = \binom{i}{j} g_t^j (1 - g_t)^{i-j}$$

where for convenience $g_t$ is written for $g(t)$. For the next transit, however, there is a difference, namely

$$P_{jk}(u) = \binom{j}{k} \left(\frac{g_{t+u}}{g_t}\right)^k \left(1 - \frac{g_{t+u}}{g_t}\right)^{j-k}$$

since $g_{t+u}/g_t$ is the conditional probability that a call which has lasted $t$ will last $u$ more; $P_{jk}(u)$ is the conditional probability of a transit from $j$ to $k$ in $u$, given the transit $i$ to $j$ in $t$.

The generating function for the double transition probabilities in this case is, then,

$$\sum_j \sum_k P_{ijk}(t, u; 0)x^j y^k = [1 + (x - 1)g_t + x(y - 1)g_{t+u}]^i \tag{16}$$

Now suppose a single call arrives at random in interval $t$. As before, the probability that it will occupy a trunk at time $t$ is $Q(t) = ht^{-1}(1 - g(t))$

and the conditional probability that it will also occupy a trunk at time $t + u$ is

$$\frac{1}{t}\int_0^t f(t + u - x)\,dx \div Q(t)$$

or

$$\frac{g(u) - g(t + u)}{1 - g(t)} = R(t, u), \quad \text{say.}$$

The corresponding generating function, with $x$ and $y$ the indicators of calls at $t$ and $t + u$, resp. is

$$1 - Q(t) + Q(t)[1 - R(t, u)]x + Q(t)R(t, u)xy$$

or

$$1 + (x - 1)(1 - g(t))h/t + x(y - 1)[g(u) - g(t + u)]h/t$$

The generating function for $c$ calls in this interval is this expression raised to the $c$'th power, since calls arrive independently; and since $c$ calls arrive with probability $e^{-at}(at)^c/c!$, the generating function for calls arriving in this interval is

$$\begin{aligned}\sum [1 + (x - 1)Q + x(y - 1)QR]^c e^{-at}(at)^c/c! \\ = \exp b[(x - 1)(1 - g(t)) + x(y - 1)(g(u) - g(t + u))]\end{aligned} \tag{17}$$

For brevity $Q$ and $R$ have been written for $Q(t)$ and $R(t, u)$.

Finally the generating function for calls arriving in $t$, $t + u$, is

$$\exp b(y - 1)(1 - g(u)) \tag{18}$$

Hence

$$\begin{aligned}\sum_j \sum_k P_{ijk}(t, u)x^j y^k = [1 + (x - 1)g(t) + x(y - 1)g(t + u)]^i \\ \cdot \exp b[(x - 1)(1 - g(t)) + (y - 1)(1 - g(u)) \\ + x(y - 1)(g(u) - g(t + u))]\end{aligned} \tag{19}$$

By similar argument, the generating function for triple transition probabilities is

$$\begin{aligned}\sum_j \sum_k \sum_t P_{ijkt}(t, u, v)x^i y^k z^t \\ = [1 + (x - 1)g_t + x(y - 1)g_{t+u} + xy(z - 1)g_{t+u+v}]^i \\ \cdot \exp b((x - 1)(1 - g_t) + (y - 1)(1 - g_u) + \\ (z - 1)(1 - g_v) + x(y - 1)(g_u - g_{t+u}) + \\ y(z - 1)(g_v - g_{u+v}) + xy(z - 1)(g_{u+v} - g_{t+u+v}))\end{aligned} \tag{20}$$

### 3. Expected Correlations

Correlation expectations, like $E[N(t)N(u)]$ in equation (3), are needed for evaluation of the moments of $M(T)$. They may be determined from the transition probability generating functions, if it is agreed, as a matter only of convenience, that the time epochs $t$, $u$, $v$, etc. are in that order ($t \leq u \leq v \leq \cdots$). Since, on the assumption of statistical equilibrium, the call probabilities at the first epoch $t$, are independent of its value, as already noticed, this value may be taken as zero without loss of generality.

Thus for the second moment it is sufficient to determine

$$\varphi(u) = E[N(0)N(u)] = \sum ip_i \sum jP_{ij}(u) \tag{21}$$

with $p_i = Pr[N(0) = i] = e^{-b}b^i/i!$

Write

$$G_u(x, y) = \sum p_i x^i \sum P_{ij}(u)y^j$$

By (12), this is the same as

$$G_u(x, y) = \exp b[x - 1 + y - 1 + (x - 1)(y - 1)g(u)]$$

or

$$H_u(x, y) = G_u(x + 1, y + 1) \equiv \exp b(x + y + xyg(u))$$

and

$$\begin{aligned} \varphi(u) &= \frac{\partial^2 H}{\partial x \partial y}\bigg|_{x,y=0} \\ &= b^2 + bg(u) \end{aligned} \tag{22}$$

In the same way the second order correlation expectation, that is

$$\varphi(u, v) = E[N(0)N(u)N(u + v)],$$

is obtained from

$$G_{u,v}(x, y, z) = \sum p_i x^i \sum \sum P_{ijk}(u, v)y^j z^k$$

and

$$\begin{aligned} H_{u,v}(x, y, z) &= G_{u,v}(x + 1, y + 1, z + 1) \\ &= \exp b(x + y + z + xyg(u) + yzg(v) + x(y + 1)zg(u + v)) \end{aligned}$$

Hence

$$\varphi(u, v) = b^3 + b^2[g(u) + g(v) + g(u + v)] + bg(u + v) \tag{23}$$

Finally, the third order correlation turns out to be

$$\begin{aligned}\varphi(u, v, w) &= E[N(0)N(u)N(u+v)N(u+v+w)] \\ &= b^4 + b^3[g(u) + g(v) + g(w) \\ &\quad + g(u+v) + g(v+w) + g(u+v+w)] \\ &\quad + b^2[g(u+v) + g(v+w) + 2g(u+v+w)] \\ &\quad + b^2[g(u)g(w) + g(u+v)g(v+w) \\ &\quad + g(v)g(u+v+w)] + bg(u+v+w)\end{aligned} \tag{24}$$

As will appear, the arrangement of terms in (22), (23) and (24) corresponds to the expansion of ordinary moments in terms of cumulants (semi-invariants); e.g. (24) corresponds to $m_4 = b^4 + 6b^2k_2 + 4bk_3 + 3k_2^2 + k_4$ with $k_i$ the $i$'th cumulant (for the Poisson of mean $b$, $k_i = b$).

## 4. Moments

Moments are obtained from these results by integrations. As already noted, equation (2), the first moment is $b$ for any holding time distribution.

Since there are two ways of ordering the epochs $t$, $u$, the second moment is

$$\begin{aligned}E[M^2(T)] &= \frac{2}{T^2}\int_0^T dt \int_0^t du\, \varphi(t-u) \\ &= b^2 + \frac{2b}{T^2}\int_0^T dt \int_0^t du\, g(t-u) \\ &= b^2 + \frac{2b}{T^2}\int_0^T dx\, g(x)(T-x)\end{aligned} \tag{25}$$

The last step is by the formula for reversing the order of integration indicated by

$$\int_0^T dt \int_0^t du = \int_0^T du \int_u^T dt$$

The variance or second central moment, which is also the second cumulant $k_2$, is then

$$\begin{aligned}\text{Var}\,[M(T)] &= E[(M(T) - b)^2] \\ &= E[M^2(T)] - b^2 \\ &= \frac{2b}{T^2}\int_0^T dx\, g(x)(T-x)\end{aligned} \tag{26}$$

Since there are $3! = 6$ ways of ordering 3 epochs, the third moment may be written

$$E[M^3(T)] = \frac{6}{T^3}\int_0^T dt \int_0^t du \int_0^u dv\, \varphi(t-u, u-v)$$

$$= b^3 + \frac{6b^2}{T^3}\int_0^T dt \int_0^t du \int_0^u dv[g(t-w) + g(u-v) + g(t-v)]$$

$$+ \frac{6b}{T^3}\int_0^T dt \int_0^t du \int_0^u dv\, g(t-v)$$

Here the first triple integral is immediately evaluated by use of the identity

$$2\int_0^T dt \int_0^t du \int_0^u dv[g(t-u) + g(u-v) + g(t-v)]$$

$$= \int_0^T \int_0^T \int_0^T dt\, du\, dv\, g(|\, t-v\,|)$$

$$= 2T\int_0^T dx\, g(x)(T-x)$$

$$= T^3 k_2/b$$

The last triple integral, by successive inversions of integration order, turns out to be

$$\frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x$$

Hence finally

$$E[M^3(T)] = b^3 + 3bk_2 + \frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x \tag{27}$$

and

$$\begin{aligned} k_3 &= E[(M(T) - b)^3] \\ &= E[M^3(T)] - 3b\,E[M^2(T)] + 2b^3 \\ &= E[M^3(T)] - 3b\,k_2 - b^3 \\ &= \frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x \end{aligned} \tag{28}$$

The fourth moment is given by

$$E[M^4(T)] = \frac{24}{T^4}\int_0^T dt \int_0^t du \int_0^u dv \int_0^v dw\, \varphi(t-u, u-v, v-w)$$

$$= b^4$$

$$+ \frac{24}{T^4}\Big\{b^3 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw[g(t-u) + g(t-v)$$

$$+ g(t-w) + g(u-v) + g(u-w) + g(v-w)]$$

$$+ b^2 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw\,[g(t-v) + g(u-w) + 2g(t-w)]$$

$$+ b^2 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw[g(t-u)g(v-w) +$$

$$g(t-v)g(u-w) + g(t-w)g(u-v)]$$

$$+ b \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw[g(t-w)]\Big\}$$

Employing the identities

$$4\int_0^T \int_0^T \int_0^T \int_0^T dt\, du\, dv\, dw\,[g(t-u) + g(t-v) + g(t-w)$$

$$+ g(u-v) + g(u-w) + g(v-w)]$$

$$= \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw\; g(|\, t-u\,|)$$

$$= 2T^2 \int_0^T dx\; g(x)(T-x) = T^4 k_2/b,$$

$$8\int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw\,[g(t-u)\, g(v-w) + g(t-v)\, g(u-w)$$

$$+ g(t-w)\, g(u-v)]$$

$$= \int_0^T \int_0^T \int_0^T \int_0^T dt\, dw\, dv\, dw\; g(|\, t-u\,|)g(|\, v-w\,|)$$

$$= 4\left[\int_0^T dx\; g(x)(T-x)\right]^2 = T^4 k_2^2/b^2,$$

and successive inversion of order of integration, the final result turns out to be

$$E[M^4(T)] = b^4 + 6b^2k_2 + 4bk_3 + 3k_2^2 + \frac{12b}{T^4}\int_0^T dx\, g(x)(T-x)x^2 \tag{29}$$

and

$$\begin{aligned} k_4 &= E[(M(T) - b)^4] - 3E[(M(T) - b)^2] \\ &= \frac{12b}{T^4}\int_0^T dx\, g(x)(T-x)x^2 \end{aligned} \tag{30}$$

It is a tempting surmise that

$$k_n = b\,\frac{n(n-1)}{T^n}\int_0^T dx\, g(x)(T-x)\,x^{n-2}$$

but this has not been proved. Note that for $g(x) = 1$, $k_n = b$, the cumulant of the Poisson, as it should.

For the two cases of chief interest, constant and exponential holding times, the function $g(x)$, *in average holding time units* (that is, $x = t/h$) is given by

c.h.t. $g(x) = 1 - x \qquad x < 1$
$= 0 \qquad x > 1$

e.h.t. $g(x) = e^{-x}$

and the results are as follows:

| Cumulant | Constant Holding Time T < 1 | T > 1 |
|---|---|---|
| $k_2$ | $b(1 - T/3)$ | $bT^{-1}(1 - 1/3T)$ |
| $k_3$ | $b(1 - T/2)$ | $bT^{-2}(1 - 1/2T)$ |
| $k_4$ | $b(1 - 3T/5)$ | $bT^{-3}(1 - 3/5T)$ |
| | Exponential Holding Time | |
| $k_2$ | $2bT^{-2}[T - 1 + e^{-T}]$ | |
| $k_3$ | $6bT^{-3}[T - 2 + (T + 2)e^{-T}]$ | |
| $k_4$ | $12bT^{-4}[2T - 6 + (T^2 + 4T + 6)e^{-T}]$ | |

It may be worth noting that, if the surmise is correct, for constant holding time

$$\begin{aligned} k_n &= b\left[1 - \frac{n-1}{n+1}T\right] \qquad T < 1 \\ &= \frac{b}{T^{n-1}}\left[1 - \frac{n-1}{n+1}\frac{1}{T}\right] \qquad T > 1 \end{aligned}$$

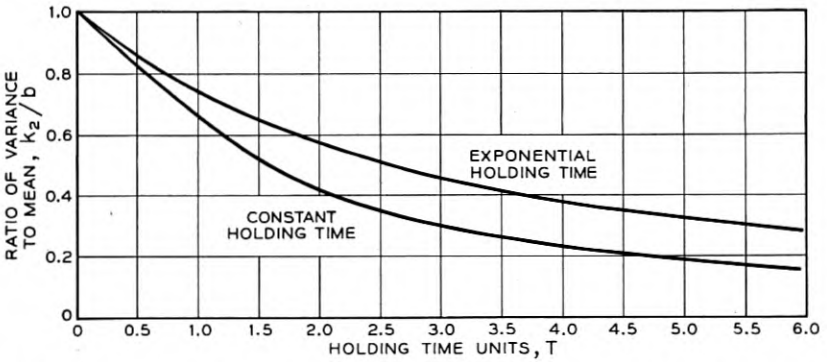



FIG. 1.—Comparison of variances of average traffic for constant and exponential holding times.

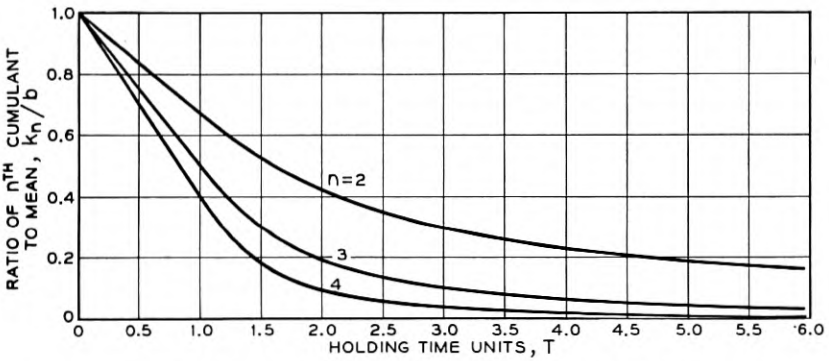



FIG. 2.—Cumulants $k_2$, $k_3$, and $k_4$ for constant holding time.

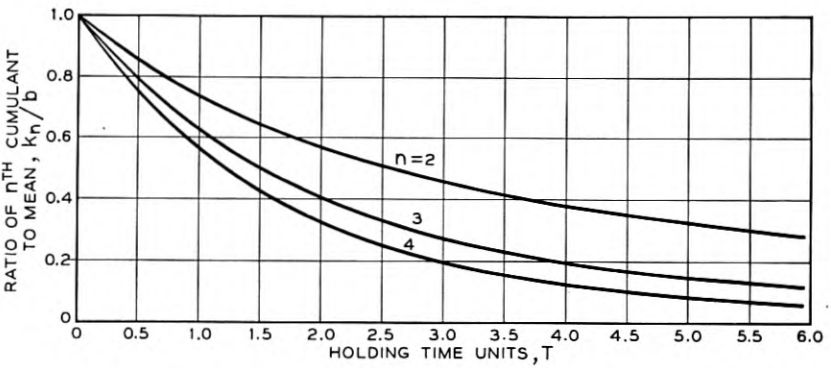



FIG. 3.—Cumulants $k_2$, $k_3$ and $k_4$ for exponential holding time.

and for exponential holding time

$$k_n = b\,\frac{n(n-1)}{T^n}\,[(n-2)!\,T - (n-1)! + e^{-T}\,(T+\alpha)^{n-2}]$$

where in the last term $(T+\alpha)^{n-2}$ is a symbolic expression or shorthand for

$$(T+\alpha)^{n-2} = \sum_0^{n-2} \binom{n-2}{m} T^{n-2-m}\,\alpha_m$$

and $\alpha_m = (m+1)!$; e.g.

$$(T+\alpha)^3 = T^3 + 6T^2 + 18T + 24$$

For small values of $T$, the two cases coalesce ($e^{-x} \approx 1 - x$) and at $T = 0$ approach $b$ as they should. For large values of $T$, and constant holding time,

$$k_n \sim b/T^{n-1}, \qquad (n = 2, 3, 4);$$

for exponential holding time

$$k_n \sim n!b/T^{n-1}, \qquad (n = 2, 3, 4).$$

For $n = 2$, these results agree with Rabe [6].

As $T$ increases, for either holding time, the cumulants are progressively smaller, and the approximation of the distribution of $M(T)$ by a normal curve (which has all cumulants, except the first and second, zero) improves. This is what follows from the central limit theorem if the subdivision of $T$ into a large number of intervals results in mutually independent random variables (cf. Rice [7] 3.9).

Figure 1 shows a comparison of the variances ($k_2$) for the two holding time cases. Figure 2 shows a comparison of the cumulants $k_2$, $k_3$ and $k_4$ for constant holding time, and Fig. 3 shows the same thing for exponential holding time.

## 5. Finite Sources—Exponential Holding Time

The generating function for transitional probabilities for $N$ subscribers, each originating calls independently with probability $\lambda$, and for exponential holding time, as given by Jensen (l.c.) is as follows:

$$P_i(t, x) = [1 + q_1(x-1)]^i[1 + q_0(x-1)]^{N-i} \qquad (31)$$

with

$$q_0 = p - pe^{-(\lambda+\gamma)t}$$
$$q_1 = p + q \quad \text{“}$$
$$p = 1 - q = \lambda/(\lambda+\gamma)$$
$$\gamma = 1/h$$

It should be noticed that for $t = \infty$, $q_0 = q_1 = p$ and

$$P_i(\infty, x) = [1 + p(x - 1)]^N \tag{32}$$

The right hand side is the binomial generating function and, as independent of $i$, is the generating function for the statistical equilibrium probabilities; that is

$$Pr\,[N(t) = k] = \binom{N}{k} p^k q^{N-k}$$

Also the process is Markovian since

$$\sum_k x^k \sum_j P_{ij}(t)P_{jk}(u) = \sum_j P_{ij}(t)[1 + q_{1u}(x - 1)]^j [1 + q_{0u}(x - 1)]^{N-j}$$

$$= [1 + (q_{0u} + q_{1t}q_{1u} - q_{1t}q_{0u})(x - 1)]^i$$

$$[1 + (q_{0u} + q_{0t}q_{1u} - q_{0t}q_{0u})(x - 1)]^{N-i}$$

and

$$q_{0u} + q_{1t}q_{1u} - q_{1t}q_{0u} = q_{1,t+u}$$

$$q_{0u} + q_{0t}q_{1u} - q_{0t}q_{0u} = q_{0,t+u}$$

Here it has been convenient to indicate by the double subscript the dependence of $q_0$ and $q_1$ on a time variable.

Moments are obtained by the process given in detail for the infinite source case. For brevity it is convenient to use the binomial cumulants which are as follows

$$\kappa_2 = Npq$$

$$\kappa_3 = Npq(q - p)$$

$$\kappa_4 = Npq(1 - 6pq)$$

and the modified time variable $T_1 = (\lambda + \gamma)T$. Then the results are

$$k_2 = 2T_1^{-2}\kappa_2[T_1 - 1 + e^{-T_1}]$$

$$k_3 = 6T_1^{-3}\kappa_3[T_1 - 2 + (T_1 + 2)e^{-T_1}]$$

$$k_4 = 12T_1^{-4}((\kappa_4 + \kappa_2^2 N^{-1})[2T_1 - 6 + (T_1^2 + 4T_1 + 6)e^{-T_1}]$$

$$- \kappa_2^2 N^{-1}[1 - (T_1^2 + 2)e^{-T_1} + e^{-2T_1}])$$

These of course bear a strong resemblance to the infinite source case (exponential holding time), to which they converge.

## Bibliography

1. W. Feller, "On the theory of stochastic processes with particular reference to applications," *Proc. Berkeley Symposium on Math. Statistics and Probability*, Univ. of California Press, 1949.
2. W. Feller, "An introduction to probability theory and its applications," New York, 1950.
3. A. Jensen, An elucidation of Erlang's statistical works through the theory of stochastic processes, "The Life and Works of A. K. Erlang," Copenhagen, 1948.
4. W. F. Newland, A method of approach and solution to some fundamental traffic problems, *P.O.E.E. Jl.*, *25*, 119–131 (1932–3).
5. Conny Palm, "Intensitätsschwangungen in fernsprechverkehr," *Ericsson Technics*, *44*, 1–189 (1943).
6. F. W. Rabe, "Variations of telephone traffic," *Elec. Comm. 26*, 243–248 (1949).
7. S. O. Rice, "Mathematical analysis of random noise," *Bell System Technical Journal*, *23*, 282–332 (1944); *24*, 46–156 (1945).

# The Reproduction of Magnetically Recorded Signals

R. L. WALLACE, JR.

(*Manuscript Received July 9, 1951*)

For certain speech studies at the Bell Telephone Laboratories, it has been necessary to design some rather specialized magnetic recording equipment. In connection with this work, it has been found experimentally and theoretically that introducing a spacing of $d$ inches between the reproducing head and the recording medium decreases the reproduced voltage by $54.6\ (d/\lambda)$ decibels when the recorded wavelength is $\lambda$ inches. For short wavelengths this loss is many decibels even when the effective spacing is only a few ten-thousandths of an inch. On this basis it is argued that imperfect magnetic contact between reproducing head and recording medium may account for much of the high-frequency loss which is experimentally observed.

## Introduction

WITHIN the last few years there has been increasing use of magnetic recording in various telephone research applications (examples are various versions of the sound spectrograph used in studies of speech and noise). Some of these uses[1] have required a reproducing head spaced slightly out of contact with the recording medium. Experimental studies were made to determine the effect of such spacing and the results were found to be expressible in an unexpectedly simple form. The general equation derived is believed to be fundamental to the recording problem and to account for much of the high-frequency loss that is found in both in- and out-of-contact systems.

This paper discusses results of the experimental study and presents for comparison some theoretical calculations based on an idealized model.

## Measurements of Spacing Loss

In order to measure the effect of spacing between the reproducing head and the medium, an experiment was set up as indicated in Fig. 1. The recording medium used was a 0.0003 inch plating of cobalt-nickel alloy[2] on the flat surface of a brass disc approximately 13 inches in diameter by $\frac{1}{4}$ inch thick.

This disc was made with considerable care to insure that the recording surface was as nearly plane and smooth as possible and that it would turn reasonably true in its bearings. Speeds of 25 and 78 rpm were provided.

[1] R. C. Mathes, A. C. Norwine, and K. H. Davis, "Cathode-Ray Sound Spectroscope," *Jl. Acous. Soc. Am.*, *21*, 527 (1949).

[2] Plating was done by the Brush Development Company.

The ring-type record-reproduce head shown in Fig. 1 was lapped slightly to obtain a reasonably good fit with the surface of the disc.

A single-frequency recording was made with the head in contact with the disc using a-c. bias in the usual way. Then the open circuit reproduced signal level was measured, first with the head in contact, and then after introducing paper shims of various thickness between the reproducing head and the medium. Thus the effect of spacing was measured at a particular frequency and recording speed. The signal was then erased and the process was repeated for other recorded frequencies and for several record-reproduce speeds. Measurements were also made for cases in which the recording and reproducing speeds were different. Considerable care was required to keep the disc and head sufficiently clean so that reproducible results could be obtained.

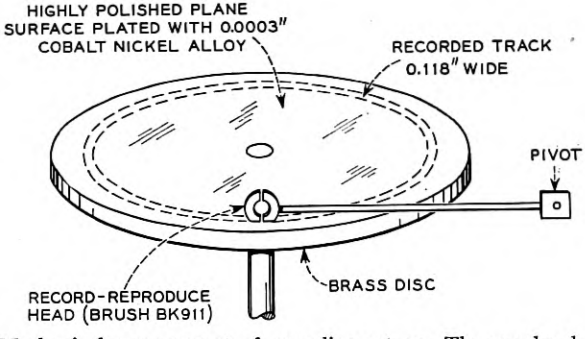



Fig. 1—Mechanical arrangement of recording set up. The one head served for recording, playback, and erase.

Figure 2 shows typical response curves measured at 21 in./sec. with the reproducing head in contact and with 0.004 inch spacing. The difference between these two curves will be called the spacing loss corresponding to this spacing and speed. From these data and more of the same sort it is found that, within experimental error, spacing loss can be very simply related to spacing and the recorded wavelength, $\lambda$, by the empirical equation,

$$\text{Spacing loss} = 55(d/\lambda) \text{ decibels} \tag{1}$$

where spacing loss is the number of decibels by which the reproduced level is decreased when a spacing of $d$ inches is introduced between the reproducing head and a magnetic medium on which a signal of wavelength $\lambda$ inches has been recorded.

The fact that this expression fits the experimental data reasonably well is indicated in Fig. 3 where spacing loss data measured at a number of different speeds, frequencies, and spacings are plotted against $d/\lambda$.

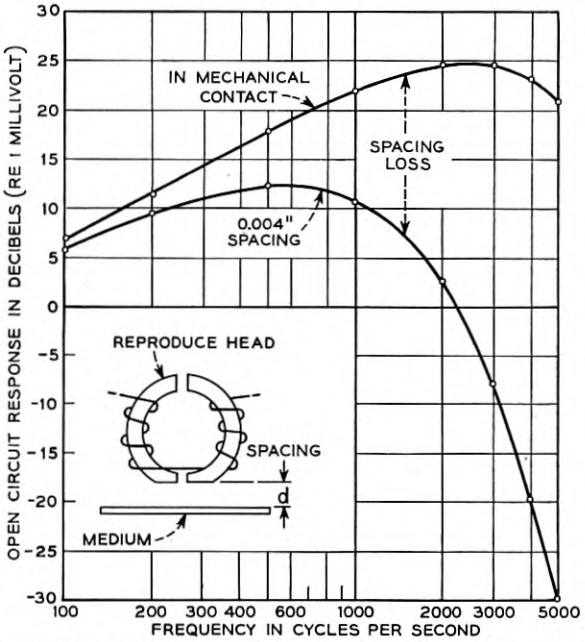



Fig. 2—Response curves taken at 21 in./sec. Recordings were made with head in contact and were played back first with head in contact and then with a spacing of 4 mils between head and disc.

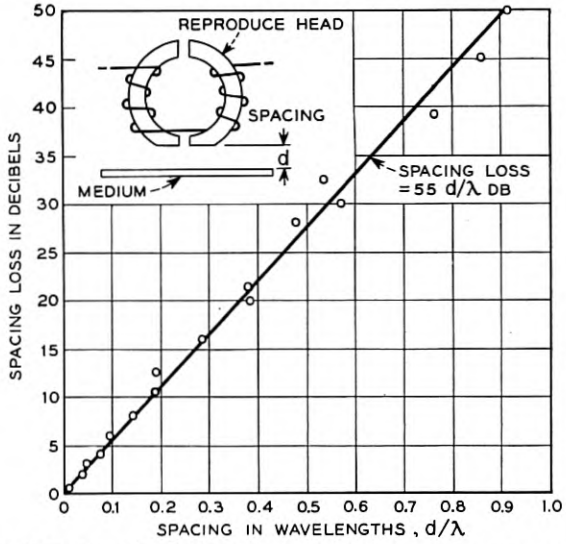



Fig. 3—Data obtained as in Fig. 2 show spacing loss approximately equal to $55(d/\lambda)$ decibels.

## Implications of the Experiment

In this section it will be assumed that equation (1) holds true in all cases where the spacing $d$ is sufficiently small and the recorded track is sufficiently wide so that end effects are negligible. If this is true, as it seems experimentally to be, then it is indeed surprising how great can be the effect of even a very small spacing when the recorded wavelength is small. For example, take the case of a 7500 cps signal recorded at 7.5 in./sec. in which case the wavelength is 0.001 inch. A particle of dust which separated the tape from the reproducing head by one-thousandth of an inch would decrease the reproduced level by 55 db. A spacing of 0.0001 inch would produce a quite noticeable 5.5 db effect and even at 0.00001 inch spacing the 0.55 db loss would be measurable in a carefully controlled experiment.

In view of the magnitudes involved, it seems probable that this spacing loss may play a significant role even in cases where the reproducing head is supposed to be in contact with the medium. For example, it has been known for some time that chattering of the tape on the reproducing head or changes in the degree of contact due to imperfect smoothness of the tape can result in amplitude modulation of the reproduced signal and thereby give rise to "modulation noise" or "noise behind the signal."

With the aid of equation (1) it is possible to estimate the magnitude of the noise provided some assumption is made about the wave form of the modulation. To take a simple case, suppose that the roughness of the tape were such as to sinusoidally modulate the spacing by a very small amount and at a low frequency. The reproduced signal would then be modulated and would contain a sideband on each side of the center frequency. The energy in these two sidebands constitutes the modulation noise in this case. If it is required that this noise be 40 db down on the signal, then one can calculate the maximum permissible excursion of the tape away from the reproducing head. This turns out to be $1.1(10)^{-5}$ cm. or about one-sixth of the wavelength of the red cadmium line! Of course, the one mil wavelength assumed in this example is about as short as is often used and the effect becomes less severe as the wavelength is increased. This is one of the reasons that speeds greater than 7.5 in./sec. are used for highest quality reproduction.

One can also make some rough qualitative inferences about the effect of the thickness of the recording medium on the shape of the response curve. As can be seen from equation (1) or from Fig. 2, low frequencies can be reproduced with very little loss in amplitude in spite of considerable spacing between the reproducing head and the medium while high frequencies (i.e. short wavelengths) may be appreciably attenuated by even 0.0001 inch

spacing between the head and the medium. With this in mind it is easy to see that at high frequencies only a thin layer of the medium nearest the reproducing head will contribute to the reproduced signal. In this case (short $\lambda$) increasing the thickness of the medium beyond a certain amount can have no effect on the reproduced level simply because the added part of the tape is too far from the head to make its effect felt. Consider the effect of increasing the thickness of the medium from 0.3 mil to 0.6 mil when the wavelength is one mil. Since the spacing loss for 0.3 mil spacing at $\lambda = 1$ mil is 16.5 db, the signal contributed by the lower half of a medium 0.6 mils thick cannot be less than 16.5 db lower than that contributed by the upper half and hence the increase in thickness can do no more than to raise the reproduced level by 1.2 db.

At a lower frequency for which $\lambda = 100$ mil, however, the corresponding spacing loss is only 0.165 db and in this case the two halves of the tape can contribute almost equally with the result that doubling the thickness of the medium can almost double the reproduced signal voltage.

Qualitatively, then, one might expect that increasing the thickness of the recording medium, other things being equal, would increase the response to low frequencies and leave the high frequency response relatively unaltered. This is in agreement with data published by Kornei.[3]

The estimates of magnitudes just given rest on assumptions which cannot be proved except by further experiments. It has been implicitly assumed, for example, that the medium is uniformly magnetized throughout its thickness and this may not be the case. It does seem perfectly safe, however, to conclude that at a wavelength of one mil that part of the medium which lies deeper than about 0.3 mil from the surface cannot contribute appreciably to the reproduced signal. Furthermore, as the wavelength is decreased beyond this point the thickness of the effective part of the tape decreases in inverse proportion to $\lambda$ with the result that the available flux also decreases. For this reason the "ideal" response curve cannot continue indefinitely to rise at 6 db per octave as it does at low frequencies. In fact, when the effective part of the tape becomes thin enough, the available flux will decrease at 6 db per octave and just cancel the usual 6 db per octave rise, giving an "ideal" response curve which rises 6 db per octave at low frequencies but which eventually becomes flat, neither rising nor falling with further increase in frequency.

Spacing loss may contribute in still another way to the frequency response characteristic of a magnetic recording system in which the reproducing head makes contact with the medium. It is well known to those who work

[3] Otto Kornei, "Frequency Response of Magnetic Recording," *Electronics*, p. 124, August, 1947.

with magnetic structures such as are used in transformers and the like that intimate mechanical contact between two parts of a magnetic circuit does not imply intimate magnetic contact. In fact, even when great care is taken in fitting such parts together, measurements invariably show an effective air gap between them and the effective width of this gap usually amounts to appreciably more than one mil. One reason for this is that the permeability of soft materials such as are used in the cores of transformers and reproducing heads is very sensitive to strain. Even the light cold working which a surface receives in being ground flat is sufficient to impair very seriously the permeability of a thin surface layer.

In view of this it is to be expected that the magnetic contact between reproducing head and medium is less than perfect. If cold working during the fabrication of the head or due to abrasion by the recording medium should result in an effective air space between head and medium amounting to as much as one mil, the effect on frequency response would be pronounced indeed. At a recording speed of 7.5 in./sec. this amount of spacing would cause a loss of 7.3 db at 1000 cps, 14.6 db at 2000 cps, 21.9 db at 3000 cps, 29.2 db at 4000 cps, etc.

