VACUUM TUBE DESIGN

Vacuum Tube Design

RCA Manufacturing Company, Inc.
Harrison, New Jersey

Privately issued by RCA Manufacturing Co., Inc.
for use of its employees and for restricted
distribution.

Photolithographed in U.S.A.

PREFACE

The material comprising the twenty-six lectures contained in this book formed the basis of a course on vacuum-tube design given by RCA engineers for company employees during the Winter of 1937 and the Spring of 1938. The lectures were intended to provide a review of the basic principles underlying the design and manufacture of vacuum tubes. They appear in the order in which they were presented and in a few instances include new material added at time of publication.

Each lecturer has treated his subject according to his own viewpoint. In general, the treatment is non-mathematical. Numerous formulas and charts of particular interest to the design engineer have been included.

Throughout the book, numerous references to sources of information have been given. References to published material list the publication; those to unpublished material are keyed. The keyed sources are available to authorized persons for reference purposes in our Library.

The Editors

Harrison, New Jersey
December, 1940

CONTENTS

Lecture 1	*FILAMENTS AND CATHODES - Part I*	E. A. Lederer	1
Lecture 2	*FILAMENTS AND CATHODES - Part II*	E. A. Lederer	11
Lecture 3	*HEATERS AND HEATER-CATHODE INSULATION*	G. R. Shaw and L. R. Shardlow	24
Lecture 4	*PHOTOELECTRIC AND SECONDARY EMISSION*	L. B. Headrick	34
Lecture 5	*LUMINESCENT MATERIALS*	H. W. Kaufmann	51
Lecture 6	*CONTACT POTENTIAL, PUMPS, AND GETTERS*	E. A. Lederer	58
Lecture 7	*METALLURGICAL PRINCIPLES*	S. Umbreit	73
Lecture 8	*METALS FOR VACUUM-TUBE CONSTRUCTION*	S. Umbreit	82
Lecture 9	*GLASS AND ITS PROPERTIES*	G. R. Shaw and C. A. Jacoby	91
Lecture 10	*CONSTRUCTION TRENDS IN RADIO TUBES*	N. R. Smith	103
Lectures 11, 12, 13, and 14	*SPACE-CURRENT FLOW IN VACUUM-TUBE STRUCTURES*	B. J. Thompson	115
Lecture 15	*ELECTRON OPTICS — Part I DETERMINATION OF ELECTRON TRAJECTORIES*	V. K. Zworykin and G. A. Morton	124
Lecture 16	*ELECTRON OPTICS — Part II ELECTRON-OPTICAL SYSTEMS WITH CYLINDRICALLY SYMMETRICAL FIELD-PRODUCING ELEMENTS*	V. K. Zworykin and G. A. Morton	136
Lecture 17	*ELECTRON OPTICS — Part III ABBERATIONS IN ELECTRON OPTICS*	G. A. Morton and E. G. Ramberg	153
Lecture 18	*RADIO RECEIVING TUBE COMPONENTS AND THEIR MANUFACTURE*	N. R. Smith	168
Lecture 19	*ANALYSIS OF RECTIFIER OPERATION*	O. H. Schade	174
Lecture 20	*THE DESIGN OF AUDIO AMPLIFIER AND POWER OUTPUT TUBES*	S. W. Dodge	194

CONTENTS (cont'd)

Lecture 21	*THE DESIGN OF RADIO-FREQUENCY AMPLIFIER TUBES*	T. J. Henry	203
Lecture 22	*THE DESIGN OF DETECTORS AND CONVERTERS*	T. J. Henry	209
Lecture 23	*THE DESIGN AND CONSTRUCTION OF TRANSMITTING TUBES*	E. E. Spitzer	218
Lecture 24	*THE DESIGN AND CONSTRUCTION OF CATHODE-RAY TUBES*	W. H. Painter	226
Lecture 25	*ELECTRON BEAMS AND THEIR APPLICATION IN RADIO TUBES*	H. M. Wagner	234
Lecture 26	*THE DESIGN AND PERFORMANCE OF RECTIFIERS*	A. P. Kauzmann	249

ERRATA SHEET
for RCA book entitled
VACUUM TUBE DESIGN

Page 25: Table I. Heading of center column should read

$$\frac{\text{Diameter (mils)}}{\sqrt{\text{mg/200 mm}}}$$

Page 114: In right column, 3rd line the year 1920 should be 1930.

Page 117: In right column, equation for d_{km} should read

$$d_{km} \approx 0.0156 \left(\frac{1}{1000\ I_b}\right)^{1/2} \left(\frac{T}{1000}\right)^{3/4}$$

Page 180: In Fig. 9, R_s should be changed to R in the circuit and in the accompanying legend. In the curves of i vs Time, the dashed curve starting at $i_{t(o)} = 0.089$ ampere should be identified as i_t. This designation may be added conveniently toward the right end of the dashed curve.

Page 183: In Fig. 12, lower part, $\frac{1}{2\omega C} >> 2\omega L$ should read $\frac{1}{2\omega C} << 2\omega L$.

Page 185: In Fig. 16, the curve e_c should be continued between point C and point O, as in Fig. 15a.

Page 190: In Fig. 21, lower part, the curve identifications in right margin should read from bottom up as follows: 100, 30, 10, 5, 2, 1, 0.2, and no designation to top (envelope) curve.

Page 243: In the boxed tabulation, the date for tube type 47 should be 1931 instead of 1929.

Page 249: In left column, line 21 the word empirical should be changed to graphical.

August 7, 1942

Lecture I

FILAMENTS AND CATHODES — Part I

E. A. Lederer

A. THERMIONIC EMISSION

1. Introduction

Electrons can be dislodged from matter by the action of incident light, by bombardment with electrons, by bombardment with positive ions, by metastable atoms, by the action of heat, or by applying intense fields. In the following we shall confine our interest only to the action of heat, commonly termed the field of thermionics.

The name implies that electron emission so obtained is a function of the temperature. But a cursory inspection of the phenomenon reveals that it is also a function of the material. The temperature function is expressed mathematically in the Richardson-Dushman equation which is known to fit experimental results within the error of measurement. The material function is contained in the equation in form of a factor A and a factor φ, the magnitude of which is at present determined only empirically.

The number of materials exhibiting thermionic emission of sufficient magnitude to be usable in various experimental and commercial devices is small. All materials, elements, and compounds holding a fair promise of becoming useful have been tested in the past three decades and the hope to add one more to the list is indeed very remote. Thermionic emission has been observed with

1) clean metals (represented by tungsten).

2) metals with minute surface contaminations (often called monomolecular films, represented by thoriated tungsten).

3) metal compounds like oxides, sometimes termed semi-conductors (represented by the oxide-coated cathode which at present is commercially the most important electron emitter).

All the materials have one property in common: they are conductors of electricity, at least at elevated temperature.

The property of metallic conduction is interpreted by the modern electron theory of metals. The carrier of the electric current is the electron, the smallest particle of electricity.

2. Some Properties of the Electron

The charge carried by an electron is

$$q = 1.591 \times 10^{-19} \text{ coulomb}$$

$$e = 4.77 \times 10^{-10} \text{ e.s.u.}$$

Thus, if a current of 1 ampere flows through a wire, and since 1 ampere = 1 coulomb per second, the number of electrons per second which pass a given cross-section of the wire is given by

$$n = \frac{1}{1.591 \times 10^{-19}} = 0.629 \times 10^{19}$$

An electron is the origin of a field of force and therefore has an equivalent mass which, expressed in grams for the electron at rest, is

$$m = 9.035 \times 10^{-28} \text{ grams}$$

The difference in mass between the moving electron and the electron at rest is very slight; only when its velocity approaches that of light does its equivalent mass increase notably.

If we assume that the electron has spherical shape, the size of its radius is given by

$$r = 1.85 \times 10^{-13} \text{ cm}$$

If an electron is moved through a potential difference of V' e.s.u., its kinetic energy changes according to the relation:

$$V'q = 1/2\ mv^2$$

Changing from e.s.u. to volts (1 e.s.u. = 300 volts), we obtain the following equation for the velocity v.

$$v = 5.94 \times 10^{7} \times \sqrt{\text{volts}} \text{ cm per sec.}$$

3. Modern Electron Theory of Metals

According to modern theory, the difference between the electrical properties of metallic conductors and insulators is that in the metal some of the electrons are free, while in the insulator they are bound to the atom. When a difference of potential is applied to the metallic conductor, the free electrons are set in motion and thus conduct current. In an insulator, however, the electrons are displaced somewhat by the electric forces but still remain part of the atomic system.