It seems certain that in a practical recording system some loss of this sort must occur. The problem of determining the magnitude of the loss or in other words the amount of the effective spacing in a practical case is, however, a difficult one. So far, no direct experimental method for its determination has been found.

## Theoretical Calculations for an Idealized Case

In the preceding section an experimentally determined spacing loss function has been discussed. It was shown that as the reproducing head is moved away from the recording medium the reproduced signal level decreases. This means that the magnetic flux through the head decreases. If the distribution of magnetization in the recording medium were known, it should be possible to compute the flux through the head and thereby to derive the spacing loss function on a theoretical basis. Unfortunately it seems almost impossible to do this calculation in an exact way because very little is known about the magnetization pattern in the medium and because the geometry of the usual ring type head makes the boundary value problem an exceedingly difficult one to solve.

It is possible, however, to obtain a solution for an idealized case which bears at least some resemblance to the practical situation and this solution will be presented. The results must, of course, be viewed with due skepticism until they can be proved experimentally or else recalculated on the basis of better initial assumptions. It is hoped, however, that in some

measure they may serve as a guide to a better understanding of the magnetic reproducing process.

## The Idealized Recording Medium

The problem will be reduced to two dimensions by assuming an infinitely wide and infinitely long tape of finite thickness $\delta$. A rectangular coordinate system will be chosen in such a way that the central plane midway between the upper and lower surfaces of the recording medium lies in the $x$–$y$ plane. It will be assumed that the medium is sinusoidally magnetized in such a way that in the medium the intensity of magnetization is given by

$$I_x = I_m \sin (2\pi x/\lambda)$$
$$I_y = I_z = 0. \tag{2}$$

Equations (2) say that the recording is purely longitudinal. In a practical case, of course, the recorded signal is neither purely longitudinal nor purely perpendicular but rather contains components of both sorts. In Appendix I it is shown that the frequency response does not depend on the relative amounts of these two components and hence that the computed results are equally valid whether the recorded signal is purely longitudinal, purely perpendicular, or a mixture of the two.

Appendix II contains calculations for the case of a round wire sinusoidally magnetized along its axis, and for a plated wire. These results, though much different in mathematical form, are shown to be very similar to the results for a flat medium.

## The Idealized Reproducing Head

Figure 4 shows a semi-practical version of the sort of idealized reproducing head which will be treated.

It consists of a bar of core material with a single turn of exceedingly fine wire around it. This head is imagined to be spaced $d$ inches above the surface of the recording medium. If the dimensions of the bar are made large enough, the amount of flux through it will obviously be as great as could be made to pass through any sort of head which makes contact with only one side of the tape and so the open circuit reproduced voltage per turn is as high as can be obtained with any practical head.

Suppose a very narrow gap is introduced in this head where the single turn coil was and that the magnetic circuit is completed by a ring of core material as shown in Fig. 5.

If the permeability of the head is very high and the gap very small then the flux which passed through the single turn coil of Fig. 4 will now pass

through the ring of Fig. 5 and can be made to thread through a coil of many turns wound on the ring. In so far as this is true, calculations based on this bar type head are applicable to ring type heads.

If the bar of Fig. 4 is now allowed to become infinite in length, width, and thickness, the flux density in it can be computed and the flux per unit width can be evaluated. This calculation is outlined in Appendix I. If the tape moves past the head with a velocity $v$ in the $x$ direction, the repro-

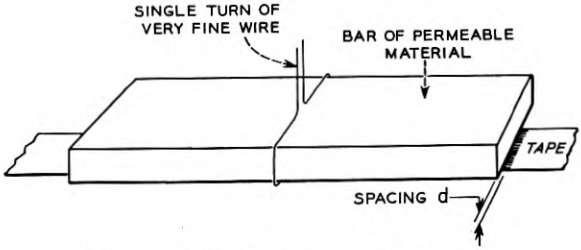



Fig. 4—Idealized bar-type reproducing head.

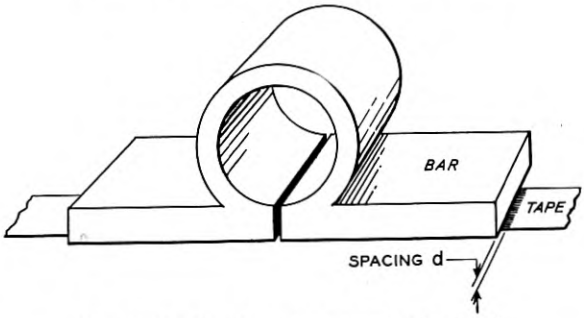



Fig. 5—Idealized ring-type reproducing head.

duced voltage should be proportional to the rate of change of flux. In the appendix this is shown to be

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu+1}\,4\pi W v I_m(1 - e^{-2\pi\delta/\lambda})e^{-2\pi d/\lambda}\cos(\omega t) \qquad (3)$$

where $\frac{d\phi_x}{dt}$ is the rate of change of flux in $W$ cm. width of the reproducing head measured in Maxwells per sec.,

$\mu$ is the permeability of the reproducing head,

$W$ is the width in cm. of the reproducing head (and of the recorded track in a practical case),

$v$ is the velocity in cm./sec. with which the recording medium passes the *reproducing* head.

$I_m$ is the peak value of the sinusoidal intensity of magnetization in the recording medium measured in gauss,

$\delta$ is the thickness of the recording medium measured in the same units as $\lambda$,

$\lambda$ is the recorded wavelength measured in any convenient units,

$d$ is the effective spacing between the reproducing head and the surface of the recording medium measured in the same units as $\lambda$, and

$\omega$ is $2\pi$ times the reproduced frequency in cycles per sec.

Note that equation (3) applies to a ring type head with no back gap. If the head has a back gap then not all the available flux will thread through the ring. Some of it will return to the medium through the scanning gap and hence will not contribute to the reproduced voltage. This does not affect the shape of the frequency response curve but does contribute a constant multiplying factor (less than unity) to the right hand side of equation (3). The value of this factor depends on the reluctances of the gaps and of the magnetic parts of the reproducing head. If the reluctance of the magnetic parts is negligible and the reluctance of the back gap is equal to the reluctance of the front gap then the available flux will divide equally in the two gaps and the factor will be one-half. This factor will not be considered further in this paper because it does not contribute to the shape of the response curve but only to the absolute magnitude of the reproduced voltage. It could be interpreted as reducing the effective number of turns on the reproducing head to a value somewhat lower than the actual number of turns.

## Spacing Loss

The term $e^{-2\pi d/\lambda}$ tells how the reproduced voltage depends on spacing. In order to compare this computed effect with the experimentally observed one it is necessary to put it in decibel form by computing twenty times the $\log_{10}$ of $e^{-2\pi d/\lambda}$. This gives

$$\text{Spacing Loss} = 54.6\ (d/\lambda)\ \text{decibels}\ .$$

This agrees very well indeed with the experimentally determined equation (1) in which the constant is 55 instead of the computed 54.6. The computed spacing loss function is plotted in Fig. 6.

### Thickness Loss

The effect of the thickness of the recording medium shows up in the term $(1 - e^{-2\pi\delta/\lambda})$. At low frequencies for which the wavelength is much greater than the thickness of the medium this reduces to $2\pi\delta/\lambda$. In this case the reproduced voltage is proportional to the thickness of the medium and to frequency. This is the familiar six db per octave characteristic.

At high frequencies, however, when $\lambda \ll \delta$ the term reduces to unity

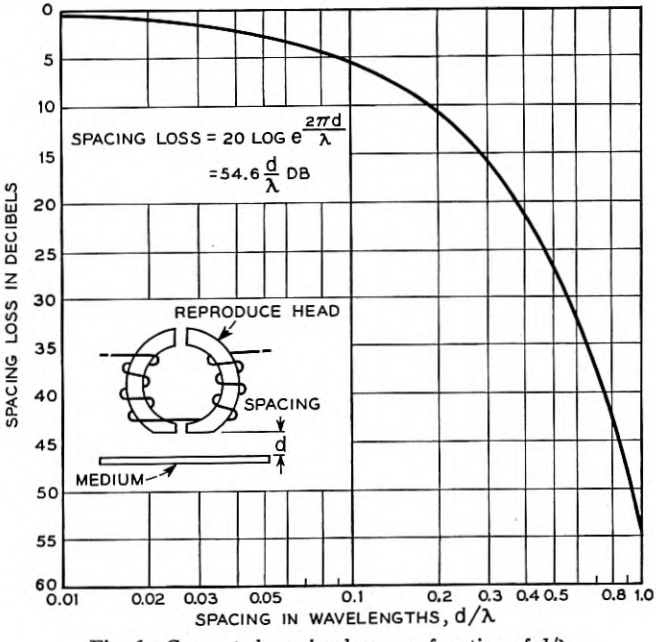



Fig. 6—Computed spacing loss as a function of $d/\lambda$.

and the computed "ideal" response is flat with frequency and independent of the thickness of the medium.

If the term $(1 - e^{-2\pi\delta/\lambda})$ is rewritten as

$$(2\pi\delta/\lambda)\left[\frac{1 - e^{-2\pi\delta/\lambda}}{2\pi\delta/\lambda}\right]$$

then the part in parenthesis accounts for a 6 db per octave characteristic and the part in brackets accounts for a loss *with respect to this 6 db per octave characteristic.* This loss, which will be called Thickness Loss[4], is given by

[4] It seems somewhat awkward to speak of "Thickness Loss" when nothing is actually lost by making the medium thick. The only excuse for this way of splitting the terms is that it makes for ease in comparing measured and computed curves.

$$\text{Thickness Loss} = 20 \log_{10} \frac{2\pi\delta/\lambda}{1 - e^{-2\pi\delta/\lambda}} \text{ db} \tag{5}$$

where $\lambda$ is the recorded wavelength and $\delta$ is the thickness of the recording medium. This function is plotted in Fig. 7.

## Comparison with Experiment

The most elementary consideration of the magnetic recording process indicates that when the recording signal current is held constant the open circuit reproduced voltage should be a function of frequency, increasing by 6 db for each octave increase in frequency. Experimental response curves tend to show this 6 db per octave characteristic when the recorded wavelength is moderately long and the frequency moderately low. This makes it possible to draw a 6 db per octave line on the measured response characteristic in such a way as to coincide with the low-frequency part of the measured response characteristic. As the frequency is increased the measured curve tends to fall more and more below the 6 db per octave line. This is because several kinds of loss come into play as the wavelength decreases or as the frequency increases. Among these losses are:

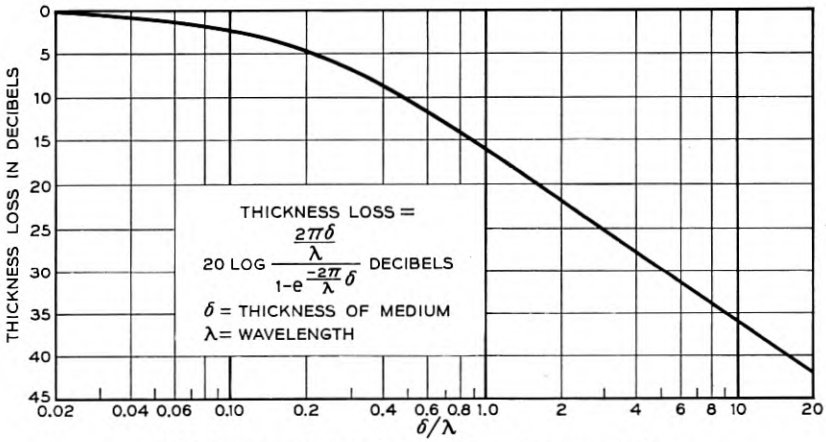



Fig. 7—Computed thickness loss as a function of $\delta/\lambda$.

1. Self demagnetization,
2. Eddy current and other losses in the recording and reproducing heads, and
3. Gap loss due to the finite scanning slit in the reproducing head.

The work presented in the first sections of this paper indicates that the

following two kinds of loss should be added to this list:

4. Spacing loss due to imperfect magnetic contact between the reproducing head and the recording medium, and
5. Thickness loss.

Of these five losses three can be evaluated quantitatively either by direct measurement or by calculation from theory. The remaining two are self-demagnetization and spacing loss.

In this section the known losses will be evaluated for a particular recording system. This leads to a response curve which can be compared with the measured curve. The difference between the two curves should be due to self-demagnetization and to spacing loss provided the above list of losses is complete.

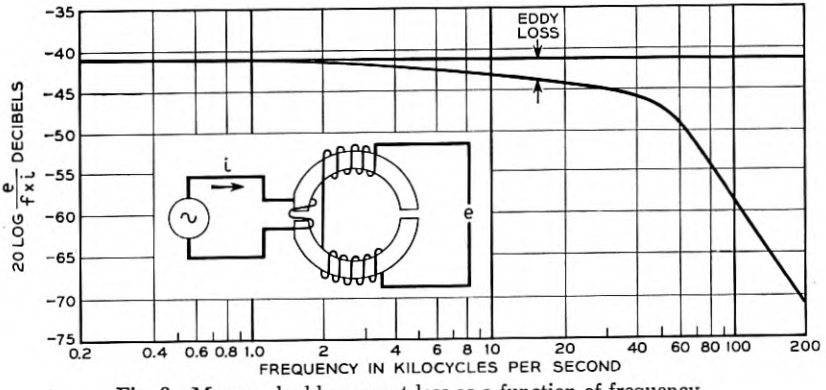



Fig. 8—Measured eddy current loss as a function of frequency.

The recording system used is the one shown in Fig. 1 with the speed set at 15.5 in./sec. for both recording and reproducing. A constant signal current of 0.1 ma was used for recording with the 55 kc bias adjusted to give maximum open circuit reproduced voltage.

Eddy current losses were measured as indicated in Fig. 8 by sending a measured constant current $i$ through a small auxiliary winding around the pole tip and measuring the open circuit voltage developed across the normal winding of the head. Any departure of this measured voltage from a 6 db per octave increase with increasing frequency is due to losses in the head which will be loosely called eddy current losses. Other kinds of loss may enter into this measurement (as, for example, loss due to the self-capacitance of the winding) but in the frequency range of interest, eddy losses predominate.

By a completely different sort of measurement,[5] J. R. Anderson has arrived at a similar value for eddy current loss in this type of head and has shown that approximately the same loss occurs in both the recording and the reproducing process. For this reason it seems proper to assume that eddy currents account for just twice the loss measured by the method of Fig. 8.

The loss due to the finite gap in the reproducing head is computed from the well known relation.[6]

$$\text{Gap loss} = 20 \log_{10} \frac{\pi g/\lambda}{\sin (\pi g/\lambda)}$$

where $g$ is the effective gap width in inches and $\lambda$ is the recorded wavelength in inches.

Thickness loss is computed from equation (5). It must be remembered that this loss was derived on the assumption of uniform magnetization throughout the thickness of the recording medium. This may be a fairly good approximation to the actual state of affairs for a thin medium such as the one being considered, but obviously if the thickness of the medium is large compared with the width of the recording gap then the recording field will not penetrate uniformly through the medium and the derived thickness loss function will not apply.

The derived equation (3) indicates that at low frequencies the reproduced voltage should be proportional to the thickness of the medium. If the thickness of the medium is increased beyond the limit to which the recording field can penetrate, this will no longer be the case and further increase in thickness will have no effect on the response.

Data presented by Kornei[3] on the cobalt-nickel plating being considered here shows that the low-frequency response is approximately proportional to the thickness of the medium for values of thickness between 0.075 mil and 0.5 mil. This may be taken as an indication of approximately uniform penetration through these thicknesses and hence tends to indicate that the derived thickness loss function should be applicable in the case of the 0.3 mil plating being considered here.

The effects of these losses are shown in Fig. 9 along with measured frequency response data. Consider first the experimentally measured response

[5] In unpublished work, J. R. Anderson of the Bell Telephone Laboratories has made use of the fact that eddy losses depend on frequency while all other magnetic recording losses depend on wavelength. By recording a single frequency and playing back at various speeds he determined the loss on playback. By recording various frequencies with recording speed adjusted to give constant recorded wavelength and using a single playback speed he evaluates the eddy loss in the recording process.

[6] S. J. Begun, "Magnetic Recording," p. 84, Murray Hill Books, Inc., New York.

data shown as circles falling near the lowest curve. Some of the measured points have been omitted to avoid crowding but enough remain to show the trend. At low frequencies these points fall along a line of approximately 6 db per octave.

A straight 6 db per octave line labeled 1 has been drawn through these points and extended as shown in the figure. This line is the base from which the various losses must be subtracted. Curve 2 shows the effect of subtracting the computed thickness loss. When eddy losses and gap loss are

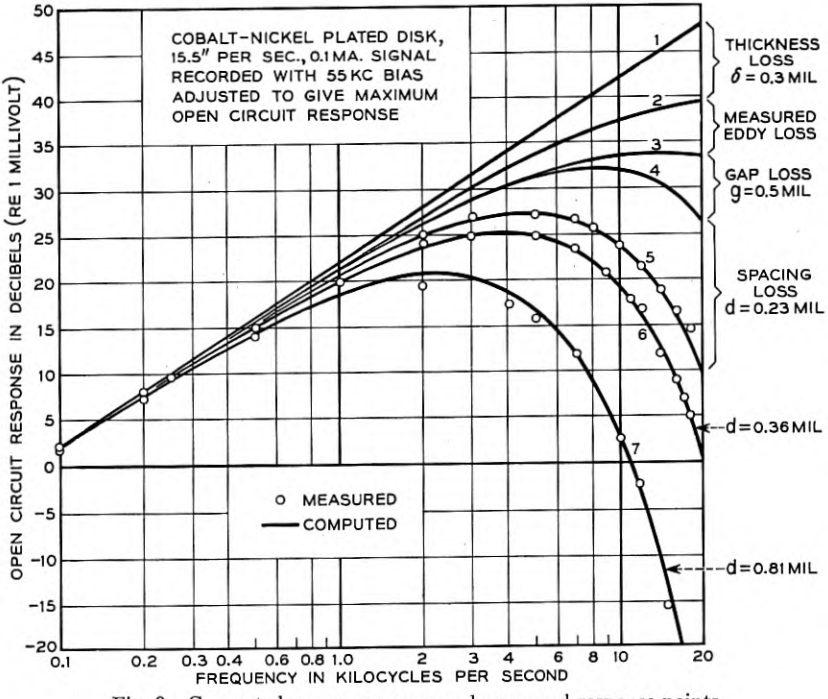



Fig. 9—Computed response curves and measured response points.

also taken into account, curve 4 is obtained. The difference between this curve and the lowest measured response points is presumably due either to self-demagnetization, to spacing loss, or perhaps to both.

There is one clue which may be of help in deciding how much of this loss should be attributed to self-demagnetization and how much to spacing loss. This clue comes from the fact that the form of the spacing loss function is known. Any part of the loss which is due to spacing must follow the equation

$$\text{Spacing Loss} = 54.6\,(d/\lambda)\ \text{db}$$

whereas there is reason to believe that the effects of self-demagnetization cannot possibly account for more than something like ten or fifteen db loss and hence could not follow the equation given above.

In view of this it seems reasonable to try as a first guess the assumption that all the unexplained loss is due to spacing.

If this assumption properly accounts for the shape of the measured response curve there will be at least some reason to suppose it may be correct; particularly so if the required amount of effective spacing seems reasonable.

The lowest solid curve, No. 7 of Fig. 9, has been computed on this basis. That is, a spacing loss corresponding to 0.81 mil effective spacing has been subtracted from curve 4. It is seen that this computed response curve fits reasonably well with the measured points. Furthermore, 0.81 mil effective spacing corresponds to quite reasonably good magnetic contact.

If this interpretation of the measured data is correct then it is obvious that the high-frequency response could be improved a great deal if more intimate magnetic contact between the reproducing head and the recording medium could be achieved. To this end an attempt was made to lap the surface of the head in such a way as to remove material very gently and slowly. After lapping, the response was appreciably improved as indicated by the set of measured points around curve 6. This curve was computed assuming an effective spacing of 0.36 mil. Note that the computed curve now fits the measured points very well indeed.

After still more lapping,[7] the measured response points around curve 5 were obtained. In this case it is necessary to assume only 0.23 mil effective spacing in order to account for the measured curve. Further lapping failed to give further improvement in response but a defect in the head which may account for this has since been found and it is believed that with great care one might actually measure something very close to curve 4.

To summarize, this is what seems to have been found. It is possible to compute a response curve taking into account gap loss, eddy current losses, and thickness loss. If this curve is compared with the final measured response curve it is found that the measured curve gives less high-frequency response than was computed. The difference between the two curves is just the right sort of function of frequency and of just the right magnitude to be accounted for by an effective spacing of 0.00023 inch between the reproducing head and the recording medium. It seems probable that the effective spacing could not have been much smaller than this value and therefore it may be correct to assume that practically all the unexplained

[7] After each lapping it was found that smaller values of bias current sufficed to give maximum reproduced voltage. This is presumably because the improved magnetic contact made the bias current more effective.

In the present case this leads to

$$dH_x = -(4\pi I_m/\lambda)\frac{(x_0 - x)}{(x_0 - x)^2 + (z_0 - z)^2}\cos(2\pi x/\lambda)\,dx\,dz$$
$$dH_z = -(4\pi I_m/\lambda)\frac{(z_0 - z)}{(x_0 - x)^2 + (z_0 - z)^2}\cos(2\pi x/\lambda)\,dx\,dz \tag{8}$$

The total field at $(x_0, z_0)$ is obtained by integrating with respect to $x$ over the range $-\infty$ to $+\infty$ and with respect to $z$ over the range $-\delta/2$ to $+\delta/2$. In carrying out the integration over $x$ it is convenient to make the substitution

$$(x_0 - x)/(z_0 - z) = p$$
$$dx = -(z_0 - z)\,dp \tag{9}$$

Neglecting terms which obviously integrate to zero, this gives

$$H_x = (4\pi I_m/\lambda)\sin(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2}\left[\int_{\infty}^{-\infty}\frac{p\sin[2\pi(z_0 - z)p/\lambda]}{1 + p^2}\,dp\right]dz$$
$$H_z = (4\pi I_m/\lambda)\cos(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2}\left[\int_{\infty}^{-\infty}\frac{\cos[2\pi(z_0 - z)p/\lambda]}{1 + p^2}\,dp\right]dz \tag{10}$$
$$z_0 \geq z$$

The integrals in brackets can be found in tables.[10] Carrying out the integration gives

$$H_x = -(4\pi^2 I_m/\lambda)\sin(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2} e^{-2\pi(z_0 - z)/\lambda}\,dz$$
$$H_z = -(4\pi^2 I_m/\lambda)\cos(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2} e^{-2\pi(z_0 - z)/\lambda}\,dz \tag{11}$$
$$z_0 \geq z$$

which integrate to

$$H_x = -2\pi I_m \sin(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}[e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}]$$
$$H_z = -2\pi I_m \cos(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}[e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \tag{12}$$
$$z_0 \geq \delta/2$$

[10] D. Bierens de Haan, "Nouvelles Tables D'Integrales Définies," p. 223, Leide, Engels, 1867.

Below the recording medium, that is for $z_0 \leq -\delta/2$,

$$\begin{aligned} H_x &= -2\pi I_m \sin (2\pi x_0/\lambda) e^{+2\pi z_0/\lambda} [e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ H_z &= 2\pi I_m \cos (2\pi x_0/\lambda) e^{+2\pi z_0/\lambda} [e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ & z_0 \leq -\delta/2 \end{aligned} \tag{13}$$

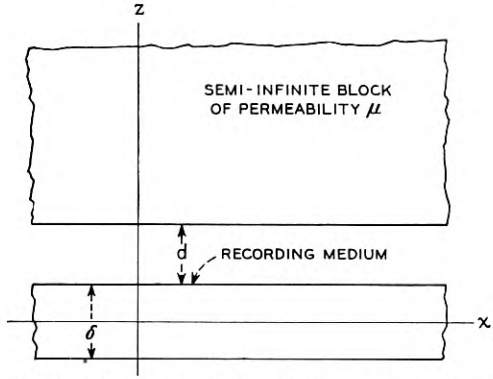



Fig. 11—Flat tape under idealized reproducing head.

Inside the recording medium,

$$\begin{aligned} H_x &= -(4\pi^2 I_m/\lambda) \sin (2\pi x_0/\lambda) \\ &\quad \cdot \left[ \int_{-\delta/2}^{z_0} e^{-2\pi(z_0-z)/\lambda}\, dz + \int_{z_0}^{\delta/2} e^{+2\pi(z_0-z)/\lambda}\, dz \right] \\ H_z &= -(4\pi^2 I_m/\lambda) \cos (2\pi x_0/\lambda) \\ &\quad \cdot \left[ \int_{-\delta/2}^{z_0} e^{-2\pi(z_0-z)/\lambda}\, dz - \int_{z_0}^{\delta/2} e^{+2\pi(z_0-z)/\lambda}\, dz \right] \end{aligned} \tag{14}$$

which integrate to

$$\begin{aligned} H_x &= -2\pi I_m \sin (2\pi x_0/\lambda)[2 - e^{-\pi\delta/\lambda}(e^{-2\pi z_0/\lambda} + e^{2\pi z_0/\lambda})] \\ H_z &= 2\pi I_m \cos (2\pi x_0/\lambda) e^{-\pi\delta/\lambda}(e^{-2\pi z_0/\lambda} - e^{2\pi z_0/\lambda}) \\ & \delta/2 \geq z_0 \geq -\delta/2 \end{aligned} \tag{15}$$

### The Fields in and Under the Reproducing Head

The idealized reproducing head amounts simply to a semi-infinite block of high permeability material with a flat face spaced a distance $d$ above the surface of the recording medium as shown in Fig. 11.

The problem of most interest is that of finding the $x$ component of magnetic induction, $B_x$, at any point $(x_0, z_0)$ in the idealized head and integrating this with respect to $z_0$ to determine the total flux passing through unit width (in the $y$ direction) of a plane $x = x_0$. This plane will then be allowed to move with a velocity $v$ by putting $x_0 = vt$ and the time rate of change of flux will be computed. Except for the effects of eddy currents, self demagnetization, gap loss, etc. (which are treated separately) this rate of change of flux should be proportional to the open circuit reproduced voltage. This is the only result of which direct use will be made but for the sake of completeness all the field components will be evaluated not only in the idealized head but also at all other points.

This problem is completely analogous to the problem of a point charge in front of a semi-infinite dielectric treated by Abraham and Becker[11] and can be solved by use of the method of images.

## The Field Inside the High Permeability Head

By analogy with the treatment of Abraham and Becker, the value of $B$ in the high permeability head is computed as though this head filled all space and as though the recording medium were polarized to a value $2\mu/(\mu + 1)$ times the actual value of polarization present. This gives directly from equations (12),

$$\begin{aligned} B_x &= -[2\mu/(\mu + 1)]2\pi I_m \sin(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}(e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ B_z &= -[2\mu/(\mu + 1)]2\pi I_m \cos(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}(e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ & z_0 \geq d + \delta/2 \end{aligned} \tag{16}$$

## The Field Below the Reproducing Head

Again by analogy with the treatment of Abraham and Becker, the field outside the idealized head is computed as though no head were present. The field is that due to the actual magnetized medium plus the field due to an image of the medium (centered about $z = 2d + \delta$). The intensity of magnetization of the image medium is $-(\mu - 1)/(\mu + 1)$ times the intensity of magnetization of the actual medium.

The field due to the image medium is computed from equations (13) after suitable modification. The required modifications are:

1. Multiply the right hand sides by $-(\mu - 1)/(\mu + 1)$ to take account of the magnitude and sign of the image magnetization as just discussed, and
2. Replace $z_0$ by $z_0 - (2d + \delta)$ to take account of the position of the image.

[11] M. Abraham and R. Becker, *The Classical Theory of Electricity and Magnetism.* p. 77, Blackie and Son Limited, London, 1937.

This gives the field due to the image plane as

$$H_{xi} = 2\pi I_m \frac{\mu - 1}{\mu + 1} \sin (2\pi x_0/\lambda) e^{2\pi(z_0 - 2d - \delta)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda})$$
$$H_{zi} = -2\pi I_m \frac{\mu - 1}{\mu + 1} \cos (2\pi x_0/\lambda) e^{2\pi(z_0 - 2d - \delta)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \tag{17}$$
$$z_0 \leq d + \delta/2$$

To this must be added the field due to the real medium which is given by equations (12) when $\delta/2 \leq z_0 \leq d + \delta/2$, by equations (15) when $-\delta/2 \leq z_0 \leq \delta/2$, and by equations (13) when $z_0 \leq -\delta/2$.

Performing this addition gives the following results:

Between the head and the recording medium,

$$H_x = -2\pi I_m \sin (2\pi x_0/\lambda) e^{-2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - 2z_0)/\lambda}\right]$$
$$H_z = -2\pi I_m \cos (2\pi x_0/\lambda) e^{-2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \cdot \left[1 + \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - 2z_0)/\lambda}\right] \tag{18}$$
$$d + \delta/2 \geq z_0 \geq \delta/2$$

Inside the recording medium,

$$H_x = -2\pi I_m \sin (2\pi x_0/\lambda) \cdot \left[2 - e^{-\pi\delta/\lambda}(e^{2\pi z_0/\lambda} + e^{-2\pi z_0/\lambda}) - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - z_0)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda})\right]$$
$$H_z = -2\pi I_m \cos (2\pi x_0/\lambda) \cdot \left[e^{-\pi\delta/\lambda}(e^{2\pi z_0/\lambda} - e^{-2\pi z_0/\lambda}) + \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - z_0)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda})\right] \tag{19}$$
$$\delta/2 \geq z_0 \geq \delta/2$$

Below the recording medium,

$$H_x = -2\pi I_m \sin (2\pi x_0/\lambda) e^{2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta)/\lambda}\right]$$
$$H_z = 2\pi I_m \cos (2\pi x_0/\lambda) e^{2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta)/\lambda}\right] \tag{20}$$
$$z_0 \leq -\delta/2$$

### The Flux per Unit Width in the Idealized Reproducing Head

The desired flux per unit width is computed from

$$\phi_x = \int_{d+\delta/2}^{\infty} B_x \, dz \tag{21}$$

where $B_x$ is given by equation (16). Performing the indicated integration gives

$$\phi_x = -\frac{2\mu}{\mu+1} 2\pi\delta I_m \sin(2\pi x_0/\lambda) \left[\frac{1 - e^{-2\pi\delta/\lambda}}{2\pi\delta/\lambda}\right] e^{-2\pi d/\lambda} \tag{22}$$

If the reproducing head moves past the recording medium with a velocity $v$ so that $x_0 = vt$,

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu+1} 4\pi v I_m (1 - e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{23}$$

where $\omega$ is $2\pi$ times the reproduced frequency. This is the result for unit width of the reproducing head. For a width of $W$ cm.,

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu+1} 4\pi W v I_m (1 - e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{24}$$

### The Case of Perpendicular Magnetization

Equation (23) was derived for the case of pure longitudinal magnetization as defined by equations (6). It will now be shown that this same result is obtained for $d\phi_x/dt$ if the magnetization is purely perpendicular, that is if

$$\begin{aligned} I_z &= -I_m \cos(2\pi x/\lambda) \\ I_x &= I_y = 0 \end{aligned} \tag{25}$$

In this case the divergence of $I$ is zero except at the surface of the tape and this magnetization is equivalent to a surface distribution of magnetic charge on the top and bottom surfaces of the tape. The magnitude of this charge density is just equal to $I_z$ so that on the top surface of the tape there is a surface density of charge given by

$$\sigma = -I_m \cos(2\pi x/\lambda) \qquad \text{at } z = \delta/2 \tag{26}$$

and on the bottom surface of the tape there is a surface density of charge given by

$$\sigma = I_m \cos(2\pi x/\lambda) \qquad \text{at } z = -\delta/2 \tag{27}$$

Since the permeability of the recording medium is assumed to be unity, this problem reduces to that of finding $d\phi_x/dt$ due to two infinitely thin

tapes of the sort to which equation (23) applies. One of these tapes is at $z = \delta/2$ and the other at $z = -\delta/2$.

The problem then is to rewrite equation (23) for a very thin tape and in terms of surface density of charge. As $\delta$ approaches zero, equation (23) reduces to

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m (2\pi\delta/\lambda) e^{-2\pi d/\lambda} \cos(\omega t) \tag{28}$$

From equation (7), the volume density of charge in this tape is

$$\rho = -(2\pi I_m/\lambda) \cos(2\pi x/\lambda)$$

But as $\delta$ approaches zero, the longitudinally magnetized tape to which equation (28) applies becomes equivalent to a surface distribution of magnetic charge of surface density equal to $\delta\rho$. This amounts, for the thin longitudinally magnetized tape, to a surface charge density of

$$\sigma_1 = -(2\pi\delta/\lambda) \cos(2\pi x/\lambda) \tag{29}$$

But the charge density on the top side of the perpendicularly magnetized tape is given by equation (26). Comparing these two values shows that the surface charge density in the thin longitudinally magnetized tape is just $2\pi\delta/\lambda$ times as great as the surface charge density on top of the perpendicularly magnetized medium. This means that $d\phi_x/dt$ due to the top side of the perpendicularly magnetized tape can be obtained by dividing the right hand side of equation (28) by $2\pi\delta/\lambda$. This gives

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m e^{-2\pi d/\lambda} \cos(\omega t) \tag{30}$$

due to the top side of the tape.

The contribution from the bottom side is obtained from equation (30) by replacing $d$ by $d + \delta$ (since the bottom side is spaced $d + \delta$ from the reproducing head) and changing the sign. Adding these two contributions gives for the total

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m (1 - e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{31}$$

This is the same as equation (23) and so the desired result has been established.

Note from equations (6) and (24) that in order to get the same result for the perpendicular and longitudinal cases it was necessary to assume a 90-degree phase difference between $I_x$ and $I_z$. The usual type of recording head lays down a pattern of magnetization which is neither purely per-

pendicular nor purely longitudinal but the two components are always in phase. This means that the two contributions to $d\phi_x/dt$ add as vectors at 90 degrees. If the intensity of magnetization in the recording medium is held constant while the relative values of perpendicular and longitudinal components are changed, the only effect on the reproduced signal is a change of phase.

## APPENDIX II

## THE FIELD DUE TO A ROUND WIRE

In Appendix I the field due to a sinusoidally magnetized flat medium such as a tape has been calculated and the rate of change of flux in an

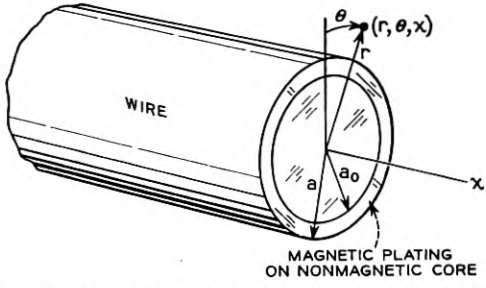



Fig. 12—Coordinate system for round wire calculations.

idealized reproducing head has been evaluated. The analogous calculations for a round wire have also been carried through and it is the purpose of this section to present some of the results. The derivation of these results seems too tedious and long to be presented here.

### The Recording Medium

Let the recording medium be a wire, the axis of which lies along the $x$ axis as shown in Fig. 12. Let the radius of the wire be $a$. To take account of plated wires as well as solid magnetic ones, let the wire have a nonmagnetic core of radius $a_0$. Let the cylindrical shell between $a_0$ and $a$ be magnetized sinusoidally in the $x$ direction so that

$$\begin{aligned} I_x &= I_m \operatorname{Sin}(2\pi x/\lambda) \\ I_r &= I_\theta = 0 \end{aligned} \tag{32}$$

By putting $a_0 = 0$ in the expressions which follow it will be possible to obtain the result for a solid magnetic wire.

### The Field in Free Space

If no reproducing head is present to disturb the field distribution, the computed field components at a point $(x_0, r_0)$ are

$$\begin{aligned} H_x &= -4\pi I_m \operatorname{Sin}(2\pi x_0/\lambda) K_0(2\pi r_0/\lambda)[(2\pi a/\lambda) I_1(2\pi a/\lambda) \\ &\qquad - (2\pi a_0/\lambda) I_1(2\pi a_0/\lambda)] \\ H_r &= -4\pi I_m \operatorname{Cos}(2\pi x_0/\lambda) K_1(2\pi r_0/\lambda)[(2\pi a/\lambda) I_1(2\pi a/\lambda) \\ &\qquad - (2\pi a_0/\lambda) I_1(2\pi a_0/\lambda)] \\ r_0 &\geq a \end{aligned} \tag{33}$$

A discussion and tabulation of the I and K functions can be found in Watson's "Theory of Bessel Functions."[12]

The field due to a solid magnetic wire is obtained by setting $a_0 = 0$ in equations (33). This gives

$$\begin{aligned} H_x &= -4\pi I_m \operatorname{Sin}(2\pi x_0/\lambda)(2\pi a/\lambda) K_0(2\pi r_0/\lambda) I_1(2\pi a/\lambda) \\ H_r &= -4\pi I_m \operatorname{Cos}(2\pi x_0/\lambda)(2\pi a/\lambda) K_1(2\pi r_0/\lambda) I_1(2\pi a/\lambda) \\ r_0 &\geq a \end{aligned} \tag{34}$$

### The Rate of Change of Flux in an Idealized Head

It has not been possible to carry out the calculations for an idealized head which is a satisfactory approximation to the grooved ring-type head often used in wire recording. The results presented below will apply only to reproducing heads which completely surround the wire. In this case the idealized head is an infinitely large block of core material of permeability $\mu$ pierced by a cylindrical hole of radius $R$ in which the wire is centered as shown in Fig. 13. At any point $(x_0, r_0)$ in the permeable medium the components of flux density can be shown to be

$$\begin{aligned} B_x &= \alpha H_x \\ B_r &= \alpha H_r \\ r_0 &\geq R \end{aligned} \tag{35}$$

where

$$\alpha = \frac{\mu}{(\mu - 1)(2\pi R/\lambda) I_0(2\pi R/\lambda) K_1(2\pi R/\lambda) + 1} \tag{36}$$

and $H_x$ and $H_r$ are given by equation (33).

[12] G. N. Watson, "A Treatise on the Theory of Bessel Functions," p. 79, 361, 698, Cambridge Univ. Press, 1922.

The total flux through a plane $x = x_0$ in the permeable medium is obtained by integrating

$$\phi_x = \int_R^\infty B_x(2\pi r)\, dr \tag{37}$$

This gives

$$\phi_x = -2\lambda^2\alpha I_m \sin(2\pi x_0/\lambda)(2\pi R/\lambda)K_1(2\pi R/\lambda)[(2\pi a/\lambda)I_1(2\pi a/\lambda) - (2\pi a_0/\lambda)I_1(2\pi a_0/\lambda)] \tag{38}$$

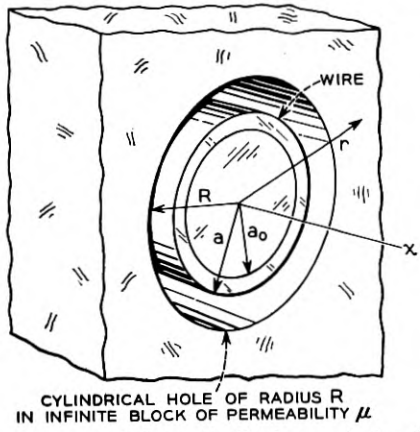



Fig. 13—Round wire surrounded by idealized reproducing head consisting of an infinite block of core material of permeability $\mu$.

If the plane $x = x_0$ moves with a velocity $v$ with respect to the wire so that $x_0 = vt$, then

$$\frac{d\phi_x}{dt} = -4\pi\lambda\alpha v I_m \cos(\omega t)(2\pi R/\lambda)K_1(2\pi R/\lambda)[(2\pi a/\lambda)I_1(2\pi a/\lambda) - (2\pi a_0/\lambda)I_1(2\pi a_0/\lambda)] \tag{39}$$

where $\omega = 2\pi f$ and $f$ is the reproduced frequency.

## Special Cases

Equation (39) can be used to compute the response of a simple reproducing head consisting of a single turn of very fine[13] wire as shown in Fig. 14.

In this case $\mu = 1$ and equation (36) shows that $\alpha = 1$. Furthermore if the wire is solid so that $a_0 = 0$, equation (39) reduces to

[13] Unless the diameter of the wire is small compared to the recorded wavelength there will be additional loss not accounted for by 39.

$$\frac{d\phi_x}{dt} = -4\pi\lambda v I_m \cos(\omega t)(2\pi R/\lambda)(2\pi a/\lambda)K_1(2\pi R/\lambda)I_1(2\pi a/\lambda) \quad (40)$$

As $\lambda$ approaches infinity, $K_1(2\pi R/\lambda)$ approaches $\lambda/2\pi R$ and $I_1(2\pi a/\lambda)$ approaches $\pi a/\lambda$ so that, for very long wavelengths, equation (40) reduces to

$$\frac{d\phi_x}{dt} = -4\pi I_m v(2\pi/\lambda)(\pi a^2)\cos(\omega t) \quad (41)$$

This relation (which could have been derived in a much simpler manner) should be useful for the experimental determination of the intensity of magnetization, $I_m$.

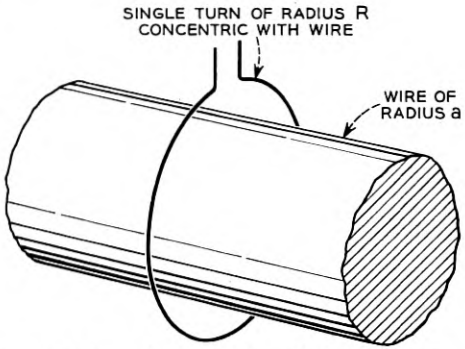



Fig. 14—Elementary reproducing head consisting of a single turn of wire.

Another case of some interest corresponds to a high permeability reproducing head which surrounds the wire. In this case $\mu$ is great enough so that equation (36) reduces to

$$\alpha = \frac{1}{(2\pi R/\lambda)I_0(2\pi R/\lambda)K_1(2\pi R/\lambda)} \quad (42)$$

If it is assumed, in addition, that the wire is solid so that $a_0 = 0$, then equations (42) and (39) give

$$\frac{d\phi_x}{dt} = -4\pi\lambda v I_m \cos(\omega t)(2\pi a/\lambda)I_1(2\pi a/\lambda)/I_0(2\pi R/\lambda) \quad (43)$$

### Comparison between Round Wire and Flat Medium Response

It is interesting to compare equation (43) with equation (24) to see how the response characteristic of a round wire compares with that of a tape. Assuming $\mu \gg 1$, the appropriate equation for the flat medium is

$$\frac{d\phi_x}{dt} = -4\pi W v I_m \cos(\omega t)(1 - e^{-(2\pi\delta/\lambda)})e^{-2\pi d/\lambda} \quad (44)$$

To compare equations (43) and (44), consider first the limiting cases of very long and very short wavelength. As $\lambda$ approaches infinity they reduce to

$$\frac{d\phi_x}{dt} = -\pi a^2(8\pi^2 v/\lambda)I_m \cos(\omega t) \tag{45}$$

for the wire and

$$\frac{d\phi_x}{dt} = -\delta W(8\pi^2 v/\lambda)I_m \cos(\omega t) \tag{46}$$

for the tape.

These two expressions are identical provided the cross section area of the wire, $(\pi a^2)$, is the same as that of the recorded track on the tape, $(\delta W)$.

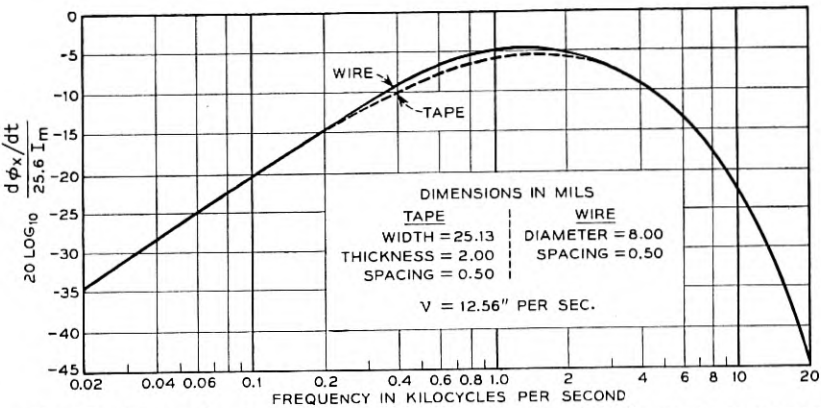



Fig. 15—Computed responses for wire and tape showing that the responses are very similar provided the dimensions of the wire and tape are suitably related.

As $\lambda$ approaches zero, the two expressions reduce to

$$\frac{d\phi_x}{dt} = -4\pi v(2\pi a)\sqrt{R/a}\,e^{-2\pi(R-a)/\lambda}I_m \cos(\omega t) \tag{47}$$

for the wire, and

$$\frac{d\phi_x}{dt} = -4\pi v(W)e^{-2\pi d/\lambda}I_m \cos(\omega t) \tag{48}$$

for the tape.

Suppose that the reproducing head makes reasonably good contact with the wire so that $\sqrt{R/a} \doteq 1$. In this case equations (47) and (48) are identical provided the circumference of the wire, $(2\pi a)$, is the same as the width of the recorded track on the tape and provided also the effective spacing between reproducing head and medium is the same in the two cases, $(d = R - a)$. In both cases only a thin surface layer of the recording medium is effective in producing high frequency response. For this reason the

high-frequency response is independent of the "thickness" of the medium and is directly proportional to the "width" of the track provided $2\pi a$ is interpreted as the width of track on a wire.

The comparisons which have just been made indicate that if the dimensions of a wire and of a tape are suitably related, the two media should give identical response at very high and very low frequencies provided they are equally magnetized. The dimensional requirements are

$$\begin{aligned} \pi a^2 &= \delta W, \\ 2\pi a &= W, \text{ and} \\ R - a &= d \end{aligned} \tag{49}$$

In order to show how the computed responses compare at intermediate frequencies, numerical calculations have been made for a special case in which equations (49) are satisfied. The case chosen is that of a wire 8 mils in diameter moving at a velocity of 12.56 in./sec. past a reproducing head which is effectively one half mil out of contact with the wire ($R - a = 0.5(10)^{-3}$ in.). By equations (49) the corresponding flat medium is a tape which is 2 mils thick and 25.13 mils wide. The tape is assumed to be moving with a velocity of 12.56 in./sec. past a reproducing head which is also effectively one half mil out of contact ($d = 0.5(10)^{-3}$ in.). In this case the numerical constants in equations (43) and (44) are equal. That is,

$$8\pi^2 av = 4\pi W v = 25.6 \text{ cm.}^2/\text{sec.}$$

and the quantity to be computed and compared for the two cases is

$$20 \log_{10} \frac{d\phi_x/dt}{25.6 I_m}$$

The computed curves are shown in Fig. 15 from which it can be seen that they coincide at low and high frequencies as planned and that furthermore they differ by no more than 1.5 db in the middle range of frequencies.

As has been pointed out, equation (43) applies only to the unusual case in which the head completely surrounds the wire. The similarity of the two curves of Fig. 15, however, suggests a way of computing approximately the response to be expected when the wire head makes contact with only a part of the circumference of the wire. It suggests that the computation be carried out as though the wire were a flat medium of suitably chosen dimensions. In order to make the high frequency end come out right one would expect that $W$ in equation (44) should be given a value equal to the length of the arc of contact between the wire and the head. To make the low frequency end come out right, $\delta$ must be given a value which makes the cross section area of the tape equal to that of the wire, i.e. such that $\delta W = \pi a^2$.

# Some Results Concerning the Partial Differential Equations Describing the Flow of Holes and Electrons in Semiconductors

By R. C. Prim, III

(*Manuscript Received June 22, 1951*)

The subject equations are investigated with the aim of establishing some general properties of the flow fields which they describe, including the existence or non-existence of classes of exact solutions having certain formal properties. The results include a number of geometric characteristics of the vector fields involved, a suggestive reformulation of the partial differential equations restricting carrier concentration and electrostatic potential, and several classes of exact solutions involving arbitrary constants and/or functions. Of particular interest is a family of solutions in closed form for the steady-state, no-recombination case involving an arbitrary harmonic function in three dimensions.

TABLE OF CONTENTS



## A. INTRODUCTION

THIS paper is concerned with the system of relations describing the flow of holes and electrons in the interior of a homogeneous semiconductor subject to the assumption of constant temperature, electrical neutrality, and constant difference in concentrations of ionized donor and acceptor centers. These relations are:

$$\operatorname{div} \overset{\circ}{||}_p = -e\left[\mathcal{R} + \frac{\partial p}{\partial t}\right] \tag{1}$$

$$\operatorname{div} \overset{\circ}{||}_n = e\left[\mathcal{R} + \frac{\partial n}{\partial t}\right] \tag{2}$$

$$\overset{\circ}{||}_p = -\mu_p e \left[ p \operatorname{grad} \mathcal{U} + \frac{kT}{e} \operatorname{grad} p \right] \tag{3}$$

$$\overset{\circ}{||}_n = -\mu_n e \left[ n \operatorname{grad} \mathcal{U} - \frac{kT}{e} \operatorname{grad} n \right] \tag{4}$$

$$n - p = n_0 - p_0 \equiv N \text{ (a constant)} \tag{5}$$

$$n,\ p \geq 0 \tag{6}$$

$$\overset{\circ}{||} = \overset{\circ}{||}_p + \overset{\circ}{||}_n \tag{7}$$

wherein

$n$: concentration of negative carriers (electrons)
$p$: concentration of positive carriers (holes)
$n_0$: thermal equilibrium value of $n$
$p_0$: thermal equilibrium value of $p$
$\overset{\circ}{||}_p$: hole current density vector
$\overset{\circ}{||}_n$: electron current density vector
$\overset{\circ}{||}$: total current density vector
$t$: time variable
$e$: magnitude of electronic charge
$k$: Boltzmann's constant
$\mu_p$: hole mobility constant
$\mu_n$: electron mobility constant
$T$: absolute temperature (assumed constant with time and uniform)
$\mathcal{U}$: potential of electrical intensity field
$\mathfrak{R}$: electron-hole recombination rate function (will usually be regarded as depending on $p - p_0$ and $n - n_0$ or equivalent variables).

These relations have fundamental application to transistor electronics, photoelectric effects, and related phenomena. Detailed discussions of their physical bases will be found in References 1 and 3. In brief, (1) and (2) are conservation conditions for the positive and negative carriers; (3) and (4) express the dependence of the local current densities on the electrostatic potential gradient and on the carrier concentration gradients (i.e., on conduction and diffusion); (5) expresses the condition of electrical neutrality under the assumption of a constant difference in concentrations of ionized donor and acceptor centers; and (6) and (7) are self evident.

The present study is directed toward the discovery of (1) general properties of the flow fields inside semiconductors and (2) families of exact solutions to the flow equations. The approach to the latter objective is through

the "inverse method" which has proved very useful in the study of various non-linear partial differential equation systems in mechanics. In the inverse method, one proceeds by formal devices suggested by the equations under study to try to find families of solutions to the equations which involve arbitrary constants or, preferably, arbitrary functions. This is done without reference to any preconceived boundary value problems. After a pool of such families of solutions is available, it can be examined from the point of view of finding boundary value problems of interest consistent with any of the solutions in hand. The likelihood of finding solutions of interest in this way is of course greatly enhanced when the solutions involve arbitrary functions. Aside from providing solutions of some useful boundary value problems, the solutions found by the inverse method constitute a reference bank of non-trivial exact solutions against which to check numerical methods and approximation schemes (based, for example, on the assumption that a particular term can be neglected) for solving problems of more immediate practical interest.

J. Bardeen has demonstrated (in Reference 2) how the steady-state behavior of contact-semiconductor combinations can be explained on the basis of the characteristics of (1) the flow field inside the semiconductor and (2) those of the barrier layer at the contact. The present study is concerned in this connection only with the first of these influences. It provides, for example, a complete solution for the spherically symmetric flow field without recombination for arbitrary currents—a generalization of the zero-total current solution given by Bardeen. In the absence of surface recombination this spherically symmetric solution provides the hemispherically symmetric flow field in the neighborhood of a point contact on a plane surface and remote from other electrodes or surfaces. This spherically symmetric solution is contained as a particular case in a family of solutions involving an arbitrary harmonic function in three dimensions. Other choices of the harmonic function can be made to yield flow fields associated with numerous electrode configurations of immediate practical interest, for example that of the type-A transistor.

The objective of the present paper is to find (or establish the non-existence of) broad classes of solutions, and not to undertake detailed studies of any particular solutions. Such detailed studies of particular cases from the family of solutions mentioned above (and from other families found in this study) will form the subject matter of papers dealing with specific flow field configurations. However, in order to illustrate the interpretation of mathematical arbitrary constants in terms of basic physical parameters, the analysis of the spherically symmetric solution mentioned above is car-

ried up to the point of actual substitution of numerical values in the formulae.

Note: In the following, functions and constants described as "arbitrary" are to be considered as being subject nevertheless to the restrictions implied by (6). In any particular case it is an elementary matter to determine these restrictions and we shall not usually carry out this detail. Also, "arbitrary" functions are subject to appropriate differentiability conditions readily evident in any particular case.

### B. Some Properties of the Current Density Vector Fields

Several interesting properties of the current density vector fields $\overset{\circ}{||}_p$, $\overset{\circ}{||}_n$, and $\overset{\circ}{||}$ are easily found from (3)–(5).

It is evident that (3) and (4) can be rewritten as

$$\overset{\circ}{||}_p = -e\mu_p\, p \operatorname{grad}\left(\mathcal{V} + \frac{kT}{e} \ln p\right) \tag{8}$$

and

$$\overset{\circ}{||}_n = -e\mu_n\, n \operatorname{grad}\left(\mathcal{V} - \frac{kT}{e} \ln n\right). \tag{9}$$

From (3), (4), and (7) we have

$$\overset{\circ}{||} = -e(\mu_n n + \mu_p p) \operatorname{grad} \mathcal{V} + kT \operatorname{grad} (\mu_n n - \mu_p p) \tag{10}$$

which because of (5) can be rewritten as

$$\overset{\circ}{||} = -e\,(\mu_n\, n + \mu_p\, p) \operatorname{grad}\left[\mathcal{V} - \frac{kT}{e} \frac{\mu_\mu - \mu_p}{\mu_n + \mu_p} \ln\,(\mu_n\, n + \mu_p\, p)\right]. \tag{11}$$

Now (8), (9) and (11) are all of the form

$$\mathbf{u} = \phi \operatorname{grad} \psi$$

and hence obviously satisfy the condition

$$\mathbf{u} \cdot \operatorname{curl} \mathbf{u} = 0.$$

Therefore we have

*Theorem 1:* $\overset{\circ}{||}_p$, $\overset{\circ}{||}_n$, and $\overset{\circ}{||}$ are surface-normal vector fields.

From (8)–(10) we find, using (5)

$$\operatorname{curl} \overset{\circ}{||}_p = -e\mu_p \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{12}$$

$$\operatorname{curl} \overset{\circ}{||}_n = -e\mu_n \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{13}$$

and

$$\operatorname{curl} \overset{\circ}{||} = -e(\mu_n + \mu_p) \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{14}$$

whence

*Theorem 2:*

$$\frac{\operatorname{curl} \overset{\circ}{||}_p}{\mu_p} = \frac{\operatorname{curl} \overset{\circ}{||}_n}{\mu_n} = \frac{\operatorname{curl} \overset{\circ}{||}}{\mu_n + \mu_p}.$$

That is, curl $\overset{\circ}{||}_p$, curl $\overset{\circ}{||}_n$, and curl $\overset{\circ}{||}$ are constant multiples of one another.

and

*Theorem 3:* $\overset{\circ}{||}_p$, $\overset{\circ}{||}_n$, and $\overset{\circ}{||}$ are irrotational if and only if

$$\operatorname{grad} p = 0 \qquad (p = p(t))$$

or

$$\operatorname{grad} \mathcal{V} = 0 \qquad (\mathcal{V} = \mathcal{V}(t))$$

or

$$\mathcal{V} = \mathcal{V}(p, t).$$

The following interesting relations can be obtained from (8) and (9) (they are really consequences of Theorem 1):

$$\operatorname{curl} \overset{\circ}{||}_p = \operatorname{grad} \ln p \times \overset{\circ}{||}_p \tag{15}$$

and

$$\operatorname{curl} \overset{\circ}{||}_n = \operatorname{grad} \ln n \times \overset{\circ}{||}_n. \tag{16}$$

Now from (3) – (5) we find

$$\overset{\circ}{||}_p \times \overset{\circ}{||}_n = e\mu_n\mu_p kT(n + p) \operatorname{grad} p \times \operatorname{grad} \mathcal{V} \tag{17a}$$

$$= \tfrac{1}{2}\mu_n\mu_p kT(n + p) \operatorname{grad} (n + p) \times \operatorname{grad} \mathcal{V} \tag{17b}$$

$$= \tfrac{1}{4}e\mu_n\mu_p kT \operatorname{grad} (n + p)^2 \times \operatorname{grad} \mathcal{V} \tag{17c}$$

$$= \tfrac{1}{4}e\mu_n\mu_p kT \operatorname{curl} [(n + p)^2 \operatorname{grad} \mathcal{V}] \tag{17d}$$

and

$$\frac{\overset{\circ}{||}_p}{\mu_p e} - \frac{\overset{\circ}{||}_n}{\mu_n e} = \operatorname{grad}\left[N\mathcal{V} - \frac{kT}{e}(n + p)\right] \tag{18}$$

and

$$\frac{\overset{\circ}{||}_p}{\mu_p e} + \frac{\overset{\circ}{||}_n}{\mu_n e} = -(n + p) \operatorname{grad} \mathcal{V}. \tag{19}$$

[*Note:* As is suggested by (18) and (19), the total carrier concentration

$$\mathcal{P} \equiv n + p = N + 2p = 2n - N \quad (\mathcal{P} \geq |N|)$$

will frequently appear as the "natural" concentration variable in the relations with which we shall be working. Hence, expressions involving $p$, or $p$ and $n$ will often be replaced in the sequel by their equivalents in terms of the variable $\mathcal{P}$. It will be noted that

$$\text{grad } \mathcal{P} = 2 \text{ grad } p = 2 \text{ grad } n.]$$

Equations (17) and (19) yield at once the following theorems:

*Theorem 4:* The vector field

$$\overset{\circ}{\mathcal{J}}_p \times \overset{\circ}{\mathcal{J}}_n = \overset{\circ}{\mathcal{J}} \times \overset{\circ}{\mathcal{J}}_n = \overset{\circ}{\mathcal{J}}_p \times \overset{\circ}{\mathcal{J}}$$

is solenoidal.

*Theorem 5:* The vector field

$$\left(\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} - \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n}\right) \text{ is irrotational with a potential } (-eN\mathcal{V} + kT\mathcal{P}).$$

*Theorem 6:* The vector field

$$\left(\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} + \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n}\right) \text{ is surface-normal (to the surfaces of constant } \mathcal{V}).$$

*Theorem 7:* $\overset{\circ}{\mathcal{J}}_p$, $\overset{\circ}{\mathcal{J}}_n$, $\overset{\circ}{\mathcal{J}}$, grad $\mathcal{V}$, and grad $p$ are coplanar vectors.

*Theorem 8:* The flow lines of any two of the fields $\overset{\circ}{\mathcal{J}}_p$, $\overset{\circ}{\mathcal{J}}_n$, and $\overset{\circ}{\mathcal{J}}$ coincide if and only if

$$\text{grad } p = 0 \qquad (p = p(t))$$

or

$$\text{grad } \mathcal{V} = 0 \qquad (\mathcal{V} = \mathcal{V}(t))$$

or

$$\mathcal{V} = \mathcal{V}(p, t).$$

Also, from (17) and (19) we obtain the curious relations:

$$\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} \times \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n} = -\frac{kT}{2} \text{ grad } \mathcal{P} \times \left(\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} + \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n}\right) \tag{20a}$$

$$= -\frac{kT}{2} \mathcal{P} \text{ curl } \left(\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} + \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n}\right) \tag{20b}$$

$$= -\frac{kT}{2} \text{ curl } \left[\mathcal{P}\left(\frac{\overset{\circ}{\mathcal{J}}_p}{\mu_p} + \frac{\overset{\circ}{\mathcal{J}}_n}{\mu_n}\right)\right]. \tag{20c}$$

Finally, by taking the divergence of (7) and making use first of (1) and (2) and then of (5), we obtain:

[*Theorem 9:* The vector field $\overset{\circ}{||}$ is solenoidal.

### C. Formulation of Partial Differential Equation System Restricting $\mathcal{P}$ and $\mathcal{U}$

A very convenient formulation of the partial differential equations restricting $\mathcal{P}$ and $\mathcal{U}$ is suggested by (18) and (19). Taking the divergence of these equations and substituting (1) and (2) into the results we obtain:

$$\text{div grad}\left(N\mathcal{U} - \frac{kT}{e}\mathcal{P}\right) = -\alpha\left(\mathcal{R} + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right) \tag{21}$$

and

$$\text{div}\,(\mathcal{P}\,\text{grad}\,\mathcal{U}) = \beta\left(\mathcal{R} + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right) \tag{22}$$

wherein for brevity we have set

$$\alpha \equiv \frac{1}{\mu_p} + \frac{1}{\mu_n}$$

and

$$\beta \equiv \frac{1}{\mu_p} - \frac{1}{\mu_n}$$

and shall henceforth assume $\beta \neq 0$, i.e., $\mu_p \neq \mu_n$. Equations (21) and (22) yield immediately a derived equation not containing explicitly the terms introduced by recombination and time variations:

$$\text{div grad}\left(N\mathcal{U} - \frac{kT}{e}\mathcal{P}\right) = -\frac{\alpha}{\beta}\,\text{div}\,(\mathcal{P}\,\text{grad}\,\mathcal{U}) \tag{23a}$$

or

$$\text{div}\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\text{grad}\,\mathcal{U} - \frac{kT}{e}\,\text{grad}\,\mathcal{P}\right] = 0 \tag{23b}$$

or

$$\text{div}\left(\left(\mathcal{P} + \frac{\beta N}{\alpha}\right)\text{grad}\left[\mathcal{U} - \frac{\beta kT}{\alpha e}\ln\left(\mathcal{P} + \frac{\beta N}{\alpha}\right)\right]\right) = 0. \tag{23c}$$

Either the set (21) and (22) or one of the forms of (23) together with either (21) or (22) constitutes a basic set of two partial differential equations determining $\mathcal{P}$ and $\mathcal{U}$. We are here considering $\mathcal{R}$ as $\mathcal{R}(\mathcal{P})$.

It will be observed that (23) is equivalent to the condition

$$\operatorname{div} \overset{\circ}{||} = 0 \tag{24}$$

established as Theorem 9.
(In terms of $\mathcal{P}$, (10) becomes

$$\overset{\circ}{||} = -\frac{e(\mu_n - \mu_p)}{2}\left[\left(\frac{\alpha}{\beta}\mathcal{P} + N\right)\operatorname{grad} \mathcal{V} - \frac{kT}{e}\operatorname{grad} \mathcal{P}\right]. \tag{25}$$

In most of the following sections we shall find it expedient to consider separately the cases $N \neq 0$ and $N = 0$ (associated respectively with semiconductors of the extrinsic and intrinsic conductivity types). *For the case* $N \neq 0$, use will be made frequently of new dependent variables $\mathfrak{U}$ and $\mathfrak{IC}$ defined by:

$$\mathfrak{U} \equiv \frac{kT}{eN}\mathcal{P} \tag{26}$$

$$\mathfrak{IC} \equiv \mathcal{V} - \frac{kT}{eN}\mathcal{P} = \mathcal{V} - \mathfrak{U}. \tag{27}$$

That is,

$$\mathcal{P} \equiv \frac{eN}{kT}\mathfrak{U} \tag{28}$$

$$\mathcal{V} \equiv \mathfrak{U} + \mathfrak{IC} \tag{29}$$

will be substituted into relations involving $\mathcal{P}$ and $\mathcal{V}$ to obtain the corresponding relations in terms of $\mathfrak{U}$ and $\mathfrak{IC}$. Incidentally, it will be noted that $\mathfrak{U}$ and $\mathfrak{IC}$ have the dimensions of voltage.

In terms of $\mathfrak{U}$ and $\mathfrak{IC}$ the basic equations (21)–(23) can be written:

$$\operatorname{div} \operatorname{grad} \mathfrak{IC} = -\frac{\alpha}{N}\left[\mathfrak{R} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right] \tag{30}$$

$$\operatorname{div} [\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})] = \frac{\beta kT}{eN}\left[\mathfrak{R} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right] \tag{31}$$

$$\operatorname{div}\left[\operatorname{grad} \mathfrak{IC} + \frac{\alpha e}{\beta kT}\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})\right] = 0 \tag{32}$$

wherein $\mathfrak{R}$ will be considered as $\mathfrak{R}(\mathfrak{U})$.