The classical theory of electric conduction was formulated by Lorenz and expanded by Thomson, Riecke, and Drude. This theory postulates that the metal consists of a rigid lattice structure built up from atoms, the interstices of which are occupied by free electrons. The electrons are thought of as being in to-and-fro motion, colliding with each other and with the atoms of the lattice. Because of this similarity, we speak of the "electron gas" in metals.

With a picture like this involving moving particles all of one kind, size, and charge, the most logical question is: what is the velocity of these particles? With such an enormous number of particles, it is utterly futile to assign

individuality to each and every particle and to chart their velocities. Instead we group them together and talk of a certain number of particles having velocities between certain limits. Exactly the same system is used in other statistics. For example, we speak of the life expectance of the population between the age limits of 20 and 21 years as being, say, 42 years. The goal of any statistical theory is to establish a distribution function by means of which we express the limit of velocity v + dv for a given number of particles n + dn.

Maxwell has computed the most probable velocity distribution of molecules in gases under ordinary conditions of pressure and temperature. And following his lead, the classical electron theory has assumed that the electrons in the metal have a Maxwellian velocity distribution. With this assumption which is the least restrictive of any we can make, a few properties of the metal could be explained, but others, such as the specific heat of the metal, showed differences between computed and observed values.

The Maxwellian distribution requires that the energy of the free electrons vanish at absolute zero, whereas from speculations in connection with the specific heat of metals, the free electrons should still possess considerable kinetic energy at 0°K. The logical inference is that if the free electron gas exists at all, its velocity distribution cannot be Maxwellian. The next step in modifying the classical theory was carried out by Sommerfeld in 1925. Digressing for a moment, I should probably mention that in 1916 a very convincing experiment proving the existence of the free electrons in metals was made by Tolman and Steward. In modifying the classical theory, Sommerfeld applied the Pauli-exclusion principle, the counterpart of which in statistical mechanics is the Fermi statistics, and could account for the first time for the correct value of the specific heat and other observed properties of the metals.

According to the Fermi statistics, the free electrons in the metal are endowed with kinetic energy even at absolute zero, and no two electrons in the metal can have the same kinetic energy. Instead, the electrons are spaced out according to a definite law. The result is that while the slowest free electron has no kinetic energy it is the only one of this sort. There is one, then, with a small value of energy, one with a larger value and so on until the fastest of the free electrons has quite a large value of kinetic energy, i.e., such as that which it would get by falling through a difference of potential of several volts. This distribution holds at absolute zero, and increasing the temperature produces very little change in the energy distribution. However, it is this small change which we use in thermionics.

An atom in a metal contains several loosely bound electrons surrounding a much more stable core consisting of the positive nucleus and a number of electrons. The loosely bound electrons are the ones which can be removed when the atom in question enters an ionic crystal. (Sodium chloride, for example, is an ionic crystal). However, in a metallic lattice, these electrons remain free because there is no electro-negative element (like chlorine in sodium chloride) to bind them. They are free to wander through the metallic crystal, and thereby carry electric current. The picture of a metal is then roughly a sea of electrons containing enough positively charged ions to make the whole thing electrically neutral. Different metals vary greatly in the fraction of the volume occupied by the positive ions. In an alkali metal, this volume may be only 10%, the remaining 90% of the volume being occupied by the free electrons. The volume occupied by positive ions in the ferrous-type metals may amount to 50% and over. The space occupied by the positive ions is used by Slater to compute some of the physical properties of the metal, such as ductility.

The electrons would diffuse right out of the metal by virtue of their velocity if it were not for some sort of restraining action. However, if an electron escapes, it leaves a positive charge of equal magnitude behind tending to prevent the escape of further electrons. For example, cathodes in rectifier tubes attain a positive charge because electrons are removed.

The attraction between the electron just outside the metal and the positive charge induced in the metal is called the image force and is numerically equal to

$$F = \frac{e^2}{(2x)^2}$$

where x is the distance of the electron from the surface of the metal. With only the image force acting, the egress of electrons would, therefore, ultimately be prevented by the building up of a surface charge but, as we find experimentally, this is not the case. Modern theory, therefore, postulates the existence of the restraining action at the surface of the metal in the form of a potential barrier, such that a definite amount of work, characteristic of the metal, must be done if the electron is moved from the interior to free external space.