It will be observed that, in the absence of recombination and time variation, (30)–(32) reduce to

$$\operatorname{div} \operatorname{grad} \mathfrak{IC} = 0 \tag{33}$$

and $[N \neq 0]$

$$\operatorname{div} [\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})] = 0 \tag{34}$$

The elegant form of this set of equations furnished the original motivation for the introduction of the variables $\mathfrak{U}$ and $\mathfrak{K}$. The comparable equations for $N = 0$ are

$$\text{div grad } \mathfrak{P} = 0 \tag{35}$$

$$[N = 0]$$

$$\text{div } [\mathfrak{P} \text{ grad } \mathfrak{V}] = 0. \tag{36}$$

### D. The Recombination Rate Function $\mathfrak{R}$

In order to avoid undue confusion in the sequel we shall at this point make some clarifying remarks concerning the function $\mathfrak{R}$. As was stated in the Introduction, we basically regard $\mathfrak{R}$ as a function of $p - p_0$ and $n - n_0$. However, because of (5), any expression in $p - p_0$ and $n - n_0$ can be replaced by one in which (say) $p$ is the only field variable quantity. It is then convenient to regard $\mathfrak{R}$ as a function of $p$ and write it $\mathfrak{R}(p)$. When dealing with expressions in terms of $\mathfrak{P}$ and of $\mathfrak{U}$, it is convenient to regard $\mathfrak{R}$ as a function of one of these variables and to indicate this fact by writing $\mathfrak{R}(\mathfrak{P})$ or $\mathfrak{R}(\mathfrak{U})$. When we do this we do not mean that $\mathfrak{R}(\mathfrak{P})$ (say) is the same algebraic function of $\mathfrak{P}$ as $\mathfrak{R}(p)$ is of $p$, but rather that $\mathfrak{R}(p)$ is the function of $p$ obtained when one substitutes $\mathfrak{P} = N + 2p$ into $\mathfrak{R}(\mathfrak{P})$.

For example, *for constant mean lifetime recombination*

$$\mathfrak{R}(p) \equiv \frac{1}{\tau_0}(p - p_0) \tag{37a}$$

$$\mathfrak{R}(\mathfrak{P}) \equiv \frac{1}{2\tau_0}(\mathfrak{P} - \mathfrak{P}_0) \tag{37b}$$

$$\mathfrak{R}(\mathfrak{U}) \equiv \frac{eN}{2\tau_0 kT}(\mathfrak{U} - \mathfrak{U}_0) \tag{37c}$$

with $\tau_0$ constant;
and *for mass-action recombination*

$$\mathfrak{R}(p) \equiv \frac{1}{n_0\tau_0}[p(p + N) - p_0 n_0] \tag{38a}$$

$$\mathfrak{R}(\mathfrak{P}) \equiv \frac{1}{2\tau_0(\mathfrak{P}_0 + N)}(\mathfrak{P}^2 - \mathfrak{P}_0^2) \tag{38b}$$

$$\mathfrak{R}(\mathfrak{U}) \equiv \frac{e^2N^2}{2k^2T^2\tau_0\,(\mathfrak{P}_0 + N)}(\mathfrak{U}^2 - \mathfrak{U}_0^2). \tag{38c}$$

### E. Addition of Arbitrary Time Functions to $\mathcal{U}$ and $\mathcal{K}$

Since only the gradient of $\mathcal{U}$ appears in the basic equations (21) and (22), it is evident that if

$$\mathcal{U} = \mathcal{U}(x, y, z, t)$$

and

$$\mathcal{P} = \mathcal{P}(x, y, z, t)$$

are a pair of functions satisfying (21) and (22), then so also are

$$\tilde{\mathcal{U}} = \mathcal{U}(x, y, z, t) + \tilde{m}(t)$$

and

$$\tilde{\mathcal{P}} = \mathcal{P}(x, y, z, t)$$

where $\tilde{m}(t)$ is an arbitrary time function.
And since $\mathcal{U} = \mathcal{U} + \mathcal{K}$, if

$$\mathcal{K} = \mathcal{K}(x, y, z, t)$$

and

$$\mathcal{U} = \mathcal{U}(x, y, z, t)$$

are a pair of functions satisfying (30)–(32), so also are

$$\tilde{\mathcal{K}} = \mathcal{K}(x, y, z, t) + \tilde{m}(t)$$

and

$$\tilde{\mathcal{U}} = \mathcal{U}(x, y, z, t).$$

These arbitrary additive functions with zero gradients are physically trivial in that they merely reflect the arbitrariness of the reference voltage level. They will, however, be retained for the sake of formal completeness whenever they appear in the subsequent analyses.

### F. Summary of Solutions for No Recombination or Time Variation

The next ten sections of this paper (Sections G–Q) contain a sequence of detailed analyses in which is determined the existence or non-existence of solution fields having certain prescribed formal properties. In most of these studies time variability and recombination are admitted and the analysis includes the establishment of the class of recombination rate functions $\mathcal{R}$ consistent with the property under consideration. In those cases where solutions are found to exist, they are expressed in the simplest convenient

terms: in closed form, or as solutions of an ordinary differential equation, or as solutions of a single partial differential equation. The solutions found usually involve arbitrary constants and/or arbitrary functions of various kinds.

The present section is intended to provide a skimpy but compact sampling of the results obtained in the next ten sections. It will be confined to a simple listing of solutions found and furthermore will contain only the forms to which these solutions reduce when recombination and time variation are excluded. (Some solutions are lost under this reduction.) A heading will indicate the section(s) from which the solution comes as well as the formal property associated with each solution.

For the sake of conciseness and simplicity the symbols denoting arbitrary constants and functions in this section are independent of those employed in the later sections. They are to be interpreted as follows:

$A$, $B$: arbitrary constants
$h(x, y, z)$: any harmonic function
(or with subscript)

$(\tilde{\mathcal{V}}, \tilde{\mathcal{P}})$: any given solution field

[G. grad $\mathcal{V} = 0$]

$$\begin{cases} \mathcal{V} = A \\ \mathcal{P} = h(x, y, z) \end{cases}$$

[H, I. grad $\mathcal{P} = 0$]

$$\begin{cases} \mathcal{V} = h(x, y, z) \\ \mathcal{P} = A \end{cases}$$

[J. grad $\mathcal{K} = 0$, $N \neq 0$]

$$\begin{cases} \mathcal{V} = A + \sqrt{h(x, y, z)} \\ \mathcal{P} = \dfrac{Ne}{kT}\sqrt{h(x, y, z)} \end{cases}$$

[K, L. $\mathcal{V} = \mathcal{V}(\mathcal{P})$, $N \neq 0$]

$$(A \neq 0) \begin{cases} \mathcal{V} = h(x, y, z) + A\Lambda\left[\dfrac{B - h(x, y, z)}{A}\right] \\ \mathcal{P} = \dfrac{Ne}{kT} A\Lambda\left[\dfrac{B - h(x, y, z)}{A}\right] \end{cases}$$

(For definition of function $\Lambda$ see Equation (87) and Figs. 1 and 2.)

[K, M. $\mathfrak{V} = \mathfrak{V}(\mathcal{P})$, $N = 0$]

$$\begin{cases} \mathfrak{V} = A \ln h(x, y, z) + B \\ \mathcal{P} = h(x, y, z) \end{cases}$$

[N, O. grad $\mathcal{P}$·grad $\mathfrak{V} = 0$]

$$\begin{cases} \mathfrak{V} = h_1(x, y, z) \\ \mathcal{P} = h_2(x, y, z) \\ \quad \text{provided} \\ \text{grad } h_1(x, y, z) \cdot \text{grad } h_2(x, y, z) = 0 \end{cases}$$

[N. grad $\mathfrak{U}$·grad $\mathfrak{K} = 0$, $N \neq 0$]

$$\begin{cases} \mathfrak{V} = \sqrt{h_1(x, y, z)} + h_2(x, y, z) \\ \mathcal{P} = \dfrac{Ne}{kT} \sqrt{h_1(x, y, z)} \\ \quad \text{provided} \\ \text{grad } h_1(x, y, z) \cdot \text{grad } h_2(x, y, z) = 0 \end{cases}$$

[P. grad $\mathcal{P}$·grad $h = 0$]

$$\begin{cases} \mathfrak{V} = \tilde{\mathfrak{V}} + h(x, y, z) \\ \mathcal{P} = \tilde{\mathcal{P}} \\ \quad \text{provided} \\ \text{grad } \tilde{\mathcal{P}} \cdot \text{grad } h(x, y, z) = 0. \end{cases}$$

## G. Solutions with $\mathfrak{V} = \mathfrak{V}(t)$

Our point of view in general is that $\mathcal{P}$ and $\mathfrak{V}$ (or $\mathfrak{U}$ and $\mathfrak{K}$) are functions of three space coordinates and time, so that $\mathfrak{V} = \mathfrak{V}(t)$ implies for example that $\dfrac{\partial \mathfrak{V}}{\partial x} = \dfrac{\partial \mathfrak{V}}{\partial y} = \dfrac{\partial \mathfrak{V}}{\partial z} = 0$. That is to say, we now seek solutions for which everywhere

$$\text{grad } \mathfrak{V} = 0. \tag{39}$$

From (21) and (22) this condition gives us the following restrictions on

$\mathcal{P}$ (and none on $\mathcal{U}(t)$):

$$\text{div grad } \mathcal{P} = 0 \tag{40}$$

and

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t} = 0. \tag{41}$$

By operating with div grad on (41) we obtain

$$\mathcal{R}''(\mathcal{P}) = 0$$

(we consistently use primes to denote differentiation with respect to the argument of a function of a single variable—e.g.,

$$\mathcal{R}''(\mathcal{P}) \equiv \frac{d^2\mathcal{R}(\mathcal{P})}{d\mathcal{P}^2}\Bigg)$$

whence,

$$2\mathcal{R}(\mathcal{P}) = A\mathcal{P} + B \tag{42}$$

(A, B: arbitrary constants). Substituting (42) into (41) we obtain

$$\frac{\partial \mathcal{P}}{\partial t} + A\mathcal{P} = -B$$

whence

$$\mathcal{P} = c(x, y, z)\epsilon^{-At} - B/A \qquad (A \neq 0) \tag{43a}$$

or

$$\mathcal{P} = c(x, y, z) - Bt \qquad (A = 0). \tag{43b}$$

From (36) it follows that

$$\text{div grad } c(x, y, z) = 0, \tag{44}$$

that is, $c$ must be harmonic.

In brief, if $\mathcal{R}(\mathcal{P})$ is of the form given in (42), any $\mathcal{U}(t)$ and (43) constitute solutions to the flow equations for any harmonic $c(x, y, z)$. Other forms of $\mathcal{R}(\mathcal{P})$ admit no solutions with $\mathcal{U} = \mathcal{U}(t)$.

It is evident that when recombination is absent time variation is also absent, and vice versa. The solutions reduce in this case to:

$$\mathcal{U} = C \qquad (C\text{: arbitrary constant}) \tag{45}$$

$$\mathcal{P} = c(x, y, z). \tag{46}$$

### H. Solutions with $\mathcal{P} = \mathcal{P}(t)$, $N \neq 0$

The condition

$$\operatorname{grad} \mathcal{P} = 0 \tag{47}$$

yields from (21) and (22)

$$\left(N + \frac{\alpha}{\beta}\mathcal{P}\right) \operatorname{div} \operatorname{grad} \mathcal{U} = 0 \tag{48}$$

and

$$\mathcal{P} \operatorname{div} \operatorname{grad} \mathcal{U} = \beta\left[\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{d\mathcal{P}}{dt}\right]. \tag{49}$$

Two cases arise for $\mathcal{R} \not\equiv 0$:

*Case 1:*

$$\mathcal{P} = -\frac{\beta N}{\alpha} \tag{50}$$

and

$$\operatorname{div} \operatorname{grad} \mathcal{U} = -\frac{\alpha}{N}\mathcal{R}(\mathcal{P})\Big|_{\mathcal{P} = -\frac{\beta N}{\alpha}} \tag{51}$$

*Case 2:*

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{d\mathcal{P}}{dt} = 0$$

or

$$D - t = \int \frac{d\mathcal{P}}{2\mathcal{R}(\mathcal{P})} \qquad (D\text{: arbitrary constant}) \tag{52}$$

and

$$\operatorname{div} \operatorname{grad} \mathcal{U} = 0. \tag{53}$$

When recombination is absent, these cases reduce to:

$$\mathcal{P} = E \qquad (E\text{: arbitrary constant}) \tag{54}$$

and (53).

When time variation is absent, Case 2 again yields (53) and (54).

It should be recalled that $\mathcal{U}$ can depend on $t$ as well as $x$, $y$, $z$; so that arbitrary functions of $t$ play the role of arbitrary constants in (51) and (53), whenever time variation is allowed.

### I. Solutions with $P = P(t)$, $N = 0$

For $N = 0$, only Case 2 of the previous section occurs, because the condition $\mathcal{P} = 0$ (implying no carriers!) is of no interest.

### J. Solutions with $\mathfrak{C} = \mathfrak{C}(t)$, $N \neq 0$

For grad $\mathfrak{C} = 0$, (30) and (32) yield:

$$\mathfrak{R}(\mathfrak{u}) + \frac{eN}{2kT}\frac{\partial \mathfrak{u}}{\partial t} = 0 \tag{55}$$

and

$$\text{div grad } \mathfrak{u}^2 = 2 \text{ div } \mathfrak{u} \text{ grad } \mathfrak{u} = 0. \tag{56}$$

Taking the div grad of (55) multiplied by $\mathfrak{u}$ we obtain

$$\text{div grad } \mathfrak{u}\mathfrak{R}(\mathfrak{u}) = 0$$

whence, because of (56)

$$\frac{4kT}{eN}\,\mathfrak{u}\mathfrak{R}(\mathfrak{u}) = F\mathfrak{u}^2 + G \qquad (F, G\text{: arbitrary constants})$$

or

$$\frac{4kT}{eN}\,\mathfrak{R}(\mathfrak{u}) = F\mathfrak{u} + G\mathfrak{u}^{-1}. \tag{57}$$

Substituting this permitted form for the recombination rate function into (55) we obtain

$$\frac{\partial \mathfrak{u}^2}{\partial t} + F\mathfrak{u}^2 = -G \tag{58}$$

whence

$$\mathfrak{u} = \sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \qquad (F \neq 0) \tag{59a}$$

or

$$\mathfrak{u} = \sqrt{f(x, y, z) - Gt} \qquad (F = 0). \tag{59b}$$

From (56), $f(x, y, z)$ is subject to

$$\text{div grad } f(x, y, z) = 0. \tag{60}$$

In summary, if and only if $\mathfrak{R}(\mathfrak{u})$ has the form (57), there are solutions for which $\mathfrak{C} = \mathfrak{C}$ (t) (arbitrary). The $\mathfrak{u}$ is given by (59) in which $f$ is an arbitrary harmonic function of $x$, $y$, $z$.

In terms of $\mathcal{P}$ and $\mathcal{U}$ these solutions are given by:

$$\mathcal{R}(\mathcal{P}) = \frac{F}{4}\mathcal{P} + \left(\frac{eN}{kT}\right)^2 \frac{G}{4}\mathcal{P}^{-1}, \tag{61}$$

$$\mathcal{P} = \frac{eN}{kT}\sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \qquad (F \neq 0) \tag{62a}$$

or

$$\mathcal{P} = \frac{eN}{kT}\sqrt{f(x, y, z) - Gt} \qquad (F = 0), \tag{62b}$$

and

$$\mathcal{U} = \mathcal{K}(t) + \sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \quad (F \neq 0) \tag{63a}$$

or

$$\mathcal{U} = \mathcal{K}(t) + \sqrt{f(x, y, z) - Gt} \qquad (F = 0). \tag{63b}$$

For no recombination ($\mathcal{R} \equiv 0$), these results specialize to (59b), (60), (62b), and (63b) with $G$ set equal to zero. It should be noted (see (55)) that absence of time variation implies absence also of recombination.

### K. Solutions with $\mathcal{U} = \mathcal{U}(\mathcal{P}, t)$, grad $\mathcal{P} \neq 0$

In Theorems 3 and 8 of Section B we have shown that some very interesting properties are implied by the condition

$$\text{grad } \mathcal{U} \times \text{grad } \mathcal{P} = 0. \tag{64}$$

In sections G–I we have treated the cases grad $\mathcal{U} = 0$ and grad $\mathcal{P} = 0$. We now turn to the remaining possibility leading to (64):

$$\mathcal{U} = \mathcal{U}(\mathcal{P}, t) \text{ with grad } \mathcal{P} \neq 0. \tag{65}$$

Substitution of (65) into (23b) leads to

$$\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e}\right] \text{div grad } \mathcal{P} + \frac{\partial}{\partial \mathcal{P}}\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e}\right](\text{grad } \mathcal{P})^2 = 0. \tag{66}$$

Two cases arise.

*Case 1:*

$$\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} = 0.$$

This condition clearly satisfies (66) and leads to

$$\mathcal{U}(\mathcal{P}, t) = g(t) + \frac{\beta kT}{\alpha e} \ln \left| \mathcal{P} + \frac{\beta N}{\alpha} \right| \quad (g(t)\text{: arbitrary function}). \tag{67}$$

The restriction on $\mathcal{P}$ is then provided by the result of substituting (67) into (21):

$$\text{div grad} \left[ \mathcal{P} - \frac{\beta N}{\alpha} \ln \left| \mathcal{P} + \frac{\beta N}{\alpha} \right| \right] = \alpha \left[ \mathcal{R}(\mathcal{P}) + \frac{1}{2} \frac{\partial \mathcal{P}}{\partial t} \right]. \tag{68}$$

Any $\mathcal{P}$ satisfying (68) constitutes with (67) a solution having the property desired.

If (65) is substituted into (25) it will be found that the condition

$$\left( N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} = 0$$

is equivalent to $\overset{\circ}{||} = 0$, so that Case 1 is characterized by zero total current.

*Case 2:*

$$\left( N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} \neq 0.$$

In this case (66) can be written in the form

$$\frac{\text{div grad } \mathcal{P}}{(\text{grad } \mathcal{P})^2} = - \frac{\partial}{\partial \mathcal{P}} \ln \left[ \left( N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} \right] = \phi(\mathcal{P}, t). \tag{69}$$

From (69) it follows that $\mathcal{P}$ must be of the form $\mathcal{P}(h, t)$ with

$$\text{div grad } h(x, y, z, t) = 0. \tag{70}$$

In summary we have

*Theorem 10:* If $\mathcal{U} = \mathcal{U}(\mathcal{P}, t)$ with grad $\mathcal{P} \neq 0$, then either $\overset{\circ}{||} = 0$ or $\mathcal{U} = \mathcal{U}(h, t)$ and $\mathcal{P} = \mathcal{P}(h, t)$ with div grad $h(x, y, z, t) = 0$.

We shall investigate the restrictions on the functions $h(x, y, z, t)$, $\mathcal{U}(h, t)$, and $\mathcal{P}(h, t)$ in the next two sections.

Theorem 10 remains unchanged if recombination is absent. If time variation is absent, it simply drops $t$ as a variable in the functions mentioned in the theorem. If both recombination and time variation are absent, the theorem can be strengthened to:

*Theorem 11:* If both recombination and time variation are absent and $\mathcal{U} = \mathcal{U}(\mathcal{P})$, then $\mathcal{U} = \mathcal{U}(h)$ and $\mathcal{P} = \mathcal{P}(h)$ with div grad $h(x, y, z) = 0$.

## L. Solutions with $\mathfrak{V} = \mathfrak{V}(h, t)$, $\mathfrak{P} = \mathfrak{P}(h, t)$, grad $\mathfrak{P} \neq 0$, div grad $h = 0$, $N \neq 0$

For formal reasons we shall work, not with the conditions $\mathfrak{P} = \mathfrak{P}(h, t)$ and $\mathfrak{V} = \mathfrak{V}(h, t)$, but with the equivalent conditions

$$\mathfrak{U} = \mathfrak{U}(h, t) \text{ and } \mathfrak{K} = \mathfrak{K}(h, t). \tag{71}$$

The condition grad $\mathfrak{P} \neq 0$ now implies $\dfrac{\partial \mathfrak{U}}{\partial h} \neq 0$.

Substitution of (79) into (30) and (32) yields—after use of (70):

$$\frac{\partial^2 \mathfrak{K}}{\partial h^2} (\text{grad } h)^2 = -\frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial h}\frac{\partial h}{\partial t}\right] \tag{72}$$

and

$$\frac{\partial}{\partial h}\left(\left[\frac{\beta kT}{\alpha e} + \mathfrak{U}\right]\frac{\partial \mathfrak{K}}{\partial h} + \mathfrak{U}\frac{\partial \mathfrak{U}}{\partial h}\right) = 0. \tag{73}$$

From (73) we get

$$\frac{\partial \mathfrak{K}}{\partial h} = \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial h}}{\dfrac{\beta kT}{\alpha e} + \mathfrak{U}} \tag{74}$$

($j(t)$: arbitrary function)

which yields upon substitution into (72):

$$\frac{\partial}{\partial h}\left[\frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial h}}{\dfrac{\beta kT}{\alpha e} + \mathfrak{U}}\right](\text{grad } h)^2 = -\frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\left(\frac{\partial \mathfrak{U}}{\partial h}\frac{\partial h}{\partial t} + \frac{\partial \mathfrak{U}}{\partial t}\right)\right] \tag{75}$$

in which $\mathfrak{U}$, $\dfrac{\partial \mathfrak{U}}{\partial h}$, and $\dfrac{\partial^2 \mathfrak{U}}{\partial h^2}$ are, of course, functions of $h$ and of $t$.

In determining the combined implications of (75) and (70) three cases arise according to whether or not $\dfrac{\partial^2 \mathfrak{K}}{\partial h^2} = 0$ or grad $(\text{grad } h)^2 = 0$.

*Case 1:*

$$\frac{\partial^2 \mathfrak{K}}{\partial h^2} \neq 0, \qquad \text{grad } (\text{grad } h)^2 \neq 0.$$

In this case no satisfactory interpretation has been found when time variability is present.

When time variation is absent, we work with the conditions

$$\mathfrak{U} = \mathfrak{U}(h); \quad \mathfrak{IC} = \mathfrak{IC}(h)$$

with

$$\text{div grad } h(x, y, z) = 0 \tag{76}$$

and arrive at counterparts of (74) and (75):

$$\mathfrak{IC}' = \frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}} \qquad \left(\gamma \equiv \frac{\beta k T}{\alpha e}\right) \qquad (H\text{: arbitrary constant}) \tag{77}$$

and

$$\mathfrak{IC}'' (\text{grad } h)^2 = \left(\frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' (\text{grad } h)^2 = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U}). \tag{78}$$

From (78) it is evident that $\mathfrak{R} \neq 0$ implies $\mathfrak{IC}'' \neq 0$ and grad $h \neq 0$. So we have

$$(\text{grad } h)^2 = \frac{-\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U})}{\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)'}, \tag{79}$$

which is of the form

$$[\text{grad } h(x, y, z)]^2 = \phi(h). \tag{79a}$$

Now from (79a) follows

$$\text{grad } h \times \text{grad } (\text{grad } h)^2 = 0 \tag{80}$$

which implies that the vector lines of the field grad $h$ are all straight. Since $h$ is harmonic, this restricts the choice of $h$ to the potential fields associated with a uniform parallel flow, a straight line source, or a point source. Hence, for suitably chosen rectangular coordinates $(x, y, z)$, circular cylindrical coordinates $(\rho, \theta, z)$ or spherical polar coordinates $(r, \theta, \phi)$, the only possibilities, are respectively

$$h = x \rightarrow (\text{grad } h)^2 = 1 \tag{81a}$$

or

$$h = \ln \frac{1}{\rho} \rightarrow (\text{grad } h)^2 = \frac{1}{\rho^2} = \epsilon^{2h} \tag{81b}$$

or

$$h = \frac{1}{r} \rightarrow (\text{grad } h)^2 = \frac{1}{r^4} = h^4. \tag{81c}$$

The possibility $h = x$ violates one defining condition for the present case (i.e., grad (grad $h)^2 \not\equiv 0$) and hence will be left for consideration in Case 3. The remaining two possibilities lead respectively to the following forms of ordinary differential equation for the determination of $\mathfrak{U}(h)$:

$$\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' + \frac{\alpha}{N} \epsilon^{-2h} \mathfrak{R}(\mathfrak{U}) = 0 \tag{82b}$$

or

$$\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' + \frac{\alpha}{N} \frac{1}{h^4} \mathfrak{R}(\mathfrak{U}) = 0. \tag{82c}$$

Given any $\mathfrak{U}(h)$ satisfying one of these equations, the associated $\mathfrak{IC}(h)$ is obtained by integration from (77):

$$\mathfrak{IC}(h) = \int \left(\frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right) dh + J \qquad (J: \text{arbitrary constant}). \tag{83}$$

It is evident from (72) that Case 1 does not exist if both recombination and time variation are absent.

*Case 2:*

$$\frac{\partial^2 \mathfrak{IC}}{\partial h^2} = 0$$

In considering this case we shall exclude the condition $\frac{\partial \mathfrak{IC}}{\partial h} = 0$ because it has been included in Section J.

From the condition $\frac{\partial^2 \mathfrak{IC}}{\partial h^2} = 0$ we have

$$\mathfrak{IC} = k(t)h + \ell(t) \qquad (k(t), \ell(t): \text{arbitrary functions}) \tag{84}$$

with $k \not\equiv 0$. This shows that $\mathfrak{IC}$ itself is a harmonic function and we can without loss of generality use it in place of $h$.

Equations (74) and (75) now yield the two conditions on $\mathfrak{U}(\mathfrak{IC}, t)$, $\mathfrak{R}(\mathfrak{U})$, and $\mathfrak{IC}(x, y, z, t)$:

$$\frac{j(t) - \mathfrak{U} \frac{\partial \mathfrak{U}}{\partial \mathfrak{IC}}}{\mathfrak{U} + \gamma} = 1 \tag{85}$$

and

$$\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\left(\frac{\partial \mathfrak{U}}{\partial \mathfrak{K}}\frac{\partial \mathfrak{K}}{\partial t} + \frac{\partial \mathfrak{U}}{\partial t}\right) = 0. \tag{86}$$

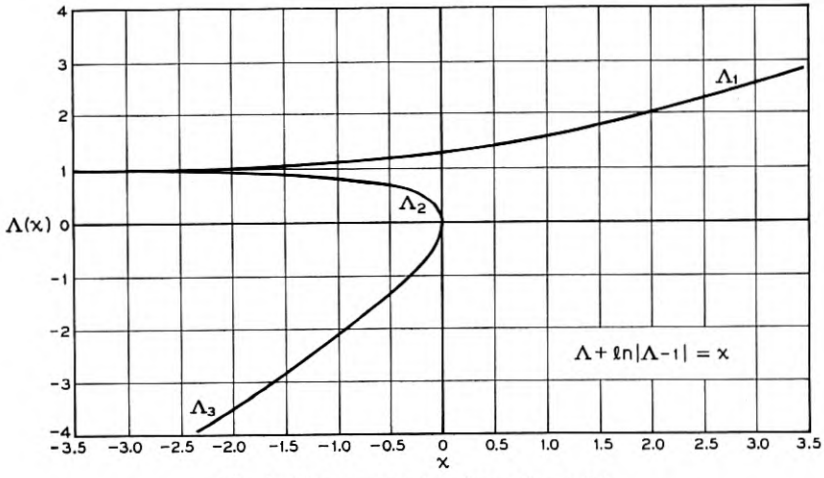



Fig. 1—The transcendental function $\Lambda\ (x)$.

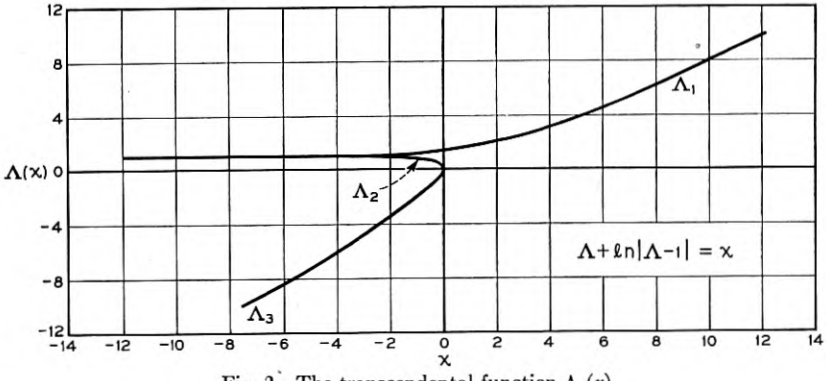



Fig. 2—The transcendental function $\Lambda\ (x)$.

For the integration of (85) we need the transcendental algebraic function of a single real variable defined by

$$\Lambda(x) + \ln \mid \Lambda(x) - 1 \mid = x. \tag{87}$$

This function is plotted in Figs. 1 and 2. It will be observed that $x$ is always a single-valued function of $\Lambda$; while $\Lambda$ is a single-valued function of $x$ for $x > 0$, a double-valued function for $x = 0$, and a triple-valued function for

$x < 0$. The single-valued monotone functions $\Lambda_1$, $\Lambda_2$, and $\Lambda_3$ are defined respectively by the restrictions $\Lambda > 1$, $1 > \Lambda \geq 0$, and $\Lambda \leq 0$. When $\Lambda$ is used without subscript it is implied that either $\Lambda_1$, $\Lambda_2$, or $\Lambda_3$ can be used. It will be useful to remember that

$$\Lambda'(x) = \frac{\Lambda(x) - 1}{\Lambda(x)}. \tag{88}$$

In terms of the function $\Lambda$, (85) integrates to

$$\mathfrak{U}(\mathfrak{IC}, t) = [j(t) - \gamma]\Lambda\left[\frac{m(t) - \mathfrak{IC}}{j(t) - \gamma}\right] \qquad (j \neq \gamma) \tag{89a}$$

(m(t): arbitrary function)

or

$$\mathfrak{U}(\mathfrak{IC}, t) = m(t) - \mathfrak{IC} \qquad (j = \gamma). \tag{89b}$$

The latter case ($j = \gamma$) corresponds to $\mathfrak{V} = \mathfrak{V}(t)$ and hence was included in Section G. Therefore in the following we shall consider only $j \neq \gamma$.

Now by making use of (89a) and (88), (86) can be rewritten in the form:

$$\frac{2kT}{eN}\frac{\mathfrak{U}\mathfrak{R}(\mathfrak{U})}{\mathfrak{U} - j + \gamma} + \frac{j'\mathfrak{U}^2}{(j - \gamma)(\mathfrak{U} - j + \gamma)} = \frac{\partial \mathfrak{IC}}{\partial t} + \frac{j'(m - \mathfrak{IC})}{j - \gamma} - m' \tag{90}$$

$$\left(\text{primes denoting here } \frac{d}{dt}\right).$$

We now observe that the right side of (90) is harmonic, while the left side is a function only of $\mathfrak{IC}$ and $t$. From this it follows that the right side can be written in the form:

$$\frac{\partial \mathfrak{IC}}{\partial t} + \frac{j'(m - \mathfrak{IC})}{j - \gamma} - m' = q(t)\left[\frac{m - \mathfrak{IC}}{j - \gamma}\right] + r(t). \tag{91}$$

From (90), (91) and (89a) follows

$$\frac{2kT}{eN}\mathfrak{R}(\mathfrak{U}) = -\frac{j'}{j - \gamma}\mathfrak{U} + \frac{\mathfrak{U} - j + \gamma}{\mathfrak{U}}\left(r + \frac{q}{j - \gamma}\mathfrak{U} + q\ln\left|\frac{\mathfrak{U}}{j - \gamma} - 1\right|\right). \tag{92}$$

Since (92) is of the form

$$\mathfrak{R}(\mathfrak{U}) = \phi(\mathfrak{U}, t),$$

The result of taking $\left(\frac{\partial}{\partial t}\right)_{\mathfrak{U} = \text{const}}$ of the right side must be zero identi-

cally in $\mathfrak{U}$. Making use of the algebraic lemma that

$$A + Bx + \frac{C}{x} + \left(D + \frac{E}{x}\right) \ln \left| \frac{x}{F} - 1 \right| \equiv 0$$

implies $A=B=C=D=E=0$, we arrive at two possibilities:

*Possibility 1:*

$$(q(t) = r(t) = 0 \quad \text{and} \quad j - \gamma = K\epsilon^{-Lt})$$

($K$, $L$: arbitrary constants)

This yields from (92), (89a) and (91):

$$\frac{2kT}{eN} \mathfrak{R}(\mathfrak{U}) = L\mathfrak{U} \tag{93}$$

$$\mathfrak{U}(\mathfrak{K}, t) = K\epsilon^{-Lt} \Lambda \left[ \frac{1}{K} \epsilon^{Lt} (m(t) - \mathfrak{K}) \right] \tag{94}$$

and

$$\mathfrak{K}(x, y, z, t) = \epsilon^{-Lt} s(x, y, z) + \epsilon^{-Lt} \int \epsilon^{Lt} \left[ Lm(t) + m'(t) \right] dt \tag{95}$$

where $s(x, y, z)$ is any harmonic function.

*Possibility, 2:*

$$(j(t) = M, q(t) = Q, r(t) = R)$$

($M$, $Q$, $R$: arbitrary constants)

This yields from (92), (89a) and (91):

$$\frac{2kT}{eN} \mathfrak{R}(\mathfrak{U}) = \frac{\mathfrak{U} - M + \gamma}{\mathfrak{U}} \cdot \left( R + \frac{Q}{M - \gamma} \mathfrak{U} + Q \ln \left| \frac{\mathfrak{U}}{M - \gamma} - 1 \right| \right) \tag{96}$$

$$\mathfrak{U}(\mathfrak{K}, t) = (M - \gamma) \Lambda \left[ \frac{m(t) - \mathfrak{K}}{M - \gamma} \right] \tag{97}$$

and

$$\mathfrak{K}(x, y, z, t) = \epsilon^{Qt/(M-\gamma)} u(x, y, z) + \epsilon^{Qt/(M-\gamma)} \int \epsilon^{-(Qt/(M-\gamma))} \left[ R + m'(t) + \frac{Q}{M - \gamma} m(t) \right] dt \tag{98}$$

where $u(x, y, z)$ is any harmonic function.

In the absence of recombination, Possibilities 1 and 2 lead to the same result: Equation (97) and

$$\mathfrak{K}(x, y, z, t) = u(x, y, z) + m(t). \tag{99}$$

In the absence of time variation, (86) shows that recombination is necessarily absent, too, so the results reduce to

$$\mathfrak{u}(\mathfrak{K}) = \tilde{A}\Lambda\left(\frac{\tilde{B} - \mathfrak{K}}{\tilde{A}}\right) \tag{100}$$

with $\mathfrak{K}(x, y, z)$ any harmonic function and $\tilde{A}$ and $\tilde{B}$ arbitrary constants. This solution for the case grad $\mathfrak{K} \neq 0$, together with that given by (59b) and (60) (with $G = 0$) for the case grad $\mathfrak{K} = 0$, constitute a veritable gold mine of useful solutions because of the arbitrary harmonic function involved. An example involving a particular choice of $\mathfrak{K}$ will be examined in Section R.

*Case 3:*

$$\frac{\partial^2 \mathfrak{K}}{\partial h^2} \neq 0, \qquad \text{grad (grad h)}^2 = 0.$$

In this case $(\text{grad } h)^2$ is a function of $t$ so that (75) can be written in the form

$$\frac{\partial h}{\partial t} = \phi(h, t)$$

From this it follows (because div grad $h = 0$) that

$$h(x, y, z, t) = \tilde{a}(t)\tilde{b}(x, y, z) + \tilde{c}(t) \tag{101}$$

with

$$\text{div grad } \tilde{b}(x, y, z) = 0. \tag{102}$$

The condition grad $(\text{grad } h)^2 = 0$ now requires further that

$$\text{grad (grad } \tilde{b})^2 = 0. \tag{103}$$

But any $\tilde{b}(x, y, z)$ satisfying both (100) and (101) can, by suitable choice of axes, be written

$$\tilde{b} = Sx \qquad (S\text{: constant}).$$

This leaves us with exactly the same totality of solutions as we could have obtained by setting $\mathfrak{U} = \mathfrak{U}(x, t)$, $\mathfrak{K} = \mathfrak{K}(x, t)$ in the first place. So we replace $h$ by $x$ in (74) and (75) and obtain:

$$\frac{\partial \mathfrak{K}}{\partial x} = \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}} \tag{104}$$

and

$$\frac{\partial}{\partial x}\left[\frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}}\right] + \frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right] = 0. \tag{105}$$

Any $\mathfrak{U}(x, t)$ satisfying (105) can be substituted into (104) to obtain $\mathfrak{IC}(x, t)$ from

$$\mathfrak{IC}(x, t) = \tilde{f}(t) + \int^{(x)} \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}}\, dx \tag{106}$$

($\tilde{f}(t)$: arbitrary function).

If recombination is absent, $\mathfrak{R}(\mathfrak{U})$ disappears from (105). If time variation is absent, $\dfrac{\partial \mathfrak{U}}{\partial t}$ disappears from (103) and $j(t)$ and $\tilde{f}(t)$ are replaced by arbitrary constants. In the latter case, the standard change of variables

$$\mathfrak{W}(\mathfrak{U}) \quad \text{for} \quad \frac{d\mathfrak{U}}{dx} \qquad \mathfrak{W}(\mathfrak{U})\frac{d}{d\mathfrak{U}} \quad \text{for} \quad \frac{d}{dx} \tag{107}$$

reduces the solution of the second order equation (105) to the solution of a first-order equation followed by a quadrature. If both recombination and time variation are absent, the substitution (107) reduces the solution of (105) to two quadratures.

A set of equations equivalent to the steady-state $\left(\dfrac{\partial}{\partial t} \equiv 0\right)$ forms of (104) and (105) has been the subject of an extensive numerical investigation by W. van Roosbroeck (Reference 1) for the recombination rate functions given in (37) and (38).

### M. Solutions with $\mathcal{U} = \mathcal{U}(h, t)$, $\mathcal{P} = \mathcal{P}(h, t)$, Grad $\mathcal{P} \neq 0$, Div Grad $h = 0$, $N = 0$

For these conditions (21) and (23b) yield

$$\frac{\partial^2 \mathcal{P}}{\partial h^2}(\text{grad } h)^2 = \frac{\alpha e}{kT}\left[\mathfrak{R}(\mathcal{P}) + \frac{1}{2}\left(\frac{\partial \mathcal{P}}{\partial h}\frac{\partial h}{\partial t} + \frac{\partial \mathcal{P}}{\partial t}\right)\right] \tag{107}$$

and

$$\frac{\partial}{\partial h}\left[\frac{\partial \mathcal{P}}{\partial h} - \frac{\alpha e}{\beta kT}\mathcal{P}\frac{\partial \mathcal{U}}{\partial h}\right](\text{grad } h)^2 = 0. \tag{108}$$

Since we do not here allow grad $h = 0$, (108) implies

$$\frac{\partial \mathcal{U}}{\partial h} = \frac{\gamma\left[\dfrac{\partial \mathcal{P}}{\partial h} - \bar{g}(t)\right]}{\mathcal{P}} \qquad \gamma \equiv \frac{\beta kT}{\alpha e} \qquad (\bar{g}(t)\text{: arbitrary function}) \tag{109}$$

*Case 1:*

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} \neq 0, \qquad \text{grad (grad } h)^2 \neq 0.$$

In this case, as in the associated case in Section L, the implications of (107) together with

$$\text{div grad } h(x, y, z, t) = 0$$

are not known when time variation is present.

When time variation is absent, we work with the conditions

$$\mathcal{P} = \mathcal{P}(h) \quad \text{and} \quad \mathcal{U} = \mathcal{U}(h)$$

with

$$\text{div grad } h(x, y, z) = 0$$

and arrive at counterparts of (107) and (108):

$$\mathcal{P}'' \cdot (\text{grad } h)^2 = \frac{\alpha e}{kT} \mathcal{R}(\mathcal{P}) \tag{110}$$

and

$$\left(\mathcal{P}' - \frac{1}{\gamma} \mathcal{P}\mathcal{U}'\right)' = 0. \tag{111}$$

Proceeding as in the analysis of Case 1 of Section L, we infer that $h$ must be of the kind given by (81b) or (81c). The associated second-order differential equations restricting $\mathcal{P}(h)$ are then, respectively:

$$\mathcal{P}'' - \frac{\alpha e}{kT} \epsilon^{-2h} \mathcal{R}(\mathcal{P}) = 0 \tag{112a}$$

and

$$\mathcal{P}'' - \frac{\alpha e}{kT} \frac{1}{h^4} \mathcal{R}(\mathcal{P}) = 0. \tag{112b}$$

The $\mathcal{U}(h)$ associated with any solution of (112) can be obtained by integration from

$$\mathcal{U}(h) = \tilde{C} + \int \frac{\gamma \mathcal{P}' - \tilde{D}}{\mathcal{P}} \, dh \qquad (\tilde{C}, \tilde{D}\text{: arbitrary constants}). \tag{113}$$

It will be noted from (110) that simultaneous absence of recombination and time variation is inconsistent with the defining conditions of this case.

*Case 2:*

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} = 0.$$

We shall exclude the possibility of $\frac{\partial \mathcal{P}}{\partial h} = 0$ because it is included in Section I. Then, proceeding as in Case 2 of Section L, we conclude that $\mathcal{P}$ itself is a harmonic function and can be used in place of $h$. (107) and (109) then become·

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t} = 0 \tag{114}$$

and

$$\frac{\partial \mathcal{U}}{\partial \mathcal{P}} = \frac{\gamma[1 - \tilde{g}(t)]}{\mathcal{P}} \tag{115}$$

or

$$\mathcal{U}(\mathcal{P}, t) = \gamma[1 - \tilde{g}(t)] \ln \mathcal{P} + \tilde{\jmath}(t) \qquad (\tilde{\jmath}(t)\text{: arbitrary function}). \tag{116}$$

Because $\frac{\partial \mathcal{P}}{\partial t}$ is harmonic and a function of $\mathcal{P}$, we have

$$2\mathcal{R}(\mathcal{P}) = -\frac{\partial \mathcal{P}}{\partial t} = E\mathcal{P} - \tilde{F}$$

($\tilde{E}$, $\tilde{F}$: arbitrary constants)

whence

$$2\mathcal{R}(\mathcal{P}) = \tilde{E}\mathcal{P} - \tilde{F} \tag{117}$$

and

$$\mathcal{P}(x, y, z, t) = \epsilon^{-\tilde{E}t}\tilde{m}(x, y, z) - \frac{\tilde{F}}{\tilde{E}} \qquad (E \neq 0) \tag{118a}$$

or

$$\mathcal{P}(x, y, z, t) = \tilde{m}(x, y, z) + \tilde{F}t \qquad (\tilde{E} = 0) \tag{118b}$$

where $\tilde{m}(x, y, z)$ is an arbitrary *harmonic* function.

If recombination is absent, these results specialize to (116) and (118b)

with $\tilde{F} = 0$. If time variation is absent it follows from (114) that recombination is absent, too, and the results specialize to

$$\mathcal{U}(\mathcal{P}) = \tilde{G} + \tilde{H} \ln \mathcal{P} \tag{119}$$

with $\mathcal{P}(x, y, z)$ any harmonic function and $\tilde{G}$ and $\tilde{H}$ arbitrary constants. These solutions play the same role for the intrinsic semiconductor ($N = 0$) that (100) does for the extrinsic ($N \neq 0$).

*Case 3:*

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} \not\equiv 0, \qquad \text{grad (grad } h)^2 = 0.$$

In this case it can be shown, just as in Case 3 of Section L, that no generality is lost by considering $\mathcal{P} = \mathcal{P}(x, t)$ and $\mathcal{U} = \mathcal{U}(x, t)$ in place of $\mathcal{P}(h, t)$ and $\mathcal{U}(h, t)$. Equations (107) and (109) then become

$$\frac{\partial^2 \mathcal{P}}{\partial x^2} = \frac{\alpha e}{kT}\left[\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right] \tag{120}$$

and

$$\frac{\partial \mathcal{U}}{\partial x} = \frac{\gamma\left[\dfrac{\partial \mathcal{P}}{\partial x} - \tilde{g}(t)\right]}{\mathcal{P}}. \tag{121}$$

Any solution of (120) when substituted into (121) gives an associated $\mathcal{U}$ from

$$\mathcal{U}(x, t) = \tilde{q}(t) + \gamma \int^{(x)} \frac{\dfrac{\partial \mathcal{P}}{\partial x} - \tilde{g}(t)}{\mathcal{P}}\, dx. \tag{122}$$

If recombination is absent, $\mathcal{R}(\mathcal{P})$ merely vanishes from (120). If time variation is absent, the functions $\tilde{g}(t)$ and $\tilde{q}(t)$ are replaced by arbitrary constants and the standard change of variables

$$\begin{aligned} \mathbf{u}(\mathcal{P}) \quad &\text{for} \quad \frac{d\mathcal{P}}{dx} \\ \mathbf{u}(\mathcal{P})\frac{d}{d\mathcal{P}} \quad &\text{for} \quad \frac{d}{dx} \end{aligned} \tag{123}$$

leads to a solution of (120) in two quadratures. An equivalent solution is given by W. van Roosbroeck in Reference 1. From (120) it follows that recombination and time variation cannot simultaneously be absent for Case 3.

## N. Construction of Solutions from Orthogonal Harmonic Fields, $N \neq 0$

There are many known examples of pairs of harmonic functions $h_1(x, y, z)$ and $h_2(x, y, z)$ that have orthogonal vector fields—that is, for which

$$\text{grad } h_1 \cdot \text{grad } h_2 = 0 \tag{124}$$

with grad $h_1 \neq 0$ and grad $h_2 \neq 0$. [E.g., the real and imaginary parts of any analytic function of a complex variable.] From any such pair of functions we can construct the following solutions of (33) and (34):

$$\mathfrak{U} = h_1 ; \qquad \mathfrak{IC} = h_2 - h_1 \tag{125}$$

and

$$\mathfrak{U} = \sqrt{h_1} ; \qquad \mathfrak{IC} = h_2 . \tag{126}$$

In terms of $\mathcal{P}$ and $\mathcal{U}$ these solutions are

$$\mathcal{P} = \frac{Ne}{kT} h_1 ; \qquad \mathcal{U} = h_2 \tag{127}$$

and

$$\mathcal{P} = \frac{Ne}{kT} \sqrt{h_1} ; \qquad \mathcal{U} = \sqrt{h_1} + h_2 . \tag{128}$$

The validity of the solution (125) is seen from (33) and this expanded form of (34):

$$\mathfrak{U} \text{ div grad } \mathfrak{U} + \text{grad } \mathfrak{U} \cdot \text{grad } (\mathfrak{U} + \mathfrak{IC}) = 0. \tag{129}$$

Similarly, the validity of (126) follows from (33) together with a different expansion of (34):

$$\text{div grad } \mathfrak{U}^2 + 2 \text{ grad } \mathfrak{U} \cdot \text{grad } \mathfrak{IC} = 0. \tag{130}$$

It is evident that a given $h_1$ and $h_2$ can be interchanged in the above solutions to yield different solutions, and also that any given $h_1$ or $h_2$ can be replaced by an arbitrary constant multiple of itself plus a second arbitrary constant.

## O. Construction of Solutions from Orthogonal Harmonic Fields, $N = 0$

We can write the differential equation system for the intrinsic semiconductor [(35) and (36)] in the form:

$$\text{div grad } \mathcal{P} = 0 \tag{131}$$

$$\mathcal{P} \text{ div grad } \mathcal{U} + \text{grad } \mathcal{P} \cdot \text{grad } \mathcal{U} = 0. \tag{132}$$

From these we verify the solution:

$$\mathcal{P} = h_1 ; \qquad \mathcal{U} = h_2 \tag{133}$$

for any harmonic $h_1$ and $h_2$ satisfying (124).

The solutions given by (127) and (133) have the property

$$\text{grad } \mathcal{P} \cdot \text{grad } \mathcal{U} = 0$$

and so may be considered, in a sense, complementary to the solutions in Sections L and M for which

$$\text{grad } \mathcal{P} \times \text{grad } \mathcal{U} = 0.$$

## P. Superposition of a Harmonic $\mathcal{C}$ Field, $N \neq 0$

Inspection of the equation system [(33), (130)] reveals the following superposition theorem for obtaining new solutions from some known solutions *for the case of no recombination or time variation:*

*Theorem 12:* If $[\tilde{\mathfrak{U}}, \tilde{\mathcal{C}}]$ is a known solution and if $h$ is any harmonic function such that $\text{grad } \tilde{\mathfrak{U}} \cdot \text{grad } h = 0$, then $[\mathfrak{U}, \tilde{\mathcal{C}} + h]$ is also a solution.

Or, in terms of $\mathcal{P}$ and $\mathcal{U}$:

*Theorem 12′:* If $[\tilde{\mathcal{P}}, \tilde{\mathcal{U}}]$ is a known solution and if $h$ is any harmonic function such that $\text{grad } \tilde{\mathcal{P}} \cdot \text{grad } h = 0$, then $[\tilde{\mathcal{P}}, \tilde{\mathcal{U}} + h]$ is also a solution.

In the latter form it is evident from Section O that the theorem holds also for $N = 0$, but does not extend the results of Section O.

## Q. A Partial Differential Equation in Terms of $\mathcal{C}$ Alone, $N \neq 0$

For $N = 0$, (21) provides a differential equation involving only one dependent variable—$\mathcal{P}$. We shall now derive an analogous—but vastly more complicated—differential equation *for the case* $N \neq 0$, $\dfrac{\partial \mathfrak{U}}{\partial t} = 0$.

For this case (30) and (32) become

$$\text{div grad } \mathcal{C} = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U})$$

and

$$\text{div}\left[\text{grad } \mathcal{C} + \frac{1}{\gamma} \mathfrak{U} \text{ grad } (\mathfrak{U} + \mathcal{C})\right] = 0,$$

or in terms of a familiar vector symbolism

$$\nabla^2 \mathfrak{IC} = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U}) \tag{134}$$

and

$$\nabla \cdot \left[ \nabla \mathfrak{IC} + \frac{1}{\gamma} \mathfrak{U} \nabla (\mathfrak{U} + \mathfrak{IC}) \right] = 0. \tag{135}$$

Now let $\mathfrak{S}(\mathfrak{U})$ be the inverse function to $\mathfrak{R}(\mathfrak{U})$, i.e. the function such that

$$\mathfrak{S}(\mathfrak{R}(\mathfrak{U})) \equiv \mathfrak{U}.$$

Then from (134) we have

$$\mathfrak{U} = \mathfrak{S}\left( -\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right). \tag{136}$$

Substitution of (136) into (135) yields after some computation

$$\mathfrak{S}\mathfrak{S}' \nabla^2 (\nabla^2 \mathfrak{IC}) - \frac{N}{\alpha} (\mathfrak{S}\mathfrak{S}'' + \mathfrak{S}'^2)(\nabla \nabla^2 \mathfrak{IC})^2 + \mathfrak{S}' \nabla \mathfrak{IC} \cdot \nabla \nabla^2 \mathfrak{IC} - \frac{\alpha}{N} (\mathfrak{S} + \gamma) \nabla^2 \mathfrak{IC} = 0 \tag{137}$$

where $\mathfrak{S}'(\psi) \equiv \dfrac{d}{d\psi} \mathfrak{S}(\psi)$, etc.

$\mathfrak{S}$, $\mathfrak{S}'$, $\mathfrak{S}''$ are considered as given functions of $\left( -\dfrac{N}{\alpha} \nabla^2 \mathfrak{IC} \right)$.

The simplest meaningful choice of $\mathfrak{S}$ is

$$\mathfrak{S}\left( -\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right) = \frac{\alpha}{N} \tilde{J} \cdot \left( -\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right) + \tilde{K} \tag{138}$$

($J$, $\tilde{K}$: prescribed constants)

corresponding to constant mean lifetime recombination. For this $\mathfrak{S}$, (137) specializes to

$$\tilde{J}(\tilde{K} - J \nabla^2 \mathfrak{IC}) \nabla^2 (\nabla^2 \mathfrak{IC}) - \tilde{J}^2 (\nabla \nabla^2 \mathfrak{IC})^2 + J \nabla \mathfrak{IC} \cdot \nabla \nabla^2 \mathfrak{IC} - (\gamma + \tilde{K} - J \nabla^2 \mathfrak{IC}) \nabla^2 \mathfrak{IC} = 0. \tag{139}$$

If any $\mathfrak{IC}$ can be found satisfying (139), the associated $\mathfrak{U}$ is given (from (136)) by

$$\mathfrak{U} = \tilde{J} \nabla^2 \mathfrak{IC} + \tilde{K}.$$

## R. Sample Application of the Results of Section L: Spherical Symmetry, $N \neq 0$

As an example of the solutions included in the results of Section L we consider the case of a spherically symmetric field about a point (or spherical) source of current.

We take, as the most general harmonic function having spherical symmetry,

$$\mathfrak{H} = \tilde{L}\frac{1}{r} + \tilde{M} \tag{140}$$

($\tilde{L}$, $\tilde{M}$: arbitrary constants).

For the time being we shall assume $\tilde{L} \neq 0$ and $\tilde{M} \neq 0$. Then from (100) and (28) and (29) we have

$$\mathfrak{V} = \tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M} - \tilde{L}/r}{\tilde{A}}\right) + M + \frac{L}{r} \tag{141}$$

and

$$\mathfrak{P} = \frac{Ne}{kT}\tilde{A}\Lambda\left(\frac{\tilde{B} - M - \tilde{L}/r}{\tilde{A}}\right). \tag{142}$$

In terms of $\mathfrak{V}$ and $\mathfrak{P}$, (3) and (4) can be written

$$\overset{\circ}{|}|_p = \frac{-\mu_p e}{2}\left[(\mathfrak{P} - N)\operatorname{grad}\mathfrak{V} + \frac{kT}{e}\operatorname{grad}\mathfrak{P}\right] \tag{143}$$

and

$$\overset{\circ}{|}|_n = \frac{-\mu_n e}{2}\left[(\mathfrak{P} + N)\operatorname{grad}\mathfrak{V} - \frac{kT}{e}\operatorname{grad}\mathfrak{P}\right] \tag{144}$$

which yield upon substitution of (141) and (142):

$$\overset{\circ}{|}|_p = \frac{1}{2}\mu_p e N\tilde{L}\left(\frac{e}{kT}\tilde{A} - 1\right)\frac{1}{r^2}\mathfrak{r}_1 \tag{145}$$

and

$$\overset{\circ}{|}|_n = \frac{1}{2}\mu_n e N\tilde{L}\left(\frac{e}{kT}\tilde{A} + 1\right)\frac{1}{r^2}\mathfrak{r}_1 \tag{146}$$

where $\mathfrak{r}_1$ is the unit radial vector. The total current density is obtained by adding (151) and (152):

$$\overset{\circ}{|}| = \frac{1}{2}eN\tilde{L}\left[(\mu_n + \mu_p)\frac{e}{kT}\tilde{A} + (\mu_n - \mu_p)\right]\frac{1}{r^2}\mathfrak{r}_1. \tag{147}$$

The currents flowing are obtained from the current densities from the relation

$$I = \Omega r^2 \overset{\circ}{||} \cdot \mathfrak{r}_1$$

where $\Omega$ is the solid angle (with respect to the origin) within which the flow field lies. (If the current source is surrounded by the homogeneous semiconductor, $\Omega = 4\pi$; if it lies on a flat surface of a large slab, $\Omega = 2\pi$, etc.) So we have

$$I_p = \frac{1}{2}\Omega\mu_p eN\tilde{L}\left(\frac{e}{kT}\tilde{A} - 1\right) \tag{148}$$

$$I_n = \frac{1}{2}\Omega\mu_n eN\tilde{L}\left(\frac{e}{kT}\tilde{A} + 1\right) \tag{149}$$

$$I = \frac{1}{2}\Omega eN\tilde{L}\left[(\mu_n + \mu_p)\frac{e}{kT}\tilde{A} + (\mu_n - \mu_p)\right]. \tag{150}$$

We shall now obtain expressions for the mathematical parameters $\tilde{B}$, $\tilde{A}$, $\tilde{L}$, and $\tilde{M}$ in terms of meaningful physical quantities: $I_p$, $I_n$, $\mathcal{V}_\infty$ and $\mathcal{P}_\infty$. (Subscript $\infty$ refers to values of variables as $r$ becomes very large.) We shall take our reference voltage as the voltage "at infinity" so that $\mathcal{V}_\infty = 0$. Setting $1/r = 0$ in (141) and (142) we obtain

$$0 = \tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right) + \tilde{M}$$

and

$$\mathcal{P}_\infty = \frac{Ne}{kT}\tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right)$$

from which follows (for $\tilde{A} \neq 0$)

$$\tilde{B} = \tilde{A}\ln\left|\frac{\tilde{M}}{\tilde{A}} + 1\right| \tag{151}$$

and

$$\tilde{M} = -\frac{kT}{eN}\mathcal{P}_\infty. \tag{152}$$

From (148) and (149) we readily find (for $\tilde{L} \neq 0$):

$$\tilde{A} = \frac{kT}{e}\frac{\mu_p I_n + \mu_n I_p}{\mu_p I_n - \mu_n I_p} \tag{153}$$

and

$$\tilde{L} = \frac{\mu_p I_n - \mu_n I_p}{\Omega \mu_n \mu_p eN}. \tag{154}$$

Finally we substitute (152) and (153) into (154) to get

$$\tilde{B} = \frac{kT}{e} \frac{\mu_p I_n + \mu_n I_p}{\mu_p I_n - \mu_n I_p} \ln \left| \frac{(N - \mathcal{P}_\infty)\mu_p I_n + (N + \mathcal{P}_\infty)\mu_n I_p}{N(\mu_p I_n - \mu_n I_p)} \right| \tag{155}$$

Equations (152)–(155) give the desired expressions for $\tilde{M}$, $\tilde{A}$, $\tilde{L}$ and $\tilde{B}$ in terms of $I_p$, $I_n$, and $\mathcal{P}_\infty$ for $\mathcal{U}_\infty = 0$ if $\tilde{A} \neq 0$ and $\tilde{L} \neq 0$.

*For* $\tilde{A} = 0$ we can repeat the above steps using

$$\mathcal{U} = \tilde{B} \tag{156}$$

and

$$\mathcal{P} = \frac{Ne}{kT}(\tilde{B} - \tilde{M} - \tilde{L}/r) \tag{157}$$

in place of (141) and (142). The result for $\tilde{L} \neq 0$ and $\mathcal{U}_\infty = 0$ is

$$I_p = -\tfrac{1}{2}\Omega \mu_p eN\tilde{L} \tag{158}$$

$$I_n = \tfrac{1}{2}\Omega \mu_n eN\tilde{L} \tag{159}$$

$$I = \tfrac{1}{2}\Omega eN(\mu_n - \mu_p)\tilde{L} \tag{160}$$

with

$$\tilde{B} = 0 \tag{161}$$

$$\tilde{M} = -\frac{kT}{eN}\mathcal{P}_\infty \tag{162}$$

and

$$\tilde{L} = \frac{-2I_p}{\Omega \mu_p eN} = \frac{2I_n}{\Omega \mu_n eN} = \frac{\mu_p I_n - \mu_n I_p}{\Omega \mu_p \mu_n eN}. \tag{163}$$

It is evident that $\tilde{A} = 0$ implies $\mathcal{U} =$ constant and $\mu_p I_n + \mu_n I_p = 0$.

The condition $\tilde{L} = 0$ makes $\mathcal{JC} =$ constant, so we use (62b) and (63b) and set

$$\mathcal{U} = \tilde{Q} + \sqrt{\tilde{R} + \tilde{S}/r} \tag{164}$$

and

$$\mathcal{P} = \frac{Ne}{kT}\sqrt{\tilde{R} + \tilde{S}/r} \qquad (\tilde{Q}, \tilde{R}, \tilde{S}\text{: arbitrary constants}). \tag{165}$$

From (143) and (144) we obtain

$$I_p = -\frac{\Omega N e^2 \mu_p}{4kT}\tilde{S} \tag{166}$$

$$I_n = -\frac{\Omega N e^2 \mu_n}{4kT}\tilde{S} \tag{167}$$

and

$$I = -\frac{\Omega N e^2}{4kT}(\mu_n + \mu_p)\tilde{S}. \tag{168}$$

From (164) and (165) we readily obtain for $\mathfrak{V}_\infty = 0$:

$$R = \left(\frac{kT}{Ne}\mathcal{P}_\infty\right)^2 \tag{169}$$

and

$$\tilde{Q} = -\frac{kT}{Ne}\mathcal{P}_\infty. \tag{170}$$

It is evident that $\tilde{L} = 0$ corresponds to the case $\mathfrak{K} =$ constant and implies $\mu_p I_n - \mu_n I_p = 0$.

The foregoing now provides a formal solution with $\mathfrak{V}_\infty = 0$ for every assignment of values to $\mathcal{P}_\infty$, $I_p$, and $I_n$. There remains the question of the requirements imposed by the condition

$$n, p \geq 0$$

which is equivalent to

$$\mathcal{P} \geq |N|. \tag{171}$$

This implies first of all that $\mathcal{P}_\infty$ must be chosen $\geq |N|$.

It is instructive to look first at the case $\tilde{L} = 0$. Equation (165) shows immediately that (171) requires the choice of the positive sign for the radical for $N > 0$ and the negative for $N < 0$ to avoid $\mathcal{P}_\infty < |N|$. We further find by substitution of (166) and (169) into (165) that (171) requires

$$r \geq \left[\frac{4}{\Omega kT\mu_p(\mathcal{P}_\infty^2 - |N|^2)}\right] NI_p. \tag{172}$$

The bracketed factor is positive. Since we are interested only in non-negative values of $r$, (172) imposes no restriction if $I_p$ is zero or not of the same sign as $N$. However, for $N$ and $I_p$ of the same sign, (172) establishes an inner radius inside which the solution does not satisfy (171). This may be regarded as establishing the minimum radius for an inner spherical electrode for

prescribed $I_p$ and $\mathcal{P}_\infty$, or alternatively as limiting the possible choices of $I_p$ and $\mathcal{P}_\infty$ for prescribed inner electrode radius. Had we chosen the constants $\tilde{Q}$, $\tilde{R}$ and $\tilde{S}$ so as to obtain prescribed values of $\mathcal{P}$ and $\mathcal{U}$ at a preselected electrode radius $r_0$, restrictions analogous to (172) on the *maximum* radius would appear.

For the case $\tilde{A} = 0$ the restriction analogous to (172) is

$$r \geq -\left[\frac{2}{\Omega\mu_p kT(\mathcal{P}_\infty - |N|)}\right] I_p . \tag{173}$$

Since the bracketed factor is positive, (173) provides no restriction for $I_p \geq 0$, but for $I_p < 0$ establishes a minimum radius of the kind just discussed.

For $\tilde{L}, \tilde{A} \neq 0$, the analog of (172) and (173) is

$$r \geq \frac{\tilde{L}/\tilde{A}}{\Lambda^{-1}\left(\frac{kT\mathcal{P}_\infty}{Ne\tilde{A}}\right) - \Lambda^{-1}\left(\frac{kT|N|}{Ne\tilde{A}}\right)} \tag{174}$$

where $\tilde{A}$ and $\tilde{L}$ are given by (153) and (154) and $\Lambda^{-1}$ denotes the inverse function of $\Lambda$—i.e.,

$$\Lambda^{-1}[\Lambda(x)] \equiv x$$

or

$$\Lambda^{-1}(\Lambda) = \Lambda + \ln|\Lambda - 1| .$$

Equation (174) is a minimum radius restriction of the same kind as those obtaining for $\tilde{A} = 0$, and $\tilde{L} = 0$, but the relationship between the minimum radius $r_0$ and $\mathcal{P}_\infty$, $I_p$ and $I_n$ is considerably more complicated than in the more degenerate cases.

It will be noted that the relation

$$\frac{kT}{eN}\frac{\mathcal{P}_\infty}{\tilde{A}} = \Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right)$$

(with $\tilde{A}$, $\tilde{B}$, $\tilde{M}$ given in terms of $\mathcal{P}_\infty$, $I_p$, $I_n$ by (152), (153), and (154)) determines which function ($\Lambda_1$, $\Lambda_2$, or $\Lambda_3$) is to be used for $\Lambda$ in any given case, because any assigned value ($\neq 0$) is taken on by one and only one of ($\Lambda_1$, $\Lambda_2$, $\Lambda_3$).

If surface recombination is negligible as well as interior recombination, this spherically symmetric solution is of use in the study of "point" contacts on a plane surface of a semiconductor. [Fig. 3 and Ref. 2.]

The results of this section can easily be duplicated for any other choice of the harmonic function $\mathcal{H}$ to obtain a great variety of specimen solutions.

Solutions based on $\mathfrak{C}$'s having a single source singularity (such as the example above) will contain four mathematical parameters, and hence will permit arbitrary selection (subject to (6)) of the physical parameters, $I_p$, $I_n$, $\mathcal{P}_\infty$, and $\mathcal{U}_\infty$. However, solutions based on $\mathfrak{C}$'s having more than one source singularity will provide only a subset of the possible assignments of the physical parameters. For example, the harmonic function associated with the electrostatic field produced by two separate point charges each equidistant from two parallel infinite plane conductors provides solutions of

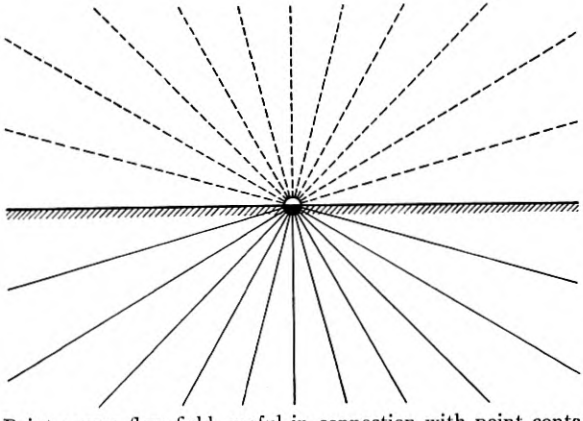


Fig. 3—Point source flow field, useful in connection with point contact theory

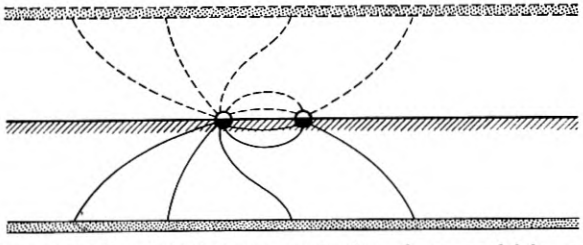


Fig. 4—Two-source flow field between conducting planes, useful in connection with Type A transistor theory.

interest in connection with the type A transistor configuration (Fig. 4). However, the family of solutions obtained contains only a five-parameter subset of the six-parameter family obtainable by arbitrary assignment of $I_{p1}$, $I_{p2}$, $I_{n1}$, $I_{n2}$, $\mathcal{P}_\infty$, and $\mathcal{U}_\infty$.