If we denote by W_a the change in potential energy of a single electron after it has been removed from the metal, then W_a is also the work required to overcome the restraining action and corresponds to the energy required for removing an electron which was initially at rest in the metal. For removing an electron possessing kinetic energy W_i in the metal, a smaller amount of energy is required and is equal to $W_a - W_i$.

It seems justifiable to assume as Sommerfeld and Nordheim did, that all those electrons which have a velocity component ρ normal to the surface greater than a value ρ_o, given by the relation

$$\frac{m\ \rho_o^2}{2} = W_a$$

will escape and, therefore, will contribute to the thermionic current. Computing the number of electrons N(W)dW which have velocity components normal to the surface between W and dW, and which impinge from internally on the metal surface of unit area in unit time, Nordheim found the relation:

$$N(W) = \frac{4\pi m}{h^3} kT \log\left(1 - \varepsilon^{-\frac{W - W_i}{kt}}\right)$$

In this equation k is Boltzmann's constant ($k = 1.37 \times 10^{-16}$ ergs/degree) and h is Planck's constant ($h = 6.55 \times 10^{-27}$ erg sec.). The graphical representation of the equation for T = 0 is given in Fig. 1. As the temperature is increased, electrons appear with greater frequency in increasingly higher energy levels. For temperature T, the distribution is shown by the dashed line. We see qualitatively that the difference between W_a

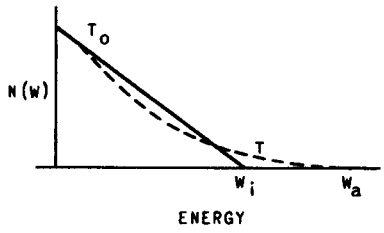


Fig. 1

and W_i must be a decisive factor in determining the number of electrons having sufficient kinetic energy to escape. This difference divided by the electronic charge, e, is called the work function,

$$\varphi = \frac{W_a - W_i}{e}$$

Thus, φxe is the energy required to remove one electron with the highest energy (W_i) at absolute zero from the metal. Or in other words, it is the latent heat of evaporation at absolute zero.

W_i can be calculated from the formula

$$W_i = \frac{h^3}{8m}\left(\frac{3n}{\pi}\right)^{\frac{2}{3}}$$

where n is the number of free electrons per unit volume in the metal.

Since metals expand when heated, it is readily seen that the electron concentration per unit volume changes with the temperature, that W_i changes but only slightly, and that consequently φ, the work function, is temperature dependent. However, no experiment has been devised as yet to support this deduction. Fortunately, W_a can be determined independently from thermionic and photo-electric measurements. It has been obtained from electron diffraction data making use of the de Broglie relation

$$\lambda = \frac{h}{m\nu}$$

connecting the wavelength λ, the electron velocity ν, and its mass m. In this equation h is Planck's constant. Davison and Germer carried out this measurement and found that W_a for Ni was 16.5 to 18 volts. Since the work function of Ni ~ 5 volts, W_i would come out to be 11 to 13 volts. If we assume two free electrons per atom, W_i as computed for Ni is 11.7 volts.

Let us consider what happens when two metals (a) and (b), as shown in Fig. 2, are in contact with each other at constant temperature. Experi-

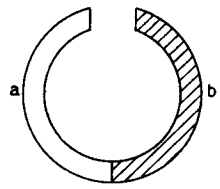


Fig. 2

ence shows that the system is in equilibrium with a characteristic potential difference V_{ab} across the external surface. At absolute zero, the maximum kinetic energy W_i is a material function and can be computed from

$$W_i = \frac{h^3}{8m}\left(\frac{3n}{\pi}\right)^{\frac{2}{3}}$$

We assume that $W_{i_b} > W_{i_a}$. This means that the maximum kinetic energy normal to the surface is greater in (b) than in (a). Therefore, an uncompensated stream of electrons would flow from (b) to (a) were it not for the barrier action at the interface. Thus, equilibrium is restored again and the loss or gain in kinetic energy at the interface is equal to $D_{ab} = W_{i_a} - W_{i_b}$.