## S. Sample Application of the Results of Section M: Spherical Symmetry, $N = 0$

We now round out the considerations of Section R by exhibiting the related solutions for $N = 0$ (i.e., the intrinsic semiconductor).

In accordance with the results of Section M, we choose for $\mathcal{P}$ the most general harmonic function with spherical symmetry:

$$\mathcal{P} = \hat{A}\frac{1}{r} + \hat{B} \qquad (\hat{A}, \hat{B}\text{: arbitrary constants}). \quad (175)$$

From (119) then, *for* $\hat{A} \neq 0$

$$\mathcal{U} = \tilde{H} \ln\left(\hat{A}\frac{1}{r} + \hat{B}\right) + \tilde{G} \quad (176)$$

and from (175), (176), (143), and (144)

$$I_p = \frac{1}{2}\Omega\mu_p e\hat{A}\left(\tilde{H} + \frac{kT}{e}\right) \quad (177)$$

$$I_n = \frac{1}{2}\Omega\mu_n e\hat{A}\left(\tilde{H} - \frac{kT}{e}\right) \quad (178)$$

$$I = \frac{1}{2}\Omega e\hat{A}\left[(\mu_n + \mu_p)\tilde{H} - (\mu_n - \mu_p)\frac{kT}{e}\right]. \quad (179)$$

From (177) and (178) we obtain

$$\hat{A} = \frac{\mu_n I_p - \mu_p I_n}{\Omega\mu_p\mu_n kT} \quad (180)$$

and

$$\tilde{H} = \frac{kT}{e}\frac{\mu_n I_p + \mu_p I_n}{\mu_n I_p - \mu_p I_n}, \quad (181)$$

and from (175) and (176) for $\mathcal{U}_\infty = 0$:

$$\hat{B} = \mathcal{P}_\infty \quad (182)$$

and

$$\tilde{G} = -\tilde{H}\ln\hat{B} = -\frac{kT}{e}\frac{\mu_n I_p + \mu_p I_n}{\mu_n I_p - \mu_p I_n}\ln\mathcal{P}_\infty. \quad (183)$$

The condition $\mathcal{P}_\infty \geq |N| = 0$ introduces the restriction (for $\hat{A} \neq 0$):

$$r \geq \left[\frac{1}{\Omega\mu_p\mu_n kT\mathcal{P}_\infty}\right](\mu_n I_p - \mu_p I_n). \quad (184)$$

Evidently this implies no real restriction for $\mu_n I_p - \mu_p I_n < 0$ (i.e., $\hat{A} < 0$), but introduces a minimum radius—of the same kind we have already discussed—when $\mu_n I_p - \mu_p I_n > 0$ (i.e., $\hat{A} > 0$).

*For $\hat{A} = 0$*, $\mathcal{P}$ is constant and, by Section I, $\mathcal{V}$ is harmonic. So we set

$$\mathcal{P} = \mathcal{P}_\infty > 0 \tag{185}$$

and

$$\mathcal{V} = \hat{C}\frac{1}{r} + \hat{D} \tag{186}$$

and obtain from (143) and (144)

$$I_p = \tfrac{1}{2}\Omega\mu_p e\hat{C}\mathcal{P}_\infty \tag{187}$$

and

$$I_n = \tfrac{1}{2}\Omega\mu_n e\hat{C}\mathcal{P}_\infty . \tag{188}$$

From (187) and (188):

$$\hat{C} = \frac{2I_p}{\Omega\mu_p e\mathcal{P}_\infty} = \frac{2I_n}{\Omega\mu_n e\mathcal{P}_\infty} = \frac{\mu_n I_p + \mu_p I_n}{\Omega\mu_n\mu_p e\mathcal{P}_\infty} \tag{189}$$

and from (186) for $\mathcal{V}_\infty = 0$,

$$\hat{D} = 0. \tag{190}$$

Evidently $\hat{A} = 0$ is associated with the condition

$$\mu_n I_p - \mu_p I_n = 0.$$

## T. Summary List of Symbols

*Coordinate Systems:*

$(x, y, z)$: ordinary rectangular cartesian coordinates.
$(\rho, \theta, z)$: ordinary circular cylindrical coordinates.
$(r, \theta, \phi)$: ordinary spherical polar coordinates.
$\mathbf{r}_1$ : unit radial vector in $(r, \theta, \phi)$.
$t$ : time variable.

*Physical Variables:*

$n$ : concentration of negative carriers (electrons).
$p$ : concentration of positive carriers (holes).
$\mathcal{P}$: total carrier concentration $\equiv n + p$.
$\mathcal{U}$: $\equiv \dfrac{kT}{cN}\mathcal{P}$ $\quad (N \neq 0)$.
$\mathcal{R}$: recombination rate function.
$\mathcal{V}$: electrostatic potential.

$\mathfrak{IC}: \equiv \mathfrak{V} - \mathfrak{U} = \mathfrak{V} - \frac{kT}{eN}\mathfrak{P} \qquad (N \neq 0).$

$\overset{\circ}{||}$ : total current density vector.

$\overset{\circ}{||}_n$: electron current density vector.

$\overset{\circ}{||}_p$: hole current density vector.

subscript "0": designates thermal equilibrium values.

subscript "$\infty$": designates values "at infinity".

*Physical Constants:*

$T$: absolute temperature.

$e$ : magnitude of electronic charge.

$k$ : Boltzmann's constant.

$\mu_n$: electron mobility constant.

$\mu_p$: hole mobility constant.

$\alpha : \equiv 1/\mu_p + 1/\mu_n .$

$\beta : \equiv 1/\mu_p - 1/\mu_n .$ (Assumed $\neq 0$)

$\gamma : \equiv \frac{\beta kT}{\alpha e}$

$N: \equiv n_0 - p_0 .$

*Other Constants and Functions:*

$A, B, \cdots, Z$ ((except $I$, $N$, and $T$)),

$\tilde{A}, \tilde{B}, \cdots, \tilde{Z}$,

$\hat{A}, \hat{B}, \cdots, \hat{Z}$: arbitrary constants

$a, b, \cdots, z$ ((except $e$, $h$, $k$, $n$, $p$, $r$, $t$, $x$, $y$, $z$)),

$\tilde{a}, \tilde{b}, \cdots, \tilde{z}$: arbitrary functions of variables designated (e.g., $j(l)$).

$h, h_1, h_2$ : harmonic functions of variables designated at place of usage.

$\Lambda$: $\Lambda(x)$ is defined by the relation $\Lambda(x) + \ln|\Lambda(x) - 1| \equiv x$. (See Figs. 1 and 2.)

$\mathfrak{S}$: $\mathfrak{S}(\mathfrak{U})$ is defined by $\mathfrak{S}[\mathfrak{R}(\mathfrak{U})] \equiv \mathfrak{U}$.

ACKNOWLEDGMENT

The author is indebted to J. Bardeen and W. van Roosbroeck for a critical reading of the manuscript and a number of valuable comments.

REFERENCES

1. W. van Roosbroeck, "Theory of the Flow of Electrons and Holes in Germanium and Other Semiconductors," Bell System Technical Journal, *29*, 4, 560–607 (October 1950).
2. J. Bardeen, "Theory of Relation Between Hole Concentration and Characteristics of Germanium Point Contacts," Bell System Technical Journal, *29*, 4, 469–495 (October 1950).
3. W. Shockley, *Electrons and Holes in Semiconductors*, New York, 1950.

# Instantaneous Compandors on Narrow Band Speech Channels

By J. C. LOZIER

(*Manuscript Received Aug. 15, 1951*)

If speech is passed through an instantaneous compressor, the original speech frequency spectrum is substantially widened. It is known that instantaneously compressed speech can be transmitted over a medium with a passband no wider than that occupied by the uncompressed speech, and the original signals recovered without distortion. The conditions required for such distortionless transmission are examined. The analysis indicates that more severe requirements must be imposed on the attenuation and phase characteristics of the system when this reduced bandwidth mode of operation is used. The practical value of this exchange of transmission requirements is a matter for experimental determination.

## Introduction

When a signal such as speech is instantaneously compressed in amplitude, harmonics and cross modulation products are generated which extend the frequency spectrum of the original signal by many octaves. It is proposed to demonstrate that the additional products thus generated are necessary for the distortionless recovery of the original signal. Then the conditions will be examined under which this broadband signal can be transmitted without distortion through a bandwidth no wider than that occupied by the spectrum of the uncompressed speech. Finally, some of the practical aspects of using intantaneous compandors on narrow band speech channels will be considered, with emphasis on the nature of the transmission requirements placed on the medium. The advantages to be obtained from the use of instantaneous compandors have already been presented by Mallinckrodt.[1]

## Bandwidth vs Distortion

If a single frequency tone is compressed by a 2 to 1 compressor,[2] and then the fundamental alone is expanded, it can be shown that the resultant 3rd harmonic distortion is only 13 db below the fundamental. Expansion of both the fundamental and the 3rd harmonic output of the compressor will reduce this 3rd harmonic distortion to 29 db below the fundamental. Expansion of the fundamental plus the 3rd and 5th harmonics will reduce the 3rd harmonic distortion in the recovered signal to 45 db below the funda-

[1] C. O. Mallinckrodt "Instantaneous Compandors," *B.S.T.J.*, Vol. XXX, No. 3, July 1951.

[2] In a 2 to 1 compressor, the output amplitude is the square root of the input amplitude. The name comes from the fact that, in such a compressor, the output amplitude will increase 1 db for each 2 db increase in input amplitude.

mental. These results are indicative of the significance of such components in the compressor output spectrum to distortion in the recovered signal.

## Frequency Analysis of Transmission under Reduced Bandwidth Conditions

It might be concluded from the results quoted above that a wideband channel is required for the distortionless transmission of instantaneously compressed speech. However, if the compressed speech is properly sampled

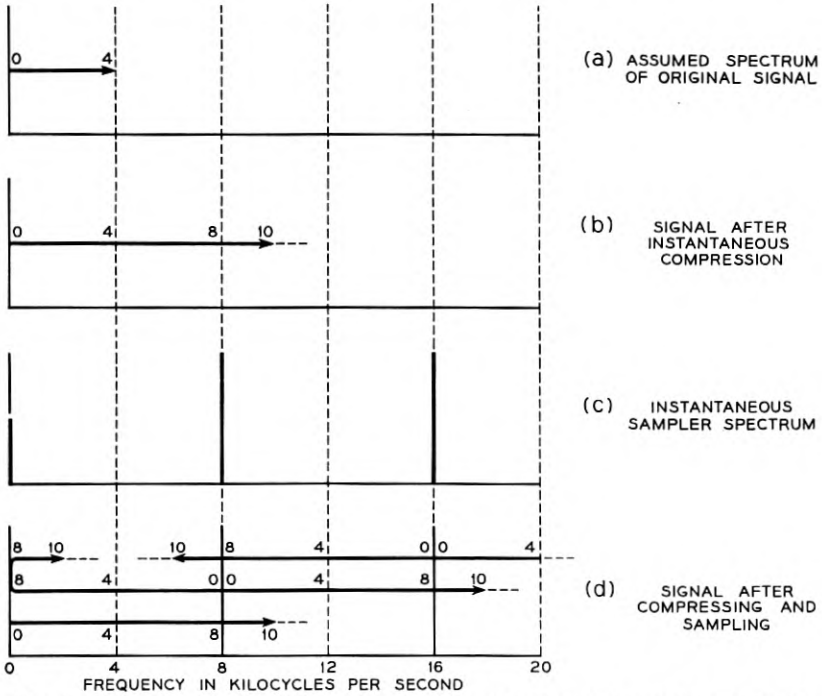



Fig. 1—Frequency analysis of instantaneous compressing and sampling of origina signal.

before transmission, and the received signals are again sampled at the receiver, the bandwidth of the intervening medium can be restricted to that of the original speech, and still the transmission can be distortionless. Hence, the sampling must transform the broadband spectrum of the compressed speech in such a way that it can be successfully transmitted over a relatively narrow band.

A steady state frequency analysis will serve to illustrate this phenomenon. Figure 1(a) shows the 4 kc frequency spectrum assumed for the original

signal, and Fig. 1(b) shows a 10 kc portion of this signal after instantaneous compression. Now the minimum sampling rate required to handle a 4 kc signal band is 8 kc. It is also the sampling rate that will allow the maximum band reduction in this case, as further analysis will show. The frequency spectrum of a sampling function with an 8 kc repetition rate has a d-c. component, an 8 kc fundamental, and all the harmonics of this repetition rate as shown in Fig. 1(c). These harmonics are all of equal amplitude and all are phased so as to add up every 125 microseconds to form the characteristic sampling waveform. Figure 1(d) shows the frequency spectrum formed by sampling the 10 kc portion of the compressed speech signal. It

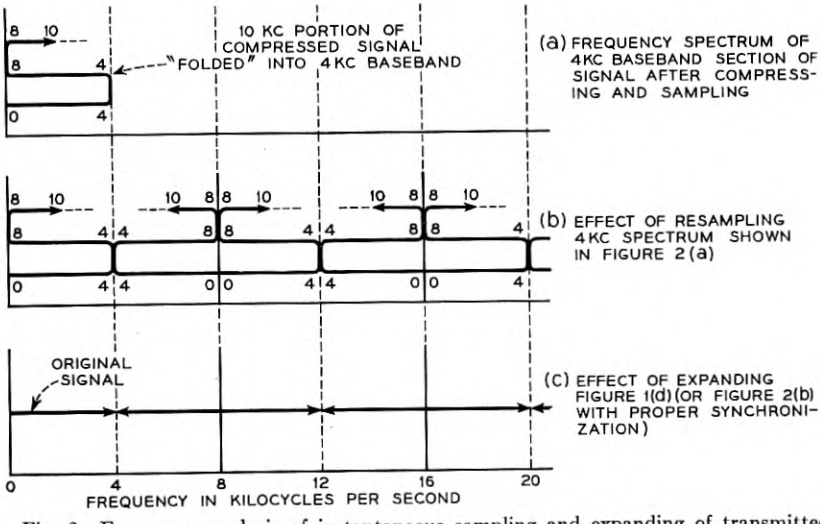



Fig. 2—Frequency analysis of instantaneous sampling and expanding of transmitted signal.

represents the product of the spectra of Fig. 1(b) and 1(c). As such, it is composed of the various component frequencies in the sampling spectrum as carrier frequencies, with the 10 kc portion of the compressed speech signals as amplitude modulated sidebands about these carriers.

Figure 2(a) shows the resulting spectrum that falls in the 4 kc baseband of Fig. 1(d). It represents that part of the compressed and sampled spectrum that would be received over an ideal 4 kc baseband channel. This spectrum is worth examining because it illustrates how the addition of instantaneous sampling makes it possible to transmit all the components in the compressed speech over a 4 kc channel. The effect may be described as a linear "folding" of the broadband spectrum back and forth over the 4 kc band. However,

although any broadband signal can be similarly folded into a 4 kc band by an instantaneous sampler with an 8 kc repetition rate, the process is not fully reversible. For example, there is no means of telling whether a 3 kc component in the folded signal comes from a 3 kc, a 5 kc, an 11 kc, or a 13 kc, etc. component in the original signal. Hence it is only a very special class of signals that can be recovered after their frequency spectra have been condensed in this fashion.

To recover the original speech in this case, the spectrum shown in Fig. 2(a) can be sampled at an 8 kc rate to produce the spectrum shown in Fig. 2(b). Now an examination of the spectra involved will show that when the second sampling is properly synchronized with the transmitting sampler, the two spectra shown in 1(d) and 2(b) will be identical. The spectrum in Fig. 1(d) represents the 8000 samples per second of the compressed speech generated at the transmitter. Thus, when the spectra of Figs. 1(d) and 2(b) are identical, samples will be recovered at the receiver which are identical to those that were generated at the transmitter. These can be converted to samples of the uncompressed speech by complementary instantaneous expansion. The spectrum of these samples is shown in Fig. 2(c). All that is necessary at this point to recover the original speech without distortion is to pass these samples through a 4 kc. low-pass filter.

## Requirements for Distortionless Transmission on Reduced Bandwidth Basis

Thus the criterion for distortionless transmission of compressed and sampled speech under these conditions is that the samples recovered at the receiver be the same as the samples of compressed speech that were generated at the transmitter. This means sending 8000 pulses per second over a 4 kc band without intersymbol interference. Nyquist[3] has shown that this is the maximum rate at which independent pulses can be transmitted over a 4 kc band and still be recovered at the receiver. At this maximum rate, the bandwidth employed does not give the transient response of one pulse time to die out before the next pulse is received. Therefore the transient response in this case must be such that, when one pulse is at its peak, the transient responses of all other pulses will be going through zero. The infinitely sharp cut-off at 4 kc, which is required to separate out the spectrum shown in Fig. 2(a) from that in Fig. 1(d), will have the required zeros in its pulse response, provided the attenuation is constant and the phase is linear with frequency.

This is the familiar $\frac{\text{Sin } x}{x}$ shape of transient response. Nyquist has shown

[3] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Transactions*, Vol. 47, Pages 617 to 644, April 1928.

also that this is just one of a whole family of transmission characteristics with a specified symmetry about the cut-off frequency, all of which have the required transient zeros. However, there is no reason to suppose that any of them would prove less sensitive to variation of the transmission characteristics from the ideal than the one described above.

It is apparent that synchronization of the transmitting and receiving samplers is required to insure that the receiving sampling is done at the exact instant that all transient responses but the desired one are zero.

## Effect of Variations from Ideal Transmission Characteristics on Distortion

In practice of course a certain amount of distortion is tolerable. To get some measure of the practicability of such reduced bandwidth transmission of compressed speech, the first step is to determine how much intersymbol interference can be tolerated in this type of signal, and then to translate this tolerance into allowable variations in the frequency characteristics of the transmission medium from the assumed ideal. However, it is hard to estimate what the allowable intersymbol interference might be in this case. In a single channel system, intersymbol interference produces a form of distortion, and the sensitivity of such signals to distortion is primarily a matter for subjective determination.

For computational purposes it will be assumed, however, that this intersymbol interference should be 20 db down in the output. It is apparent that a 5% variation in the amplitude of a sample before expansion will produce a 10% variation in the expanded sample. On this basis 5% intersymbol interference in the medium between transmitter and receiver is the maximum allowable. Using Wheeler's theory of paired echoes[4], it can readily be shown that a sinusoidal variation in the phase vs. frequency characteristics of the medium, with an amplitude of $\frac{1}{10}$ of a radian (5.7 degrees), will cause a pair of echoes each of which will have a peak equal to 5% of the original sample. Similarly a sinusoidal deviation in the attenuation vs. frequency characteristic of 0.9 db from the ideal will also cause a pair of echoes with an amplitude of 5%.

In estimating the average effect of such echoes, it cannot be expected that the intersymbol interference from a given echo will be appreciably less than its peak amplitude would indicate. The principal reason is that, in order to realize the savings in bandwidth, the pulses are 125 microseconds apart, which is as close together as the 4 kc band will permit. Reference to the $\frac{\text{Sin } x}{x}$

[4] H A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *I.R.E.*, June 1939.

form of transient response indicates that the width of pulses (and hence of echoes) received over a 4 kc band, is such that they will be within 65% of their peak amplitude for a full 125 microseconds. Thus such echoes will cause at least 65% of their peak interference to at least one subsequent pulse. This illustrates why it is so difficult to control intersymbol interference in pulse systems operating under minimum bandwidth conditions.

Assuming from this argument that the interference from echoes should be taken at their peak values, the tolerable phase deviations from linearity must be measured in tenths of a radian in this case. On ordinary speech channels the tolerable phase deviations from linearity are measured in radians, which represents a difference of one or two orders of magnitude.

Another estimate of the allowable intersymbol interference may be obtained by comparing it to quantizing noise on a PCM system. A 5-digit PCM system has 32 quantizing levels, and the average uncertainty in the recovered pulse amplitude is one half of a quantum step, or approximately 1.6%. The 10% intersymbol interference requirement chosen above represents approximately 6 times as much deviation in recovered pulse amplitude. Again only subjective measurements can tell whether intersymbol interference in this case is six times more tolerable than quantizing noise. However, a 5-digit PCM system is not a high quality circuit by Bell System standards.

The distortion effects due to lack of synchronization of the transmitting and receiving samplers have been ignored in the discussion so far, on the assumption that it would not prove too difficult in practice to make it a relatively negligible source of intersymbol interference. However, it may not prove to be a negligible factor from the economic standpoint, when an attempt is made to prove in a system of this type.

## Multichannel Aspects

In the case of multichannel time division systems, the addition of instantaneous campandors seldom requires an increase in the transmission requirements of the medium. In multichannel PAM and PPM systems, for example, intersymbol interference causes intelligible crosstalk between channels, and the requirements on such crosstalk usually calls for the intersymbol interference to be some 60 db down in the recovered speech. In such cases the addition of an instantaneous compandor can serve to reduce this requirement on the line to some 40 db, through the so-called "Compandor Advantage"[5]. It is fair to point out, however, that such systems are seldom, if ever, operated as minimum band pulse systems.

[5] C. O. Mallinckrodt, "Instantaneous Compandors," *B.S.T.J.*, Vol. XXX, No. 3, July 1951.

## Conclusions

It has been shown that distortionless transmission of instantaneously compressed speech over a frequency band no wider than that required for the uncompressed speech does involve the transmission of a broad-band signal over a relatively narrow-band channel. This is made possible by the use of an instantaneous sampler which serves to "fold" the spectrum of the compressed speech at the transmitting end so that the entire spectrum is contained within the desired bandwidth. The criterion for distortionless transmission of these "folded" signals is shown to be one of recovering at the receiving end the precise samples of compressed speech that were generated at the transmitter. To accomplish this distortionless recovery of the transmitted pulses it is necessary, first, that the transmission medium cause no intersymbol interference, and, second, that the signals at the receiver must be sampled in synchronism with the sampling at the transmitter.

It was also shown that the full reduction in bandwidth can be realized only by pulse operation under minimum bandwidth conditions. It was estimated that the accuracy of control of the steady state phase and attenuation vs. frequency characteristics that would be required to maintain the intersymbol interference below an acceptable level would be hard to meet in practice, primarily because of having to operate under such minimum bandwidth conditions.

# The Evolution of Inductive Loading for Bell System Telephone Facilities

By THOMAS SHAW

(*Concluded from July 1951 issue*)

## PART VI: CONTINUOUS LOADING

### *General*

CONTINUOUS loading, i.e., the addition of uniformly distributed inductance, was studied theoretically in the Bell System several years before theoretical work started on coil loading. This early work of John Stone Stone, then a member of the headquarters technical staff of the American Bell Telephone Company, resulted in the issue to him on March 2, 1897 of a *U.S. Patent* (575,275) describing a "bi-metallic" wire cable.

Later on, when the commercial development was authorized, cost considerations made it desirable to start with laboratory experiments on an "electrically equivalent" artificial line using small lumped inductances. In planning these experiments, it soon became apparent that only a small amount of distributed inductance could be obtained with the best magnetic material then available, namely, iron. Recognition of the important advantages inherent in the use of large amounts of inductance, and of the absence of limitations regarding the magnitude of inductance that could be provided in coil form, then shifted the development emphasis to the as yet unsolved problem of spacing inductance coils in relation to wavelength. This theoretical problem was quickly solved by G. A. Campbell, who was then in charge of the project, and accordingly the laboratory artificial line was designed to demonstrate the practicability of coil-loading (early in 1899). The Bell System development work on continuous loading was then suspended for some time.

During the next two decades, coil loading was found to be economically suited to all Bell System needs for inductive loading, even on short intermediate submarine cables required at shallow water crossings of rivers and bays. Shortly after the First World War, however, it became necessary to undertake the development of continuously loaded cable to meet an urgent demand for telephone communication with Cuba. Exploratory theoretical studies and laboratory investigations had been started shortly before the war, but were discontinued during the war. The exploratory work included consideration of the possible use of a new nickel-iron magnetic alloy which

was then under development by the Research Department of the Western Electric Company, and which later on became widely known as permalloy.[23]

*Key West-Havana Submarine Telephone Cable System*

This project required three different submarine cables ranging in length from about 100 to 105 nautical miles, each being a great deal longer than any previously designed for telephone transmission, and a large fraction of the route was in deep water, reaching a maximum depth somewhat over 6000 ft. The difficulties to be expected in protecting loading coils from injury under the great hydrostatic pressure involved, and the complications that would be encountered during installation and in subsequent maintenance work, prevented coil loading from receiving consideration. Moreover, the great water pressure also eliminated consideration of paper insulated cable.

Since the cables were intended for use in telephone circuits connecting remote points in the United States with Havana and remote points in Cuba, the over-all system design requirements were very formidable. In addition to a two-way telephone circuit in each cable, provision also was made for three carrier telegraph circuits above the voice range, and for direct current grounded telegraphy below the voice. These complex requirements brought in difficult problems regarding telegraph flutter interference and other types of non-linear distortion.

The fundamental design studies resulted in a decision to install single core, continuously loaded, cables using gutta-percha insulation, and having a concentric system of copper tapes wrapped around the insulated conductor, for use as a return conductor. (These cables were the first to be installed with this feature.) Iron-wire type continuous loading was chosen largely because the desired project in-service dates did not allow sufficient time for the additional research and development work, and the additional manufacturing preparations, that would have been necessary in order to use permalloy tape loading. The manufacturing situation presented serious problems, because it was necessary to plan for manufacture abroad, since no American company had facilities for making deep-sea submarine cable. Moreover, iron-wire type continuous loading (as proposed by C. E. Krarup* of Denmark) was old in the European telephone art, having been used in several short submarine cables, and some underground cables.

In the Cuban Straits cables under discussion, the central copper conductor had a diameter of about 0.140 inch. About this was closely wrapped a single layer of 0.008 inch soft iron wire and three layers of gutta-percha type insulation having a total thickness of about 0.135 inch. A thin copper tape directly on this core furnished protection against damage by the teredo,

* *E.T.Z.*, April 17, 1902.

and was part of the system of copper tapes previously mentioned which served as a return conductor.

The effective permeability of the iron-wire loading material was about 115. The distributed inductance of about 4.35 millihenrys per nautical mile resulted in a low nominal impedance of about 115 ohms. The energy losses in the loading material were the principal factors in limiting the top of the working frequency band to about 4000 cycles. At 1000 cycles per second, the bare line equivalent was of the order of about 22 db (for the mean value of the longest and shortest cable). At 4000 cycles it was about 2.2 times as great.

Space limitations prevent a more complete description and discussion here. Comprehensive information regarding all features of the project is given in a 1922 *A. I. E. E.* Paper[42] prepared by Messrs. W. H. Martin, G. A. Anderegg and B. W. Kendall. Engineers of the A. T. &. T. Co. and W. E. Co. were responsible for the electrical design of the cables, method of operation, and arrangement of the repeaters and other terminal apparatus. The cables were manufactured late in 1920 and installed early in 1921 by The Telegraph Construction and Maintenance Co. Ltd. of London, for the Cuban-American Telephone and Telegraph Company. The latter organization is jointly owned by the A. T. &. T. Co. and the Cuban Telephone Co. (a subsidiary of the International T. &. T. Co.)

*1930 Non-Loaded Cable:* Since the 1921 cables were not suitable for carrier telephone operation (largely because of excessive losses and non-linear distortion at high frequencies), it became necessary during 1930 to install a fourth cable between Key West and Havana in order to meet the demand for additional facilities. Advantage was taken of advances in the communication art, notably an improved cable insulation (paragutta), improved repeaters and carrier telephone systems, to design a non-loaded cable system which would be suitable for carrier operation. The initial carrier set-up provided three carrier telephone circuits, using a type C4 system which had originally been developed for open-wire lines. Early in 1942, a seven-channel system was substituted. Comprehensive information regarding the 1930 cable and its use of the 3-channel carrier telephone system is given in a 1932 *A. I. E. E.* paper by Messrs. Affel, Gorton and Chesnut.[43]

*High Speed Transoceanic Loaded Telegraph Cables*

During the First World War when the need for increasing the message-carrying capacity of existing non-loaded transoceanic telegraph cables became urgent, the Bell System engineers who worked on this problem finally came to the conclusion that to obtain a great advance in the existing art it would be necessary to have much better cables.

In July 1919 the continuing interest in this problem crystallized in a Western Electric proposal to use permalloy continuous loading in new transoceanic telegraph cables. Since this remarkable new magnetic alloy[23] had been invented and developed by Western Electric engineers, they were already familiar with its extraordinary high permeability characteristics, and had confidence in their ability to use it in providing a high impedance loading which would make practicable a great increase in message-carrying capacity. Loading with iron-wire would not have any advantage in telegraph speed, because of its low permeability. Intensive research work quickly started on the permalloy loaded cable design and installation problems, and on the related terminal apparatus and operating problems. The success attained in these efforts resulted in disclosures to the Western Union Telegraph Company regarding the great increase in telegraph signaling speed that could be obtained with the proposed new permalloy loading. In due course the Telegraph Company made arrangements with the Telegraph Construction and Maintenance Company Ltd. of London for the manufacture and installation of a 120-mile trial length, using loading material supplied by the Western Electric Company and applied and treated under the direction of Western Electric engineers. In October 1923 this experimental length was laid in deep water near the south shore of Bermuda. The trial installation tests were so satisfactory that the Western Union company arranged for the manufacture and installation of a 2300-mile cable to connect New York with Horta in the Azores. As with the trial length, the loading material was supplied by the Western Electric Company, and it was applied and treated under Western Electric supervision.

The new cable was laid during September 1924. After refined adjustments in the terminal apparatus, a speed of over 1900 letters per minute was obtained. This speed is about four times the carrying capacity of an ordinary non-loaded cable of the same length. At this point a brief statement of general theory is indicated: The effect of the inductance is to oppose the setting up of a current and to maintain it once it has been established, thus preserving a definite wave front as the signal impulse travels over the cable. The individuality of the signal impulses is retained, and thus the much higher speed becomes possible.

The permalloy loading material was applied in tape form in a close helix around a stranded copper conductor. The tape was 0.006 inch thick and 0.125 inch wide. The alloy was composed of about 79% of nickel and 21% of iron and a small amount of manganese, suitably heat treated. It provided an inductance of about 54 millihenrys per mile, slightly over 12 times that obtained by the use of iron wire in the Cuba cables previously described. The permeability of the loading was about 2300, or about 20 times that of

the iron wire used on the Cuba cables. An important feature of the cable not previously mentioned was a layer of viscous insulating material (under the regular gutta-percha insulation) which protected the strain-sensitive permalloy from the stresses caused by hydrostatic pressure in the great depths of the ocean.

Demand for other high-speed loaded submarine cables quickly followed the successful demonstration of the New York-Horta cable and several were installed during 1926, reaching a total of about 15,000 miles of high-speed cables. The new installations included the Horta-Emden cable manufactured and installed by the Norddeutsche-Seekabelwerke A.G. for the Deutsch Atlantische Telegraphengesellschaft, and the New York-Bay Roberts-Penzance cable manufactured and installed by the T. C. & M. Company for the Western Union Telegraph Company. These particular cables used an improved form of permalloy supplied by the Western Electric Company containing about 80% nickel, 17.5% iron, 2% chromium, and 0.5% manganese. This alloy had an initial permeability of about 3700 and provided a higher impedance loading than that used on the first high-speed cable. In consequence, the newer cables were capable of speeds of about 2500 letters per minute.

Other high-speed continuously loaded cables, installed in 1926 and subsequent years, used permalloy material manufactured under Western Electric Company patent license, in some instances under a special foreign trade name.

Comprehensive information regarding all features of the high-speed cable projects specifically mentioned above is given in two papers by O. E. Buckley, published in 1925[44] and 1928[45], respectively.

In passing, it should be observed that the permalloy loaded cables under discussion were not intended for, and were not suitable for telephone communication. For this purpose, a new family of magnetic alloys, the perminvars, was developed.[46] Their composition centered on 47% nickel, 25% cobalt, 20% iron, 7.5% molybdenum, and 0.5% manganese. When used as a thin loading tape, this alloy has electrical and magnetic properties especially suitable for telephone transmission, including very low hysteresis which is very advantageous in the control of all forms of non-linear distortion.

*A Proposed Transatlantic Telephone Cable*

During the late 1920's, there was worked out a design of a perminvar loaded cable suitable for voice frequency telephony between Newfoundland and Ireland (1800 nautical miles). It was of the single core type with a concentric return conductor. Four layers of very thin perminvar tape

provided the loading, and the loaded conductor was insulated with paragutta. The suitability of the design for use in deep water was verified by temporarily dropping a 20-mile length on the sea floor in a deep water section of the Bay of Biscay.

The general business depression of the early 1930's resulted in a postponement of the cable project because of its great cost. Later on the project was postponed indefinitely because, in the face of improvements in transatlantic radio telephone communicaion, so expensive a cable to carry a single conversation could no longer be justified.

Additional information regarding this cable project is included in Dr. O. E. Buckley's 1942 paper, "The Future of Transoceanic Telephony," constituting the 33rd Kelvin Lecture before the Institution of Electrical Engineers.[47]

*Continuous Loading for Paper Insulated Telephone Cables*

*Tape and Wire Loading:* When permalloy and perminvar first became available, theoretical studies were undertaken to determine the prospects of economic competition with coil loading on ordinary paper insulated telephone cables. Special consideration was given to the use of the magnetic alloys in situations where coil loading is most expensive, namely, in submarine intermediate cables at river crossings, many of which involve high-frequency carrier telephone operation. None of these studies, however, gave sufficient promise to warrant commercial development work.

*Electroplated Permalloy Loading:* During the middle 1920's, the Bell Telephone Laboratories started research work on a radical new concept of continuous loading using electroplated permalloy, which gave some promise of being less expensive than magnetic alloy tape or wire loading. The process involved the electrolytic deposition simultaneously of suitable proportions of nickel and iron on the copper conductor, and the use of special heat treatments to obtain the desired characteristic (magnetic and electrical) properties of permalloy. In due course, methods were devised for separating the concentric magnetic layer from the conductor, and for breaking it up into longitudinally discontinuous pieces, so as to secure the most advantageous properties for telephone transmission service, and to provide mechanical flexibility in handling.

The experimental work was concentrated on small copper conductors, partly because of the more simple process problems, and partly because such combinations appeared to have the best prospects of competing with coil loading from the plant cost standpoint. (N.B.—The amount of permalloy loading material required to provide a specified inductance per unit length, and its cost, is a direct function of the conductor diameter.)

The requirements for and the possibilities of using electroplated loading in the exchange area services were given priority in the theoretical cost studies—largely because of their extensive use of small conductor cables. These studies indicated some attractive possibilities of using light-weight electroplated loading on fine wires (26 and 24-gauge) as substitutes for larger size wires without loading, provided satisfactory solutions could be worked out for the circuit balance and magnetic instability problems. The balance problem arises from the difficulty of securing sufficient uniformity among the loaded conductors used as wire and mate in the individual pairs. This is complicated by the sensitivity of the permalloy continuous loading to magnetization by steady and intermittent superposed signaling currents. On the larger-size exchange cable wires that are not now used extensively without coil loading, the comparative cost estimates were not attractive for the electroplated loading.

The inflexibility of continuous loading is an adverse general factor, since it is not feasible to decrease or increase the weight of the loading after manufacture, in order to accommodate changes in transmission requirements made desirable by changes in performance standards or alterations in circuit layouts. Also, there would be inflexibility in conforming to changing requirements in complement sizes of loaded circuits in areas where it is necessary to have loaded and non-loaded circuits within the same cable sheath.

Theoretically, one of the flexibility limitations of the continuous loading could be reduced by using coil loading in combination with it, in order to extend its transmission range. However, this would reduce the width of the transmission band below that obtainable with the coil loading on a circuit not having continuous loading—the decrease in effective cutoff being a complex function of the ratio of distributed inductance to coil inductance. Combinations of high cutoff, low impedance, coil loading with low inductance continuous loading could be designed to have satisfactory band width properties. For a given grade of transmission performance, however, such combinations appear to be inherently more expensive than coil loading or continuous loading by themselves.

The experimental work on electroplated continuous loading for exchange area cables was carried on somewhat intermittently during the 1930's. At no time did the prospects of securing satisfactory over-all transmission performance, at costs which would encourage competition with coil loading, appear to be sufficient to warrant an all-out sustained attack on the many difficult technical problems involved. Although the development project has not been permanently abandoned, it had to be discontinued in the late 1930's on account of the great pressure of more urgent work.

The use of electroplated loading as a substitute for coil loading on toll cables, or on incidental cables in open-wire lines, did not appear to be attractive when the cost estimates and the complex requirements on circuit balance, stability, non-linear distortion and flexibility were taken into account.

*Summary*

Enough has been told in the preceding pages to support the earlier statements regarding the low importance of continuous loading in the growth of the Bell System, relative to that of coil loading. Obviously, the success attained by the intensive development and in the very extensive use of economical types of coil loading is an important factor in this situation. That these extent-of-use relations are not due to a lack of interest in continuous loading is well demonstrated by the Bell System initiative in developing the permalloy continuous loading that made high-speed telegraphy practicable in long submarine telegraph cables, and by the other development work summarized in this review.

## PART VII: EXTENT OF USE AND ECONOMIC SIGNIFICANCE

### Introduction

Up to now, this account of coil loading has been in terms of individual developments and their significance with respect to the prior art and current developments in related fields, with occasional information regarding their importance and extent of use.

It is now appropriate to supply and analyze some general statistics regarding the total amount of loading which has been used, in a rough appraisal of the importance of coil loading in the growth of the Bell Telephone System. Some important qualifications of the statistics are commented upon in advance of the presentation of actual figures.

The statistics here given and discussed are for the most extensive and most important applications of coil loading, namely, for voice-frequency loading over cable circuits. They are grouped in two principal categories: (1) non-phantom type coils used on non-quadded exchange area cables, and to a relatively very small extent on toll cables, and (2) side circuit and associated phantom coils used on quadded long-distance and interurban toll cables, and to a relatively very small extent on entrance cables in open-wire lines and on long quadded exchange cables.

The figures used are based on production statistics up to the end of 1949. The important significance of the production figures is that they measure at the time of manufacture the current demands for additional loaded facilities required by the growth of the telephone system, and the up-to-

then accumulated total demand. In general, the loading coils were manufactured to meet specific customers' orders; manufacture for merchandise stock in anticipation of future orders was seldom undertaken, except during periods of extraordinarily high, sustained, demand. On this basis practically all of the coils that were manufactured were installed in the telephone plant.

The production statistics of course include a considerable number of coils which were installed shortly after manufacture and which were taken out of service many years later to facilitate the use of improved transmission systems that required different types of coils, or to permit the use of carrier systems on the unloaded toll cable circuits. In general, complete potting complements were not taken out of service in preparation for carrier systems operation; i.e., a large fraction of the disconnected loading coils remain in the cases in which they were originally potted and installed, and the other coils in the same cases are still in service. It is important to remember that the displaced loading coils played an important part in the improvement and growth of telephone service in their own period of commercial use. The unavailability of statistics regarding displaced loading makes it impossible to supply accurate information regarding the total number of loading coils now being used for regular telephone service. It seems probable, however, that about 80% or more of all the toll cable coils that have been manufactured are in service, or installed in circuits which will be used as soon as traffic growth requires them. The corresponding percentage figure for exchange area coils is probably higher. The number of loading coils taken out of service because of incipient defects that were not detected in the factory inspection tests, or which became unserviceable in consequence of service injuries, or which have been junked because of obsolescence, is a very small fraction of the total number of coils that have been manufactured for Bell System use

## General Production Statistics, Voice-Frequency Cable Loading

### *Total Production*

The grand total production figure (up to the end of 1949) for all types of voice-frequency loading coils for Bell System use is of the order of 20.7 million. Approximately 54% of this total (about 11,270,000 coils) are non-phantom type coils, used almost entirely on exchange area non-quadded cables. Nearly 9,500,000 coils are side circuit or phantom loading coils used on quadded toll and toll entrance cables. Approximately three-quarters of the grand total have been manufactured during the last two decades.

The greatly varying rates in the growth of loading coil production are shown, (a) in terms of accumulated total production through 1949 in

Table XIX and (b) in annual totals during the period 1920–1949, plotted in Fig. 35.

*Annual Production Totals*

In general, the average and peak figures of annual production prior to 1920 were very small relative to those in the 1920–1949 period covered by the chart. For example, the maximum annual production of side circuit and phantom toll cable coils prior to 1925 was in the war year 1918,* and the maximum annual production of non-phantom exchange area cable coils

TABLE XIX

ACCUMULATED TOTAL PRODUCTION(1)—VOICE FREQUENCY CABLE LOADING COILS (IN MILLIONS OF COILS)

| At End of Year | Side Circuit(2) and Phantom Coils | Non-Phantom Coils | Total |
|---|---|---|---|
| 1915 | 0.31 | 0.22 | 0.53 |
| 18 | 0.52 | 0.30 | 0.82 |
| 20 | 0.64 | 0.35 | 0.99 |
| 22 | 0.73 | 0.39 | 1.12 |
| 1924 | 0.95 | 0.53 | 1.48 |
| 26 | 1.49 | 0.79 | 2.28 |
| 28 | 2.69 | 1.32 | 4.01 |
| 30 | 5.59 | 2.06 | 7.65 |
| 1934 | 6.44 | 2.60 | 9.04 |
| 38 | 6.65 | 3.21 | 9.86 |
| 40 | 7.04 | 3.81 | 10.85 |
| 42 | 7.82 | 5.15 | 12.97 |
| 1944 | 8.14 | 5.48 | 13.62 |
| 46 | 8.49 | 6.69 | 15.18 |
| 48 | 9.33 | 9.76 | 19.09 |
| 49 | 9.46 | 11.27 | 20.72 |

Notes: (1) All production figures are approximate values.
(2) Commercial production of side and phantom coils did not start until 1910. Up to that time non-phantom coils were used for toll cable loading (and for exchange area cables).

prior to 1923 was in the war year 1917.† Thus, with occasional exceptions, the production data for the years prior to 1920 could not be accurately plotted on the chart without using a confusing scale.

In the beginning, the use of loading was small relative to its subsequent use because the Bell System cable plant was small. For nearly a decade the expanding toll cable plant used fewer coils than the exchange plant. From then on, in the two-decade period 1913–1932, toll cable loading dominated

* 117,000 coil peak in 1918; 187,000 coil total in 1925.
† 33,000 coil peak in 1917; 38,000 coil total in 1923.

in the extent of use, reaching its all-time peak in growth during 1930. The four-year period of most rapid expansion of toll cable loading coincided with: (a) the full scale introduction of four-wire repeatered loaded (H44-25) circuits for long haul long-distance facilities, (b) the introduction of permalloy-core loading coils which resulted in large loading economies, and (c) the planned use of relatively large circuit-groups in order to speed up the long-distance service.

The business depression of the early 1930's terminated the rapid expansion period in all types of loading. Several years later, when business conditions

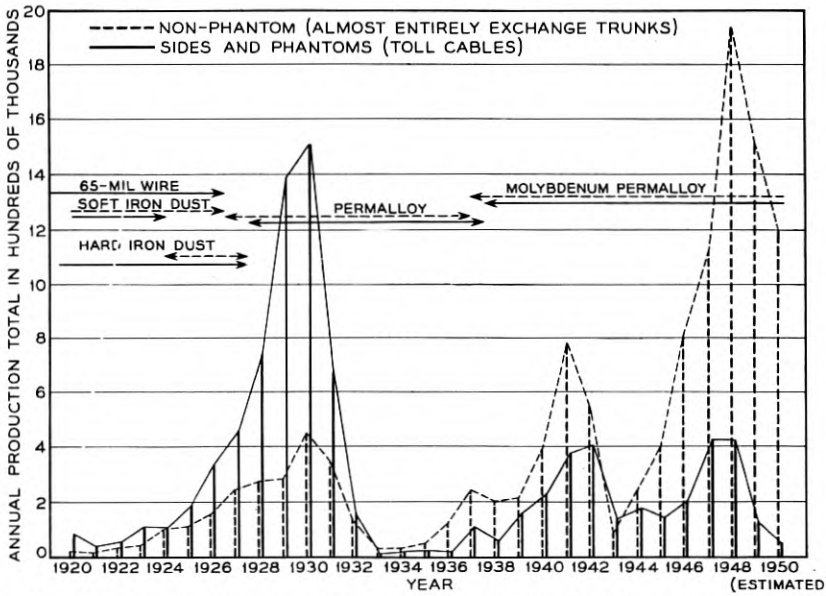



Fig. 35—Annual production totals of voice-frequency cable loading coils for Bell Telephone facilities.

improved sufficiently to require another large expansion in the toll cable facilities, the demand for new long-haul circuits was taken care of generally by the use of Type K carrier systems on non-loaded cable pairs and pairs from which loading was removed; and the use of new toll cable loading was largely restricted to short-haul repeatered and non-repeatered circuits. Thus it happened that, during the 1939–1942 period of rapid plant expansion, the production of exchange area loading coils substantially exceeded that for toll cable loading in the struggle to meet the demands for additional facilities required by the war effort.

The post-war drive to meet the greatly increased demands for long-

distance telephone service, and the provision of a tremendous amount of new exchange plant to take care of more than eleven million new Bell System telephone stations, made it necessary to build up the production rates to higher values than those during the war period. An important factor in the new heavy demands was the desire to restore the speed of service to the pre-war standards.

The post-war demand for exchange area loading has been greatly in excess of that in any previous spurt in demand, reaching its peak value during 1948, and has been very large in relation to the toll cable loading requirements. The post-war rapid build-up of a backbone network of coaxial cables, together with the expanding use of carrier systems in existing and new cables of the conventional types, and the introduction of microwave radio relay systems have held down the demand for new toll cable loading to relatively small quantities for use on relatively short circuits.

*Relative Costs, Toll and Exchange Loading*

Production statistics by themselves do not indicate the relative economic importance of exchange area and toll cable loading. Except in the early years when coils of the same size were used for both types of loading, the toll cable loading coils have been considerably more expensive than the exchange area loading coils. During the periods of maximum production and use portrayed in Fig. 35, the average prices per potted toll cable loading coil have ranged up to about twice or three times as large as those per potted exchange area coil. Consequently, the total plant investment in toll cable loading is substantially greater than the total investment in exchange area loading, notwithstanding the somewhat greater total use of exchange area loading, as indicated by the production statistics. This is consistent with the fact that more expensive types of cable are used for the toll circuits and the service requirements are more difficult.

*Analysis in Relation to Core Materials*

There now follows a rough breakdown of total production in terms of core materials, in recognition of the importance of the cores in determining the coil performance characteristics and costs:

In general, the production percentage figures in Table XX do not discriminate between types of facilities (toll or exchange area). If separate percentage-of-total figures should be derived for toll facilities and for exchange area facilities, those for toll facilities would substantially exceed the tabulated figures for total iron-wire, iron-dust, and permalloy-powder core loading coils, especially in the case of the latter, and the percentage-of-total figure for exchange area molybdenum-permalloy core coils would greatly exceed that for toll cable loading.

In considering the two different permeability types of iron-wire and of iron-dust core-materials, it is important to note that in each case the lower permeability material had a much more extensive total use than the higher permeability material, and that it was used in the more important facilities.

It is of special interest from the plant-cost standpoint that nearly two-thirds of the compressed molybdenum-permalloy powder core coils (up to the end of 1949) are the reduced cost designs using Formex-insulated conductors in their windings, this being an important factor in coil-size reduction. The other molybdenum-permalloy core coils are larger-size coils using a combination of textile and old type of enamel conductor-insulation.

It is highly significant with respect to the economics of the Bell System plant growth that over one-third of all voice-frequency loading coils manufactured up to the end of 1949 are of the lowest-cost types ever standardized

TABLE XX

ESTIMATED DISTRIBUTION OF ACCUMULATED TOTAL LOADING COIL PRODUCTION UP TO END OF 1949 IN TERMS OF CORE MATERIALS

| Core Material | Percentage of Total Loading Coil Production | Approx. Period (1) of Commercial Manufacture |
|---|---|---|
| Fine Iron-Wire | 1.5 | 1901–1927 |
| Compressed Powdered-Iron | 10.5 | 1916–1928 |
| Compressed Powdered-Permalloy | 33. | 1927–1938 |
| Compressed Powdered Molybdenum-Permalloy | 54. | 1937– |
| Non-Magnetic (Carrier loading) | 2. | 1920– |

Note (1): For more definite dates in relation to different types of facilities, and in relation to the two different permeability values of the iron-wire and iron-dust materials, reference should be made to Table III (page 158).

for general use. This total includes about 60% of the total production (through 1949) of all types of exchange area loading coils.

*Loaded Circuit Mileage Estimates*

To add some substance to the significance of the production statistics on voice-frequency loading, it is desirable to record some rough estimates regarding the aggregate length of the cable circuits which have been loaded.

For exchange area loading, a weighted average coil-spacing between the 6000 ft. and 3000 ft. values now standard can be assumed. Considering the time elements in the evolution of loading practices, as discussed in Part III of this review, it is reasonable to assume an average coil-spacing somewhat longer than the mean value of the two standard spacings, say about 5000 ft. On this assumption, the aggregate loaded cable-mileage which corresponds with an assumed production total of 11,200,000 coils is of the general order of 10,500,000 pair miles.

The 3000-ft. spacing has been used much less extensively on toll cable circuits than in the exchange plant, on which basis the weighted average coil-spacing for quadded toll cable loading is somewhat longer than the weighted average value for exchange area loading. Within the accuracy required for the present general estimates, 5500 ft. seems to be a reasonable estimate for the average coil spacing in quadded toll cable loading. On this basis, and assuming a production total of about 9,500,000 side circuit and phantom coils, the aggregate loaded toll cable circuit-mileage is of the order of 9,900,000 miles. Keeping in mind the substantially universal use of quadded cables and of phantom group loading for long-distance and inter-urban toll cables, the aggregate mileage of loaded toll cable quads is of the order of 3,300,000 miles. Because of the extensive installation of loaded H 44-25 four-wire repeatered circuits during the period 1925–1931, the *loaded "facility"* mileage-aggregate is considerably less than the *loaded "circuit"* mileage-figure above given. Meanwhile, much of the loaded H 44-25 4-wire circuit mileage has been converted for short haul two-wire circuit usage, and much has been unloaded to permit the operation of Type K carrier systems. The available data on these plant changes do not permit accurate estimates regarding the mileage of loaded four-wire and two-wire types of toll cable circuits now in commercial use. It is again appropriate, however, to call attention to the important part in the growth and improvement of the telephone service which the displaced loading coils played in their own period of commercial use.

## Economic Significance

Since loading has been used only when it permitted the use of cheaper facilities than would otherwise have been feasible, the great economic value of loading in the growth of the Bell Telephone System is indicated by the circuit mileage-figures given above. Other factors, however, would have to receive consideration in a complete appraisal, namely, the contributions of loading to nation-wide customer satisfaction that have resulted from improved transmission performance and higher speed of service. In turn, these factors themselves have been greatly influenced by the unit plant-cost reductions made possible by the use of loading.

For example, if loading had not been available when new or additional facilities became desirable, it is highly questionable as to whether it would have been economically feasible to work to the high-grade transmission-performance standards that have been readily achieved at reasonable costs with the cheaper loaded facilities. Moreover, it is even more questionable whether it would have been economically feasible to provide as many facilities without loading as were actually installed on a loaded basis.

Because of the speculative uncertainties involved in making assumptions regarding relative transmission-performance and relative plant-size, with and without loading, and because of the practical difficulties involved in evaluating in monetary terms the differences in transmission performance and in speed of service, no complete appraisal of the economic value of coil loading has ever been attempted for the exchange area plant. These, and additional special complications subsequently discussed, have also prevented accurate appraisals of the economic value of toll cable loading.

*Exchange Area Loading*

During the first two decades or so of the use of exchange area loading, rough estimates of its economic significance were sometimes made by comparing the total costs of the loaded facilities with the much higher cost of the non-loaded cable plant which otherwise would have been required to meet the same trunk-loss limits at 800 or 1000 cycles. Depending on the period under study, the estimated aggregate plant-cost reduction figures ranged up to and beyond $100,000,000. These estimates included the plant-cost reductions that resulted from the use of less expensive pole lines for aerial cables, and less expensive conduit systems made possible by utilizing a smaller total number of cables, each having a larger number of pairs. If similar studies should be made now, the corresponding hypothetical plant-cost reduction figure would probably be many times as large as the figure previously mentioned. These figures ignore the superior over-all transmission in loaded trunk plant that results from the much more favorable distortion characteristics. Also they assume equal sizes of trunk plant, with and without loading. Because of these qualifications, and because of the magnitude of the cost-reduction estimates, it is difficult to define their real significance.

A better understanding may perhaps be obtained from consideration of the cable data given in Table XXI, following. This compares some of the most important types of cable on which loading has been used with the types which would probably have been required for transmission reasons, if loading had not been available.

The large savings which loading permitted in the use of cable copper and in the amount of lead sheath per cable pair, are indirectly indicated by the tabulated data. Moreover, with loading on finer-wire cables a given total number of facilities can be provided with a much smaller total number of cables, thus permitting the use of less expensive conduit systems. This factor is extremely important in some routes of congested sections of large metropolitan areas such as Manhattan and the loop section in Chicago, where there might well be a question as to the *physical practicability*, dis-

regarding costs, of installing enough large-conductor, non-loaded cables to provide as many facilities as those made available in existing loaded small-conductor cables.

*Toll Cable Loading:* An accurate appraisal of the economic value of toll cable loading would have the specific complications mentioned above in the discussion of exchange area loading, and in addition certain intricate difficulties briefly discussed below.

In the aggregate, a very much larger amount of loading has been used on repeatered facilities than on non-repeatered voice-frequency circuits. The over-all plant-cost reduction and the transmission and speed of service

TABLE XXI
LOADED AND NON-LOADED EXCHANGE AREA CABLES
RELATIVE USE OF DIFFERENT TYPES

| Degree of Use[a] | Loaded Exchange Area Cable | | | Alternative Types of Non-Loaded Cables | | |
|---|---|---|---|---|---|---|
| | Conductor Size B & S ga. | Weight-Lbs.[1] Copper Pair-Mile | No. Pairs Full Size Cable | Conductor Size B & S ga. | Weight-Lbs.[1] Copper Pair-Mile | No. Pairs Full Size Cable |
| Very Extensive | 22 | 21.0 | 909[2] | 19 | 42.0 | 455[2] |
| | | | | 16 | 84.3 | 152[3] |
| Very Extensive | 24 | 13.2 | 1515[2] | 22 | 21.0 | 909[2] |
| | | | | 19 | 42.0 | 455[2] |
| Substantial | 19 | 42.0 | 455[2] | 16 | 84.3 | 152[3] |
| | | | | 13 | 168.8 | 75[3] |
| Small | 26 | 8.3 | 2121[2] | 24 | 13.2 | 1515[2] |
| | | | | 22 | 21.0 | 909[2] |

Notes: (1) These weights include a small allowance for the effect of pair-twist and stranding, in increasing the conductor length, relative to the cable sheath-length.
(2) High-capacitance cables–(approx.) 0.082 (±) mf/mi.
(3) Low-capacitance cables–(approx.) 0.066 mf/mi.
(a) In the very extensive installations of exchange area loading during the 1928–1949 period, a very large fraction of the total use was on 22 and 24-gauge cables in nearly equal quantities.

improvements that have resulted from the use of loading in combination with voice-frequency repeaters must of course be jointly credited to the repeaters and the loading. Since as yet no rationally acceptable procedure for allocating the pro-ratio credits has evolved, very questionable arbitrary allocations would become necessary. Moreover, very debatable uncertainties would be involved in making assumptions regarding the types of facilities which would have been employed if loading and repeaters could not have been jointly used on small-gauge toll cable conductors.

In appraising the economic importance of toll cable loading it is therefore necessary to revert to general terms, namely, its great extent of use as

indicated by the previously discussed production and circuit-mileage statistics.

In short-haul, non-repeatered, toll cable circuits, loaded 19 ga. conductors are generally used for service which would have required 16 or 13 gauge conductors without loading. The plant-cost savings in cable, copper, and lead are much greater per unit length than the average savings realized in the loaded exchange area cables. The aggregate mileage in this type of toll plant, however, is but a small fraction of that in the loaded exchange area plant.

Until the commercial exploitations of lower-cost carrier telephone systems started during the late 1930's, the loaded repeatered voice-frequency cable facilities satisfactorily met the quantitative and qualitative needs for the rapidly expanding long-distance telephone services along dense traffic routes where the use or the extension and expansion of the open-wire plant would have been unduly expensive, even on a carrier basis. In such backbone routes, and also along slow-growing tributary routes, and for short-haul toll facilities, the repeatered and non-repeatered loaded toll cables have provided more economical service than could have been obtained in an open-wire plant, and with increased dependability. Also, as previously indicated, larger circuit groups have been economically feasible, with valuable results as regards the speed of service.

In concluding this part of the review, it is noteworthy that the phantom-group loading almost universally used on voice-frequency repeatered and non-repeatered toll cable facilities is a major factor in the plant economies that have resulted from the commercial exploitation of the phantom working principle. These particular plant-cost savings constitute an important contribution to the aggregate economies achieved by toll cable loading.

## PART VIII: SUMMARY AND CONCLUSION

### *General*

The story of coil loading told in the present review is one of continuing evolution whereby its inherent capabilities have been substantially realized in its adaptation to the growing and changing needs of exchange area facilities and of interurban and long-distance communications by wires, throughout the Bell Telephone System. Also, full advantage has been taken of the opportunities offered by the development of better core-materials and new manufacturing techniques and tools to improve the loading apparatus and reduce its cost.

It was inevitable that by far the most important uses of coil loading would be for voice-frequency telephony over cable circuits. The very low

ratio of distributed inductance to distributed capacitance, incidentally resulting in low impedances, and the relatively high conductor resistances of cable circuits, gave loading its greatest opportunities in exercising its natural functions of reducing the circuit attenuation and attenuation-frequency distortion. Clearly appreciated from the beginning, these possibilities have been advantageously realized to a very great extent, and they still have substantial economic importance for future voice-frequency applications in the continuing growth of the exchange area non-quadded cable plant, and short, quadded interurban toll cables.

*Open-Wire Loading*

The higher ratios of distributed inductance to distributed capacitance in the open-wire lines made the reduction of attenuation-frequency distortion a relatively minor objective in the use of loading, attenuation reduction being the primary objective. Incidentally, the relatively high impedances of the non-loaded lines that resulted from their higher ratios of inductance to capacitance limited the attenuation reduction obtainable by coil loading to smaller percentage values than those obtainable on cable circuits. However, full advantage of these important, though limited, possibilities was realized in the expanding open-wire plant during the decade that preceded the commercial introduction of vacuum-tube repeaters. The early uses of these repeaters on open-wire lines were on circuits having improved loading designed especially for use in conjunction with repeaters. In 1915, this combination of loading and repeaters made transcontinental telephony economically feasible, and for several years greatly increased the demand for loading. The importance of open-wire loading soon started to decline, however, as a result of improvements in the repeaters, their circuits, and auxiliary networks, which made it possible to secure considerably better voice-frequency transmission on long lines at a lower total cost by discarding loading and using more repeaters. The climactic event in this new trend was the beginning of the operation of the first transcontinental circuits on a non-loaded basis during 1920. During the middle and late 1920's the general removal of open-wire loading was expedited to increase the plant flexibility and facilitate the commercial exploitation of carrier telephone and telegraph systems over non-loaded lines.

Since, for transmission-cost reasons, it is not feasible to develop suitable loading for long lines over which carrier systems are operated, there is no reason to expect any new leases of life for open-wire coil loading. Notwithstanding its small extent of use relative to that for cable loading, and the relatively short period during which it was standard practice, open-wire loading was a necessary and a vitally important factor in the rapid expansion of long-distance telephony that began nearly five decades ago.

*Toll Cable Loading*

The pattern of the commercial evolution of loading practices for long-distance cable systems has been generally similar to that for open-wire loading, but with important quantitative and qualitative differences, and especially in the relative time-elements. These various differences have been mainly due to the previously mentioned inherent differences in the basic transmission properties of non-loaded cables and non-loaded open-wire lines.

Prior to the availability of vacuum-tube repeaters, loading was an essential factor in the establishment of a very important expanding network of storm-proof, intercity, toll cables; coarse-gauge conductors and expensive coils were used for distances ranging up to about 250 miles, 16 ga. conductors and less expensive coils being satisfactory for terminal business over short distances. Without using loading, these early toll cable systems would not have been economically feasible.

In the early uses of repeaters on toll cables the cable circuits also used loading. These combinations permitted improved transmission performance and important extensions in transmission range. In this general connection, it is of interest to note that it was not economical to use non-loaded conductors for toll cable transmission until cable carrier telephone systems became available about two decades after the commercial introduction of the vacuum-tube repeater. For voice-frequency transmission, the use of repeaters without loading would have been unduly expensive, due to the high costs of the additional repeaters and the much more expensive distortion-correcting networks and regulating networks that would have been required.

In the early part of the period that intervened between the introduction of vacuum-tube repeaters and of cable carrier systems, the substantially continuous development of improved loading, and of improved repeaters and auxiliary equalizing and regulating networks, provided improved facilities of several different types especially proportioned on a minimum cost basis to meet the transmission-service needs of different geographical distances.

High-velocity, four-wire, H 44-25 19 gauge circuits were very extensively used for long-haul facilities ranging up to about 2000 miles in length. It is of interest that the timely completion of the development of the first cable-carrier system stopped the contemporary efforts to make additional improvements in the H 44-25 voice-frequency loaded four-wire circuits so that they would be suitable for transcontinental distances. These improvements would have involved the use of velocity distortion corrective networks.

Nineteen gauge two-wire circuits having lower-velocity, higher-impedance, loading than that employed on the above mentioned four-wire

circuits were very extensively used for short-haul repeatered and non-repeatered facilities.

A large curtailment in the demand for loading on new long cable circuits immediately followed the commercial exploitation of the Type K cable-carrier system, which started during the middle 1930's. The drastic nature of this impact was subsequently increased by the standardization of a still more economical (K2) cable carrier system,[48] and by the post-war extensive installation of coaxial cable systems. The very recent development of a relatively inexpensive short-haul carrier system (Type N), which uses two pairs in the same cable for its opposite-direction paths, promises an additional substantial reduction in the need for new loaded toll cable facilities, even for short distances. However, it seems probable that the demand for new loading may continue indefinitely on a low-level basis for more or less special short-haul situations where carrier telephony may be more expensive.

During the past two decades or so, loading cost-reduction has been carried so far that the prospects of further substantial cost-reductions are not now in sight. It seems improbable that any further design cost-reduction could be large enough to reverse the present general trend towards a large dependence upon carrier telephony for new short-haul toll cable facilities.

*Exchange Area Loading*

During the period covered by the present review, telephone transmission over exchange area cables has been entirely on a voice-frequency basis. Moreover, the use of vacuum-tube repeaters in conjunction with loading (or on non-loaded cables) has been statistically insignificant in comparison with the very extensive use of loading. In consequence, exchange area loading does not have to share with developments in repeaters and in carrier systems the great credit which it has earned with respect to the improvement of exchange area transmission performance and the reduction of plant cost.

The simple pattern in the evolution of exchange area loading practices, relative to those for toll cable loading, is of course basically due to the shortness of the circuits and the relatively uncomplicated service-requirements.

In certain important respects, the improvements achieved by the nearly continuous development work are generally similar in the two types of loading, notably: (1) the improvement in transmission quality obtained by increasing the transmission band-width, and (2) the successive facility-cost reductions resulting from the successive developments of lower-cost loading apparatus. These plant-cost reduction activities were carried out to a greater degree in the exchange area loading. It is especially noteworthy

that the most important apparatus-cost reduction developments were completed in time for exploitation during periods of peak demand for new coils.

With respect to the effects of other developments in reducing the demand for exchange area loading, the introduction of improved subscriber sets during the 1930's warrants special mention. By permitting higher losses in the trunks, somewhat longer non-loaded trunks could be used.

Looking towards the future, the prospective use of a new low-cost repeater of an entirely new type (El telephone repeater) is expected to reduce the demand for the heavier weights of loading. Also, the new Type N short-haul cable carrier system, referred to on page 1240, may have some considerable use on relatively long non-loaded exchange trunks along heavy traffic routes. It is also of interest that a greatly improved telephone set (500-type) now in the final stages of development will probably reduce the need for loading on long subscriber loops.

Although it is not possible at present to make accurate quantitative estimates of the ultimate effects of the just mentioned new developments upon the future demand for new exchange area loading, there is no reason to believe they will be so drastic as the effects of carrier system developments upon the ultimate future demand for toll cable loading. It seems especially probable that the low-cost H-spaced loading will continue indefinitely to be an important factor in the economy of design of new exchange area cable plant to provide telephone service for a continually increasing number of subscribers.