Electrons passing from (a) to (b) are accelerated at the joint and those passing from (b) to (a) are retarded. An electron coming from (a) to (b) with maximum kinetic energy W_{i_a} is accelerated by D_{ab} and enters (b) with energy W_{i_b}. An electron with zero energy in (a) arrives with energy D_{ab} in (b). Electrons coming from (b) with maximum kinetic energy W_{i_b} are retarded by D_{ab} and arrive in (a) with energy $W_a = W_{i_b} - D_{ab}$; those with kinetic energy D_{ab} arrive in (a) with zero kinetic energy. No electrons with energy values smaller than D_{ab} can pass the interfacial region.

Equilibrium demands that the number of elec-

trons leaving with any kinetic energy be equal to the number entering with the same energy. This condition shows that the distribution function in (a) and (b) must be of the same form, one curve displaced with respect to the other by D_{ab}, as shown in Fig. 3.

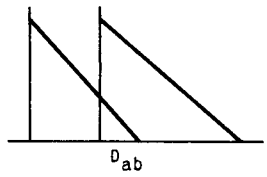


Fig. 3

4. The Emission Formula

In the light of modern electron theory, we shall now investigate qualitatively the process of electron emission from a clean metal. We treat the interior of the metal as a region of uniform potential with a barrier at the surface beyond which the potential changes to that of an electron outside of the metal. In Fig. 4, B indicates the surface of the metal to be thought of as a plane perpendicular to the plane of the paper.

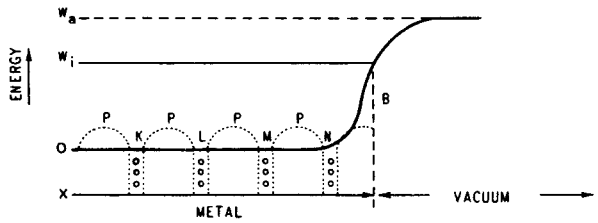


Fig. 4

Thus, the direction X is normal to the surface of the metal. The metal atoms are arranged in a three-dimensional lattice and the energy of an electron depends upon its location with respect to this lattice. We are interested in the forces acting upon an electron moving in the X-direction, and take the line integral of force, i.e., the potential of an electron as a function of its location along the X-direction. Actually the potential undergoes rapid changes near the atoms which are shown qualitatively by the dotted lines P.

The innermost levels K, L, M, etc., are all occupied. Electrons may be pictured way down in the potential valleys. The farther out the electrons in the structure of the single atom are, the more they are influenced by forces exerted by neighboring atoms. The valency electrons are thought of as being held so loosely and the binding forces acting upon them so weak and the influence of other atoms so great that they are assumed to move freely within the lattice. By doing so, they maintain the proper average charge to keep all regions electrically neutral.

The line marked O indicates the neutral zone through the potential hills and valleys. If we assume that there is no surface contamination on the metal, the difference between W_a and O can be assumed to be fixed. W_i indicates the highest energy levels at 0°K and is characteristic of the metal.

We must find now the number of electrons with X-energy between W_x and $W_x + dW_x$ impinging on the barrier surface per unit area in unit time. Classical as well as quantum mechanics show that all electrons with $W_x < W_a$ are reflected and, therefore, cannot contribute to the measured thermionic current. Electrons with higher energy than W_a have a definite probability of escape. When the energy W_x is very high, this probability $D(W_x)$ is very near unity, but when $W_x = W_a$, it must fall to zero.

The electron current from the clean metal surface can be expressed by

$$i = e \int_{W_x = W_a}^{W_x = \infty} D(W_x)\, N(W_x)\, dW_x \ .$$

$N(W_x)\, dW_x$, the number of electrons impinging in the X-direction on unit area in unit time, has been computed by L. Nordheim and is

$$N(W_x)\, dW_x = \frac{4\pi m k T}{h^3} \ln\left(1 + \varepsilon^{-\frac{W_a - W_i}{kT}}\right)$$

On substituting and integrating and assuming that the probability of escape $D(W_x) = 1$, we obtain

$$i = \frac{4\pi m k^2 e}{h^3} T^2 \varepsilon^{-\frac{W_a - W_i}{kT}}$$

Denoting $\frac{4\pi m k^2 e}{h^3}$ by A_o, we have

$$i = A_o T^2 \varepsilon^{-\frac{W_a - W_i}{kT}}$$

The emission formula arrived at in this manner is remarkable not only in that it agrees in form with the earlier formula derived thermodynamically by Richardson, but also in that the material constant A_o is identical with that obtained by Dushman and Laue if corrected for electron spin (see paragraph on the Constant A).