*Loading for Incidental Cables in Open-Wire Lines*

The impedance-matching loading systems used on entrance and intermediate cables have made vital contributions to the excellence of the over-all performance of the open-wire transmission systems. These are of great importance relative to the amount and the cost of the loading actually used.

In consequence of the increasing utilization of open-wire carrier systems, the voice-frequency loading is much less important than it was two to three decades ago. However, an indefinitely continuing, though small, demand seems certain, because of the valuable transmission improvements which the loading makes available at low cost.

The demand for additional carrier loading is expected to continue in a somewhat rough proportion to the number of additional open-wire carrier systems that are installed. However, in consequence of the high cost of the loading for multi-channel systems (which is much higher than formerly in consequence of greatly increased labor and material costs), it seems probable that more and more consideration will be given (especially on "long"

incidental cables) to the use of lower-cost transmission-improvement treatments, even though they are not so good as loading in certain respects.

*Cable Program Circuit Loading*

During the 1930's and early 1940's, there were extensive applications of loading on the cable sections of nation-wide chain networks used for transmitting AM broadcast program material. Now that high-grade program transmission circuits may be obtained by carrier methods on broadband cable carrier systems, the future demand for 8-kc loaded cable program circuits will be largely limited to special situations where the carrier program circuits are not economical.

It is expected also that there will be a moderate, continuing demand for the recently developed loading that provides a 15-kc band for the transmission of FM program broadcast material, principally on studio-transmitter circuits, and on end links in toll cable networks, where carrier program circuits may be uneconomical.

*Continuous Loading*

Over the years, a substantial amount of exploratory development work on continuous loading for ordinary types of paper-insulated cable has been done, but with negative results so far as commercial applications in the Bell System are concerned; it has not yet been found feasible to compete with coil loading in service performance and cost.

However, continuous loading has had a few applications in single core submarine cables, in deep water installations where coil loading is not feasible. The three 1921 cables between Key West and Havana are the only continuously loaded cables to become a part of the Bell System. They use iron wire as the loading material. Several years later, permalloy tape continuous loading developed by the Bell Telephone Laboratories made possible a great increase in the message-carrying capacity of transoceanic telegraph cables. During the middle 1920's, an aggregate of about 15,000 nautical miles of the new type, high speed, cable was installed for use by non-affiliated telegraph and cable companies.

Late in the 1920's, a perminvar type loaded cable suitable for voice-frequency telephony between Newfoundland and Ireland was developed by the Bell Telephone Laboratories. The business depression of the early 1930's intervened to cause a temporary postponement of the project; later on, an indefinite postponement resulted from improvements in transatlantic radio-telephony.

From the foregoing, it is clear that the importance of continuous loading has been low relative to that of coil loading in the growth of the Bell Telephone System.

## Conclusion

During the half century that has intervened since its invention, coil loading has played a very important part in making nation-wide telephony possible and in helping to make possible the great growth in the business which has occurred. Although the application of coil loading to new circuits has now been greatly curtailed, due in large part to the development of carrier systems, coil loading still has an important field of application in exchange area telephone plant and for some rather special circuit applications.

The reader may take it for granted that the organization which has developed and used loading to the maximum degree of utility in the present telephone plant will be on the alert in the future to make full use of loading in situations wherever loaded circuits provide a more economical solution of the transmission service needs than the other available procedures. It is also reasonable to expect that new types of loading and new loading apparatus will be developed to the extent that may be economically warranted.

## Bibliography (*Concluded*)


23. H. D. Arnold and G. W. Elmen, "Permalloy, An Alloy of Remarkable Magnetic Properties," *Journal of the Franklin Institute*, Vol. 195, 1923.
42. W. H. Martin, G. A. Anderegg, and B. W. Kendall, "Key West-Havana Submarine Cable System," *Trans. A.I.E.E.*, Vol. XLI, 1922.
43. H. A. Affel, W. S. Gorton, and R. W. Chesnut, "A New Key West-Havana Carrier Telephone Cable," *B.S.T.J.*, Vol. XI, April 1932.
44. O. E. Buckley, "The Loaded Submarine Telegraph Cable," *B.S.T.J.*, Vol. IV, July 1925; *Electrical Communication*, Vol. 4, No. 1, 1925, *Journal A.I.E.E.*, Vol. XLIV, No. 8, 1925.
45. O. E. Buckley, "High Speed Ocean Cable Telegraphy," *B.S.T.J.*, Vol. VII, April 1928. Presented at the International Congress of Telegraphy and Telephony in Commemoration of Volta, Lake Como, Italy, September 1927.
46. G. W. Elmen, "Magnetic Alloys of Iron, Nickel and Cobalt," *Jl. Franklin Institute* Vol. 207, p. 583, 1929.
47. O. E. Buckley, "The Future of Transoceanic Telephony," The Thirty-third Kelvin Lecture of the Institution of Electrical Engineers, April 23, 1942; *The Journal of The Institution of Electrical Engineers*, Vol. 89, Part 1, 1942. (Bell Laboratories reprint, Monograph B-1346).
48. H. S. Black, F. A. Brooks, A. J. Wier and J. G. Wilson, "An Improved Cable Carrier System," *Trans. A.I.E.E.*, Vol. 66, 1947.

In addition to the published articles referred to in the text or footnotes, the following will be of interest:

W. Fondiller, "Commercial Loading of Telephone Cable," *Electrical Communication*, Vol. 4, No. 1, July 1925.

George Crisson, "Irregularities in Loaded Telephone Circuits," *B.S.T.J.*, Vol. IV, October 1925.

F. L. Rhodes, "Beginnings of Telephony," Harper and Brothers, New York, 1929.

L. G. Abraham, "Circulating Currents and Singing on Two-Wire Cable Circuits," *B.S.T.J.*, Vol. XIV, October 1935.

L. L. Bouton, "Four-Wire Circuits in Retrospect," *Bell Lab. Record*, December 1938.

S. G. Hale, "Splice Loading Developments," *Bell Lab. Record*, January 1951.

## Abstracts of Bell System Technical Papers Not Published in This Journal

*A Full Automatic Private Line Teletypewriter Switching System.** W. M. BACON[1] and G. A. LOCKE.[1] *Elec. Engg.*, v. 70, pp. 408–413, May, 1951.

ABSTRACT—A full automatic teletypewriter message switching system has been developed for use in private line networks involving one or more switching centers and a multiplicity of local or long distance lines, each of which may have one or more stations. This system provides fast teletypewriter communication from any station to any other station or group of stations in the network.

*Crossbar Tandem System.** R. E. COLLIS.[1] *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 997–1004, 1950.

*A Study of Nuclear and Electronic Magnetic Resonance.** K. K. DARROW.[1] *Elec. Engg.*, v. 70, pp. 401–404, May, 1951.

ABSTRACT—Since the discovery of magnetic resonance in solids, liquids, and gases in 1945, the phenomenon has been used in the determination of nuclear magnetic moments and magnetic field strengths, as well as in the study of crystal structure and relaxation times.

*The Genesis of Submarine Cables.* L. ESPENSCHIED.[1] Bibliography. *Elec. Engg.*, 70, pp. 379–383, May, 1951.

ABSTRACT—It was a century ago that the first submarine cable was laid between Dover and Calais. To mark this centenary the author reviews some of the events leading up to this achievement which made possible further advances in the communications field, such as laying of the transatlantic cable by the Great Eastern escorted by four ships, as shown in the picture.

*Borocarbon Film Resistors.* R. O. GRISDALE,[1] A. C. PFISTER[1], and G. K. TEAL.[1] *Natl. Electronics Conference, Proc.*. v. 6, pp. 441–442, 1950.

ABSTRACT—The carbon film type of resistor is particularly useful at high frequencies, for not only can it be made to have small reactance but it is, in effect, all skin so that there is no increase in resistance at high frequencies due to skin effect. The film is also well cooled through its intimate contact with the core and this makes possible the dissipation of large amounts of power per unit area. While primarily developed for high frequency applications in this country, the pyrolytic carbon resistor possesses other

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

characteristics which have led and are leading to greatly expanded fields of application. Principal among these are the tolerances of one per cent or better attainable in production, the stability in use, the relatively small and predictable temperature coefficient of resistance, and the low noise level. These properties result in large part from the ultimate crystalline structure of the carbon films.

*Some Methods of Solving Hyperbolic and Parabolic Partial Differential Equations.* R. W. HAMMING.[1] International Business Machines Corp. Computation seminar. *Proceedings, Dec., 1949, Ed. by C. C. Hurd. N. Y., I.B.M.*, pp. 14–23, 1951.

ABSTRACT—The main purpose of this paper is to present a broad, nonmathematical introduction to the general field of computing the solutions of partial differential equations of the hyperbolic and parabolic types, as well as some related classes of equations. I hope to show that there exist methods for reducing such problems to a form suitable for formal computation, with a reasonable expectation of arriving at a usable answer.

I have selected four particular problems to discuss. These have been chosen and arranged to bring out certain points which I feel are important. The first problem is almost trivial as there exist well-known analytical methods for solving it, while the last is a rather complicated partial differential-integral equation for which there is practically no known mathematical theory.

*Electrography and Electro-Spot Testing.**H. W. HERMANCE[1] and H. V. WADLOW.[1] *Physical Methods in Chemical Analysis; Ed. by W. G. Berl. N. Y., Academic Press*, v. 2, pp. 155–228, 1951.

*Correlation Energy and the Heat of Sublimation of Lithium.* C. HERRING.[1] Letter to the editor. References. *Phys. Rev.*, v. 82, pp. 282–283, Apr. 15, 1951.

*Some Theorems on the Free Energies of Crystal Surfaces.** C. HERRING.[1] References. *Phys. Rev.*, v. 82, pp. 87–93, Apr. 1, 1951.

ABSTRACT—Although the interpretation of experiments in such fields as the shapes of small particles and the thermal etching of surfaces usually involves problems of kinetics rather than mere equilibrium considerations, it is suggested that a knowledge of the relative free energies of different shapes or surface configurations may provide a useful perspective. This paper presents some theorems on these relative free energies which follow from the Wulff construction for the equilibrium shape of a small particle, and some relations between atomic models of crystal surfaces and the surface free energy function used in this construction. Equilibrium shapes of

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

crystals and of non-crystalline anisotropic media are classified, and it is pointed out that the possibilities for crystals include smoothly rounded as well as sharp-cornered forms. The condition is formulated for thermodynamic stability of a flat crystal face with respect to formation of hill-and-valley structure. A discussion is presented of the limitations on the applicability of the results imposed by the dependence of surface free energy on curvature; and it is concluded that these limitations are not likely to be serious for most real substances, though they are serious for certain idealized theoretical models.

*The Crystal Structures of $NiO \cdot 3BaO$, $NiO \cdot BaO$, $BaNiO_3$ and Intermediate Phases With Composition Near $Ba_2Ni_2O_5$; With a Note on NiO.** J. J. LANDER.[1] References. *Acta Cryst.*, v. 4, pp. 148–156, Mar., 1951.

ABSTRACT—The crystal structures of $NiO \cdot 3BaO$, $NiO \cdot BaO$ and $BaNiO_3$ have been determined from X-ray diffraction data, and data are given for phases with composition near that represented by $Ba_2Ni_2O_5$. In each of these structures nickel behaves in a novel fashion. A coplanar triangular arrangement of oxygen around nickel is found in $NiO \cdot 3BaO$. In $BaNiO_3$ nickel has a valence of four and the structure is a close-packed hexagonal stacking of planar arrangements found in perovskite 111 planes. The compound $NiO \cdot BaO$ has a magnetic moment corresponding to two unpaired electrons, whereas the deduced coplanar square arrangement of oxygen around nickel suggests that there should be no unpaired electrons. Compounds with composition near $Ba_2Ni_2O_5$ contain an amount of oxygen which is a continuous function of temperature and possibly contain mixtures of bi- and tetravalent nickel.

The problem of NiO having octahedral co-ordination of oxygen is considered.

*New Ferroelectric Tartrates.* B. T. MATTHIAS[1] and J. K. HULM. Letter to the editor. *Phys. Rev.*, v. 82, pp. 108–109, April 1, 1951.

*A Negative Impedance Repeater.** J. L. MERRILL, JR.[1]. *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 1461–1466, 1950.

*Interexchange Tandem Trunking in the Los Angeles Metropolitan Area.* W. F. PFEIFFER[1]. *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 1071–1079, 1950.

ABSTRACT—Twenty-four years have elapsed since the first large-scale machine switching tandem system was designed and installed for service in Los Angeles. As an intermediate switching center, the tandem office enabled operators to use the dial method of operation for establishing interexchange telephone connections over the associated trunking network. During the intervening years, it has facilitated the rapid handling of tele-

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

phone calls between the various communities in and around the city. Step-by-step tandem equipment was employed and, as the volume of calls grew, the trunk capacity was increased by installing additional switching equipment. In 1946 it became evident that the abnormal rate of growth required additions substantially beyond the practical size limit of the step-by-step tandem unit. To solve the resulting problem, it became necessary to reorganize the tandem trunking system and select a multiunit tandem switching plan. It also provided an opportunity to consider the application of the more recently developed crossbar tandem switching system. This paper reviews the factors affecting the general problem of interexchange trunking which have led to the development of the present tandem network in the Los Angeles metropolitan area. It describes the major elements of a system which now employs a total of five tandem switching units, three of which are crossbar tandem offices.

*p-n Junction Rectifier and Photo-cell.* W. J. PIETENPOL[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 120–121, Apr. 1, 1951.

*Formulas for the Determination of Residual Stress in Wires by the Layer Removal Method.** W. T. READ, JR.[1]. *Jl. Applied Phys.*, v. 22, pp. 415–416, Apr., 1951.

ABSTRACT—The distribution of residual axial stress in a beam or wire of circular cross section is derived as a function of the moment required to straighten the wire after removal of successive layers of material. Application of the formulas involves two graphical differentiations and integrations of experimental curves.

*Observation of Magnetic Domains by the Kerr Effect.* H. J. WILLIAMS[1], F. G. FOSTER[1], and E. A. WOOD[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 119–120, Apr. 1, 1951.

*Particle Size in Suspension Polymerization.** F. H. WINSLOW[1] and W. MATREYEK[1]. Bibliography. *Ind. & Engg. Chem.*, v. 43, pp. 1108–1112, May, 1951.

ABSTRACT—Control of size and geometrical form of densely cross-linked hydrocarbon polymers yields fluid spherical powders useful as dielectrics and in rheological studies. Such studies also bear on polymer forms important in ion exchange resins.

Several significant factors influencing the preparation of polymer spheroids have been established on a semi-quantitative basis: Polyvinyl alcohol proved to be a highly efficient stabilizer for polymer spheroid preparations. Under comparable conditions, (a) high molecular weight grades, (b) partially hydrolyzed grades, and (c) high concentrations of stabilizer were

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

associated with spheroids of lower mean diameters. These generalizations cover suspension stabilization down to roughly 0.1% stabilizer. The concentration limits where suspending action begins are, however, of special interest. Here it was found that the number of polyvinyl alcohol molecules present became important—that is, for equal weight concentrations in the vicinity of 0.005%, low molecular weight polymer (19,000) produced stabilized (although large) spheres whereas the usual high molecular weight polymer (95,000) was ineffective.

Close to the maximum possible yield of well-formed spheroids was reproducibly obtained in narrow size distribution and with average spheroid diameters ranging from 5 microns to several millimeters in diameter—a thousand-fold variation in dimensions.

*Elastic and Electromechanical Coupling Coefficients of Single-Crystal Barium Titanate.* W. L. BOND[1], W. P. MASON[1], and H. J. MCSKIMIN[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 442–443, May 1, 1951.

*Making Small Spheres.* W. L. BOND[1]. *Rev. Sci. Instruments*, v. 22, pp. 344–345, May, 1951.

*Submarine Telephone Cable With Submerged Repeaters.* J. J. GILBERT[1]. *Electronics*, v. 24, pp. 164, 168, 172+, June, 1951.

*Electrode Reactions in the Glow Discharge.** F. E. HAWORTH[1]. References. *Jl. Applied Phys.*, v. 22, pp. 606–609, May, 1951.

ABSTRACT—The reactions which occur at silver electrodes in a normal glow discharge in air have been determined. These are: (1) formation of $AgNO_2$ and some $Ag_2O$ at the anode at the rate of 3.4 $\mu$g/coulomb; (2) loss of metal from the cathode by chemical action at the rate of 3.5 $\mu$g/coulomb (probably the same reaction as (1) with subsequent loss of the reaction products by the greater heating of the cathode, but this hypothesis has not been established); and (3) normal sputtering loss at the cathode at the rate of 0.4 $\mu$g/coulomg. These processes result in building a conducting layer on the anode. If the electrode separation is so small that the anode extends into the region of the cathode fall, then the high electric field pulls the newly formed and not very coherent growth upon the anode across into a bridge between the electrodes.

*Storing Video Information.**A. L. HOPPER[1]. *Electronics*, v. 24, pp. 122–125, June, 1951.

ABSTRACT—Comparison of signal amplitudes along adjacent television scanning lines can be made by storing the video information of one line for 63.5 microseconds. Storage is done in an ultrasonic delay line employing a fused silica bar with quartz transducers.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

*Cross Sections for Ion-Atom Collisions in He, Ne, and A.* J. A. HORNBECK[1] and G. H. WANNIER[1]. Letter to the editor. *Phys. Rev.*, v. 82, p. 458, May 1, 1951.

*Ferromagnetic Resonance.**C. KITTEL[1]. Bibliography. *Jl. de Physique*, v. 12, pp. 291–302, Mar., 1951.

*Theory of Antiferroelectric Crystals.* C. KITTEL[1]. References. *Phys. Rev.*, v. 82, pp. 729–732, June 1, 1951.

ABSTRACT—An antiferroelectric state is defined as one in which lines of ions in the crystal are spontaneously polarized, but with neighboring lines polarized in antiparallel directions. In simple cubic lattices the antiferroelectric state is likely to be more stable than the ferroelectric state. The dielectric constant above and below the antiferroelectric curie point is investigated for both first- and second-order transitions. In either case the dielectric constant need not be very high; but if the transition is second order, e is continuous across the Curie point. The antiferroelectric state will not be piezoelectric. The thermal anomaly near the Curie point will be of the same nature and magnitude as in ferroelectrics. A susceptibility variation of the form $C/(T + Z)$ as found in strontium titanate is not indicative of antiferroelectricity, unlike the corresponding situation in antiferromagnetism.

*Theory of Antiferromagnetic Resonance* C. KITTEL[1]. Letter to the editor. *Phys. Rev.*, v. 82, p. 565, May 15, 1951.

*Barium-Nickel Oxides With Tri- and Tetravalent Nickel.** J. J. LANDER[1] and L. A. WOOTEN[1]. *Am. Chem. Soc., Jl.*, v. 73, pp. 2452–2454, June, 1951.

ABSTRACT—The compound $BaNiO_3$ and intermediates with composition ranging between $Ba_3Ni_3O_8$ and $Ba_2Ni_2O_5$ have been prepared. $BaNiO_3$ is black, stable in alkali, and has a structure made up of layers identical with the 111 planes of a perovskite but stacked in a close-packed hexagonal fashion. At 730° in 730 mm. of oxygen, the structure changes to that associated with the series $Ba_3Ni_3O_8$ to $Ba_2Ni_2O_5$ in which the oxygen content appears to decrease continuously with temperature increasing to 1200°, at which point sharp melting is observed. These materials are black and stable in alkali with an hexagonal structure for which the details have not been determined. Resistivities and magnetic susceptibilities are reported. A wide range in composition, temperature and reaction atmosphere was studied but only one additional compound was observed. Attempts to isolate this compound were not successful.

*The Phase System BaO–NiO.** J. J. LANDER[1]. *Am. Chem. Soc., Jl.*, v. 73, pp. 2450–2452, June, 1951.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

ABSTRACT—The phase system BaO–NiO has been studied largely by means of X-ray diffraction. The two compounds NiO BaO and NiO 3BaO occur in the system. Their preparation and properties are described. NiO BaO is black, stable in air, orthorhombic, and melts at 1240°. NiO 3BaO is gray-green, unstable in air, hexagonal, and melts at 1160°. A eutectic melting at 1080° is observed between these compounds, but none between NiO 3BaO and BaO. Intersolubility of all solid phases in the system is small, even at high temperatures, but quantitative data have not been obtained.

*A Phenomenological Derivation of the First- and Second-Order Magnetostriction and Morphic Effects for a Nickel Crystal.** W. P. MASON[1]. References. *Phys. Rev.*, v. 82, pp. 715–723, June 1, 1951.

ABSTRACT—In order to account for experimental results which showed that the saturation elastic constants of a single nickel crystal varied with the direction of magnetization, a phenomenological investigation has been made of the stress, strain, and magnetic relations for single nickel crystals. The variation in elastic constants is shown to be a "morphic" effect caused by the change in the crystal symmetry due to the magnetostriction effect. In the energy equation this effect is represented by additional terms which involve squares and products of both the magnetic intensities and stresses. These terms are as large as the magnetostrictive terms when the stresses are of the order of $10^{10}$ dynes/cm$^2$. The energy equation has been used to derive the first- and second-order magnetostrictive effect, and the resulting terms agree with Becker and Döring's empirical constants for saturation conditions. For smaller magnetic intensities the terms divide up into first- and second-order terms which vary differently with magnetic field intensity. It is shown that the morphic effects involve six measurable constants, and some of these are evaluated experimentally.

*Dielectric Properties of Sodium and Potassium Niobates.** B. T. MATTHIAS[1] and J. P. REMEIKA[1]. *Phys. Rev.*, v. 82, pp. 727–729, June 1, 1951.

ABSTRACT—The following paper deals with evidence of ferroelectricity in $KNbO_3$ and $NaNbO_3$. Temperatures at which both materials undergo crystallographic changes and corresponding changes in dielectric constant and loss tangent are reported. Photographs of dielectric hysteresis loops and values of saturation polarization taken at various points over a temperature range are given for $KNbO_3$.

*Ferroelectricity.* B. T. MATTHIAS[1]. Bibliography. *Science*, v. 113, pp. 591–596, May 25, 1951.

ABSTRACT—Under the name of Ferroelectrics one classifies those materials which exhibit dielectric anomalies phenomenologically similar to the mag-

* A reprint of this article may be obtained on request.

Bell Tel. Labs.

netic behavior of the ferromagnetics. Perhaps it would have been more logical to use the term Rochelle electrics, thus emphasizing the similarity in the dielectric behavior to that of Rochelle salt, for which this behavior was first discovered by J. Valasek.

In this discussion the known ferroelectrics will be listed, and the various theories that have been created to explain them will be examined.

*Theory of Ferroelectric Behavior of Barium Titanate.* P. W. ANDERSON[1]. References. *Ceramic Age*, v. 57, pp. 29–30, 33+, April, 1951.

*Criterion for Superconductivity.* J. BARDEEN[1]. Letter to the Editor. *Phys. Rev.*, v. 82, pp. 978–979, June 15, 1951.

*Magnetic Domain Patterns.** R. M. BOZORTH[1]. Bibliography. *Jl. de Physique*, v. 12, pp. 308–321, March, 1951.

*Electron Temperature vs Noise Temperature in Low Pressure Mercury-Argon Discharges.* M. A. EASLEY[1] and W. W. MUMFORD[1]. Letter to the Editor. *Jl. Applied Phys.*, v. 22, pp. 846–847, June, 1951.

*The Origin of Bombardment-Enhanced Thermionic Emission.** J. B. JOHNSON[1]. References. *Phys. Rev.*, v. 83, pp. 49–53, July 1, 1951.

ABSTRACT—Measurements on bombardment-enhanced thermionic emission from oxide cathodes show that (a) the effect is not related to normal fading and recovery of thermionic emission; (b) the emitted electrons have energies in the thermal range rather than in the secondary range. Calculations indicate that the electron bombardment releases more than enough internal secondaries to account for the effect as increased thermionic emission. A more comprehensive theory is needed for explaining why the observed effect is not even larger.

*Dipolar Domains in Paramagnetic Crystals at Low Temperatures.* C. KITTEL[1]. Letter to the Editor. *Phys. Rev.*, v. 82, pp. 965–966, June 15, 1951.

*Methods of Measuring Adjacent-Band Radiation from Radio Transmitters.** N. LUND[1]. *I.R.E. Proc.*, v. 39, pp. 653–656, June, 1951.

ABSTRACT—A review of three possible methods of measuring or estimating adjacent-band radiation characteristics of a radio transmitter is given. These three methods differ in the type of signal applied to the transmitter and may be termed the two-tone, normal signal, and thermal noise methods. Measurements on a multichannel single-sideband transmitter using each of these methods are presented to show that there is a good correlation between the normal signal and thermal noise methods.

An empirical method for calculating the slope of the adjacent-band radiation as a function of frequency from the measured two-tone distortion values

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

is given, and the measured and calculated slopes are shown to be in fairly good agreement.

*Microwave Spectrum in $NO_2$.* K. B. Mc Afee, Jr.[1]. Letter to the Editor. *Phys. Rev.*, v. 82, p. 971, June 15, 1951.

*A Simple Electronic Differential Analyzer as a Demonstration and Laboratory Aid to Instruction in Engineering.* M. H. Nichols[1] and D. W. Hagelbarger[1]. *Jl. Engg. Education*, v. 41, pp. 621–630, June, 1951.

*Telecommunications.* H. S. Osborne[1]. *Ordnance*, v. 36, pp. 87–90, July–August, 1951.

*Triangular Permutation Numbers.* J. Riordan[1]. References. *Am. Math. Soc., Proc.*, v. 2, pp. 429–432, June, 1951.

*Measurements of Dynamic Internal Dissipation and Elasticity of Soft Plastics.** H. C. Rorden[1] and A. Grieco[1]. *Jl. Applied Phys.*, v. 22, pp. 842–845, June, 1951.

Abstract—In order to measure the mechanical properties of soft plastics over wide frequency and temperature ranges two new techniques have been devised. The first one, which operates in the frequency range of a few cycles, uses a horizontal oscillating pendulum. The shear impedance of the sample is measured by mounting a small pad of the material between the vibrating pendulum and a fixed platform and determining the change in frequency and the change in the decrement caused by the sample. From these measurements the shear mechanical resistance and reactance of the specimen can be determined. The other technique, which is applicable in the frequency range from 100 cycles to 10,000 cycles, makes use of a vibrating tuning fork. Two identical samples are mounted between a stationary weight and the moving tines, and the shear mechanical impedance is determined by determining the change in frequency and change in decrement caused by the specimen. These two techniques have been applied to measuring the shear properties of a number of soft plastics including Pyralin, Koroseal, Keldur, polyvinyl butyral, Thiokol, and gum rubber. All of these show relaxation effects. The polyvinyl butyral appears to be approaching a crystalline elastic stage at the low frequency of 1000 cycles, while gum rubber remains in a quasi-configurational stage from 2 cycles to 1000 cycles.

*The Mobility of Electrons in Silver Chloride.** J. R. Haynes[1] and W. Shockley[1]. References. *Phys. Rev.*, v. 82, pp. 935–943, June 15, 1951.

Abstract—Techniques are described which utilize the "print out effect" to obtain both the direction and velocity of photoelectrons in silver chloride crystals in an electric field. Hall mobility of the electrons is calculated from their change in direction produced by crossed electric and magnetic fields.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

Drift mobility of the electrons is obtained by measurement of their velocity in known electric fields. The value obtained for the Hall mobility ($R\sigma$) multiplied by $8/3\pi$ is 51 $cm^2$/volt sec at 25°C. The values obtained for the drift mobility are shown to be a function of temperature. A value of 49.5 $cm^2$/volt sec was obtained at 25°C, which is within experimental error of $(8/3\pi)R\sigma$, indicating that acoustical scattering is the principal mechanism and that temporary trapping is unimportant. A summary of the behavior of conduction electrons in silver chloride, calculated from the results of these experiments, is included.

*p-n Junction Transistors.** W. SHOCKLEY[1], M. SPARKS[1], and G. K. TEAL[1]. References. *Phys. Rev.*, v. 83, pp. 151–162, July 1, 1951.

ABSTRACT—The effects of diffusion of electrons through a thin p-type layer of germanium have been studied in specimens consisting of two n-type regions with the p-type region interposed. It is found that potentials applied to one n-type region are transmitted by diffusing electrons through the p-type layer although the latter is grounded through an ohmic contact. When one of the p-n junctions is biased to saturation, power gain can be obtained through the device. Used as "n-p-n transistors" these units will operate on currents as low as 10 microamperes and voltages as low as 0.1 volt, have power gains of 50 db, and noise figures of about 10 db at 1000 cps. Their current-voltage characteristics are in good agreement with the diffusion theory.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

## Contributors to This Issue

B. S. Biggs, B.A., Southwest Texas Teachers College, 1927; M.A., University of Texas, 1931, Ph.D., 1933; Civil Research Laboratory, Carnegie Institute of Technology, 1933–1936. Bell Telephone Laboratories, 1936–. With the Laboratories he has worked chiefly on the synthesis of wood preservatives, on dielectric materials and on other phases of organic chemistry. He is a member of the American Chemical Society and of Sigma Xi.

G. T. Ford, B.S., Michigan State College, 1929; M.A., Columbia, 1936. Bell Telephone Laboratories, 1929–. With the Laboratories he has worked on gas tubes, thermistors, general vacuum tube development, and electron tubes for broad band amplifiers. He is a member of the Institute of Radio Engineers.

R. W. Friis, B.E.E., University of Minnesota, 1930. Bell Telephone Laboratories, 1930–. With the Laboratories Mr. Friis has been concerned with transoceanic and ship-to-shore radio telephone, fire-control radio transmitters, and the microwave radio-relay system. He is a Senior Member of the Institute of Radio Engineers.

J. C. Lozier, B.A., Columbia, 1934. R.C.A. Mfg. Co., 1935–1936. Bell Telephone Laboratories, 1936–. Mr. Lozier's work with the Laboratories has been principally transmission development for radio and carrier telephone systems, the theory and design of servomechanisms, and the theory of feedback systems such as compandors and regulators. He is a Senior Member of the Institute of Radio Engineers.

R. C. Prim, III, B.S.E.E., University of Texas, 1941; M.A. and Ph.D., Princeton, 1949. General Electric Company, 1941–44; Naval Ordnance Laboratory, 1944–49. Bell Telephone Laboratories, 1949–. Here his work has been chiefly mathematical research on non-linear partial differential equations and as a consultant on military projects. Dr. Prim is a member of the Amer. Math. Soc., the Amer. Phys. Soc., Sigma Xi and Tau Beta Pi.

John Riordan, B.S., Yale, 1923. Amer. Tel. and Tel., 1926–34; Bell Telephone Laboratories, 1934–. With the American Company and subsequently with the Laboratories, Mr. Riordan has been concerned chiefly with

transmission theory, the application of Boolean algebra to switching, number theory in cable splicing, and combinatorial and probability studies of traffic. He is a member of the Amer. Math. Soc., Math. Assoc. of America, Inst. of Math. Statistics, and Fellow of the Amer. Assoc. for the Advancement of Science.

A. A. ROETKIN, B.E.E., Ohio State University, 1927; M.Sc., 1929. Bell Telephone Laboratories, 1929–. With the Laboratories Mr. Roetkin has worked on overseas radio telephone receivers, ultra-high frequency, point-to-point radio telephone service, pulse multiplex microwave radio repeaters for the armed forces, and microwave radio-relay systems. He is a member of the Institute of Radio Engineers.

THOMAS SHAW, S.B., Massachusetts Institute of Technology, 1905. American Telephone and Telegraph Company, Engineering Department, 1905–19; Department of Development and Research, 1919–33. Bell Telephone Laboratories, 1933–48. Mr. Shaw's active telephone career was mainly concerned with loading problems in telephone circuits, including the transmission and economic features of the loading apparatus. The article which is concluded in this issue was started shortly before his retirement in 1948.

K. D. SMITH, B.A., Pomona College, 1928; M.A., Dartmouth, 1930. Bell Telephone Laboratories, 1930–. Consultant to National Defense Research Council, 1941–44. Awarded Joint Army-Navy Certificate of Appreciation for Scientific Achievement following World War II. With the Laboratories Mr. Smith has been concerned with the coaxial cable system, radar bombing equipment, broad band microwave radio system, and transistors. He is a Senior Member of the Institute of Radio Engineers.

R. L. WALLACE, JR., B.A. summa cum laude, physics and mathematics, University of Texas, 1936; M.A., physics, 1939; Special Research Associate, Harvard, 1941–45. Bell Telephone Laboratories, 1946–. Mr. Wallace's work with the Laboratories has been chiefly concerned with magnetic recording and transistors. He is a member of the Acoustical Society of America, Phi Beta Kappa, and Sigma Chi.

E. J. WALSH, Bell Telephone Laboratories, 1928–. Mr. Walsh's work with the Laboratories has been chiefly on vacuum tube design, magnetrons, proximity fuse tubes, reflex oscillators and close-spaced fine-wire grid tubes.