In older publications, the formula is often given as follows:

$$i_o = A T^2 \varepsilon^{-\frac{b_o}{T}}$$

or

$$i_o = A' T^{\frac{1}{2}} \varepsilon^{-\frac{b_o}{T}}$$

where b_o is a constant defined below. We know now that the T^2 formula is better founded theoretically than the Richardson $T^{1/2}$ formula derived by him from the classical theory. The accuracy with which temperatures and emissions are measurable is not great enough to establish anything beyond the fact that the temperature variation of emission is dominated by a factor of the type $\varepsilon^{-b_o/T}$.

The relation between B_o and φ is given by the equation

$$b_o = \frac{\varphi e}{k}$$

Substituting the proper value for tungsten, we get for φ_W

$$\varphi_W = \frac{b_o k}{q} = \frac{52600 \times 1.371 \times 10^{-23}}{1.591 \times 10^{-19}}$$

$$= 4.53 \text{ volts}$$

(The electronic charge has been expressed in coulombs and the constant k in joules/degree). The velocity necessary for the egress of all the electrons from tungsten is, therefore:

$$V_{max} = 0.594 \times 10^8 \sqrt{4.53}$$

$$= 1.3 \times 10^8 \text{ cm/sec. (approx.)}$$

The temperature necessary for the egress of all electrons from tungsten is

$$T_{max} = 11604 \times \varphi_W = 52600 \text{ degrees.}$$

Obvious limitations do not permit us to heat tungsten cathodes to much more than 2800°K at which temperature it has been computed that only about one electron out of every 100000 electrons has sufficient velocity to escape. We see, therefore, that the emission currents drawn within realizable temperature ranges practically do not influence the electron concentration in the metal.

5. The Constant A

That the constant A must approach universal constancy was shown first by Richardson. Laue and Dushman calculated A, making certain assumptions, and found the universal value:

$$A = 60.2 \text{ amp./cm}^2\text{/degree}^2$$

Recently, in view of the electron spin, discovered by A. Compton and by means of which Pauli could account for the magnetic properties of the alkali metals, it has been found that the value of A should be corrected by a factor of 2, so that

$$A_o = 120 \text{ amp./cm}^2\text{/degree}^2$$

That this value is in poor agreement with the observed value as extrapolated from Richardson's equation for various emitters is probably due to our inability to produce chemically clean surfaces. Another explanation was offered by Dr. W. B. Nottingham according to which the work function depends upon the crystal orientation. Since we use polycrystalline wires, we observe average and not optimum work function. There are only very few metals which can be degassed sufficiently in vacuum so as to clean their surface to a satisfactory degree. These metals are tungsten, molybdenum, and tantalum. An inspection of the following table will be convincing that the agreement between theory and actual observed value is best in the three cases just mentioned.

Table I

EMITTER	A amp/cm²/degree²	WORK FUNCTION φ Volts
W	60 - 100	4.54
Re	200	5.1
Pt	170000	6.27
Mo	55	4.15
Ta	60	4.1
Ba	60	2.11
W-O-Ba	0.18	1.34
W-Th	3.0	2.63
W-Ba	1.5	1.56
W-O	5×10^{11}	9.2

6. The Transmission Coefficient

Electrons impinging upon a potential barrier, such as we assume to exist on the surface of a metal, may be in part reflected, and in part they may "pass over" the barrier. Thus, the relation between the reflection r and transmission coefficient D is given by the relation

$$D = 1 - r$$

If the kinetic energy of the impinging electron is smaller than the energy required to pass over the potential barrier, the electron is reflected according to classical mechanics. In the light of wave mechanics, according to which electrons behave also as waves, there is a certain probability that the electron wave train can pass through the barrier. This probability depends upon the potential distribution on both sides of the barrier and upon its thickness. The phenomenon is the analogue of the optical phenomenon in which a light beam in an optically dense medium is not totally reflected when interrupted

by a layer of optically less dense material provided the latter's thickness does not exceed a few wavelengths. For practical purposes and always if the kinetic energy of the impinging electron is larger than the height of the barrier, the transmission coefficient is very nearly unity and, therefore, has been neglected in our emission formula.

In making certain simplifying assumptions, Dushman has calculated the thickness of the potential barrier on thoriated tungsten and found it not to exceed 3 angstroms. Since the diameter of a thorium atom is 5.1 angstroms, the correlation appears to be satisfactory.

We must remember, however, and this is well supported by observation, that a minute surface contamination on the emitter may change both the work function and the constant A profoundly. We seek the explanation in the shape and height of the potential barrier and may expect more practical information from a study of surface contamination by electron diffraction.

7. Testing of the Emission Formula

The material to be tested, preferably in the form of a thin wire, is mounted inside a cylindrical anode consisting of three parts as shown in Fig. 5. The two outer parts of the anode, called guard-rings, serve to confine the emission measurement to the central portion of the filament which is at uniform temperature.

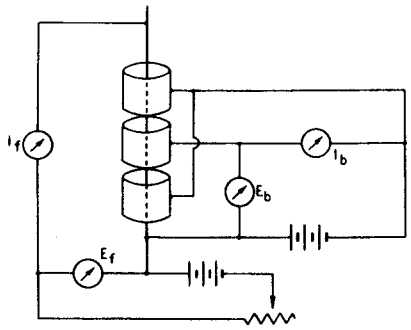


Fig. 5

In order to test our emission equation, it is necessary that the potential difference between filament and anode be zero, and that the emission current be measured as a function of the filament temperature. Because of the voltage drop along the filament and the contact potential difference between filament and anode, the measurement is carried out with an anode potential ranging between 30 and 600 volts and is then extrapolated to zero.

If we denote the quantities which we observe with the subscript "ob", the emission formula as measured is

$$i_{ob} = a\, A_{ob}\, T^2\, \varepsilon^{-\frac{e\varphi_v}{300\ kT}}$$

$$\text{or}\quad \log \frac{i_{ob}}{T^2} = \log a + \log A_{ob} - \frac{e\varphi_v}{690\ kT}$$

Plotting $\log(i_{ob}/T^2)$ as ordinates and $1/T$ as abscissa, we obtain a straight line having the slope

$$-\frac{e\varphi_v}{690\ kT} = b_v$$

Inasmuch as we used anode voltage and because, as Schottky has pointed out, the applied voltage aids the escape of electrons and decreases the φ, we have to correct for this. Schottky's relation is:

$$\log i = \log i_o + \frac{1.906\ E^{\frac{1}{2}}}{T}$$

where E is expressed in volts/cm. For concentric cylinders, we have

$$E = \frac{V}{r \ln \frac{R}{r}}$$

where r = radius of filament and R = radius of anode, both in centimeters.

B. EMISSION FROM CLEAN METALS

Experience has shown that electron emission of sufficient magnitude for practical purposes can be had from clean metals only at temperatures in excess of about 2000°K. This limitation in temperature range limits the number of clean metal emitters to very few among which tungsten, tantalum, and molybdenum are the only metals of practical importance.

Metals cannot be regarded as clean unless their surface and their interior as well, have been substantially freed from contamination. All three of the previously mentioned metals have been used as thermionic emitters, but only tungsten and tantalum are used commercially. The difficulty encountered with molybdenum is its comparatively high vapor pressure and consequently cathodes made of it have short life. Tantalum, like molybdenum, has a slightly lower work function than tungsten. Tantalum is mechanically not strong enough and "sags" under the combined influences of temperature and the electrostatic forces between cathode and anode.

The thermionic properties of tungsten are known with greater accuracy than those of any other metal. Not only has tungsten the highest melting point (3655°K) of all metals but its chemical compounds formed with various gases, particularly oxygen, are more easily removed by heat treating than those of tantalum for example. Furthermore, the temperature scale of tungsten has been studied with great accuracy.

In high-power transmitting tubes, tungsten is the only metal rugged enough to withstand all the mechanical and chemical requirements. As seen from Table I, the work function of tungsten increases more than twofold after exposure to oxygen. Although the oxygen contamination can be removed by heating the tungsten to a temperature in excess of 1800°K, it is obvious in order to maintain a steady emission level with respect to small diameter. Wires of larger size can be operated at a higher temperature for the percentage loss due to evaporation becomes smaller with increasing bulk of the emitters.

Practical design data for pure tungsten filament, as employed in our transmitting-tube section, will be given by Mr. Spitzer. The properties of pure tungsten wire which are of importance to the radio-tube designer are given in Table II.

Table II *

Temperature °K	Resistivity microhms/cm	Total Radiation Intensity watts/cm^2	Electron Emission amp/cm^2	Rate of Vaporization grams/cm^2/sec
300	5.64	.0015	-	-
600	13.54	.048	-	-
900	22.58	.379	-	-
1200	37.02	1.691	-	-
1400	38.52	3.82	5.75×10^{-9}	-
1600	45.22	7.77	8.05×10^{-7}	3.7×10^{-20}
1800	52.08	14.22	3.92×10^{-5}	6.22×10^{-17}
2000	59.10	23.72	8.92×10^{-4}	2.32×10^{-14}
2200	66.25	37.18	1.14×10^{-2}	2.90×10^{-12}
2400	73.55	55.8	1.02×10^{-1}	1.58×10^{-10}
2600	81.0	80.8	6.48×10^{-1}	4.64×10^{-9}
2800	88.5	112.9	3.21	8.28×10^{-8}
3000	96.2	153.9	-	9.92×10^{-7}

* Data given in this table apply to well-aged tungsten.

time, that the vacuum in the device must be of a high order. Expressed in common units, the vacuum should be 10^{-6} mm or less. Attainment of such a vacuum level is carried out either by the use of a getter, or if this is not permissible, by internal bombardment of the transmitting tube during exhaust. This latter procedure sputters electrode material on the wall of the envelope while the gases as evolved are carried off by the pump. The sputtered metal on the inner surface of the bulb adsorbs gases freed after sealing off and thus aids in maintaining the vacuum. Furthermore, the tungsten acts in such a manner as to be self-gettering in that, during operation, a small amount of evaporation and sputtering is unavoidable. The adsorption of gases by the sputtered metal or the getter is considerably enhanced by ionization.

Thus, in view of the known properties of tungsten, we can derive logically an aging schedule which consists of two steps:

1) Removal of gas residue in the device by the combined action of sputtering and ionization during which the tungsten cathode is maintained at slightly over the operating temperature.

2) Cleaning the surface of the tungsten by flashing for a short time at a temperature near 3000°K.

Since evaporation at high temperature seriously limits the life of a tungsten cathode, it is not practical to exceed 2400°K for wires of

C. ELECTRON EMISSION FROM CONTAMINATED METALS

1. Synopsis

When the surface of a metal is contaminated with a film of foreign material, the surface polarization is profoundly changed and concurrently the work function and the transmission coefficient are altered. While such contamination occurs abundantly in nature, only a few cases are of practical importance and hence have been subjected to detailed study. The most important of these cases are:

Tungsten - Thorium
Tungsten - Caesium
Tungsten - Oxygen - Caesium
Tungsten - Oxygen - Barium

The first two cases may be regarded as applying to one class of emitter and the last two cases to another class. The latter class leads to the so-called oxide-coated cathode which will be discussed separately because of its technical importance. In all four cases, tungsten is used as the base material which is most natural in view of what has been said previously about its properties, particularly since it is so readily cleaned of undesired contamination by flashing.

2. Thoriated Tungsten

a. The Main Phenomenon — In order to prevent offsetting, tungsten filaments were "doped" with 0.75 to 2% thoria even in the early stages of tungsten incandescent-lamp manufacture. In 1913 Langmuir and Rogers discovered that the thermionic emission of such "thoria-doped filaments" was, after certain treatment, several thousand times larger than the emission of a pure tungsten filament under the same conditions. Various treatments were tried but the one finally adopted to obtain maximum thorium emission is as follows:

The filament is flashed for a minute or two at a temperature higher than 2700°K to reduce some of the thoria in the wire to thorium metal. At this high temperature, any impurity on the surface of the wire is vaporized. Any thorium metal diffusing to the surface is immediately evaporated so that the activity of the filament is substantially that of pure tungsten.

If now the temperature is reduced to a value between 2000° and 2200°K, the rate of diffusion is quite high but the rate of evaporation is decreased to such an extent that thorium atoms can accumulate on the surface as an adsorbed layer. In order that the rate at which the surface becomes covered with thorium can be observed, the emission is tested at frequent intervals at a comparatively low temperature (1600°K). At this temperature the rate of diffusion becomes negligibly small. Fig. 6 shows qualitatively what takes place at various temperatures and it also indicates the ranges for reduction, diffusion, operation, and deactivation.

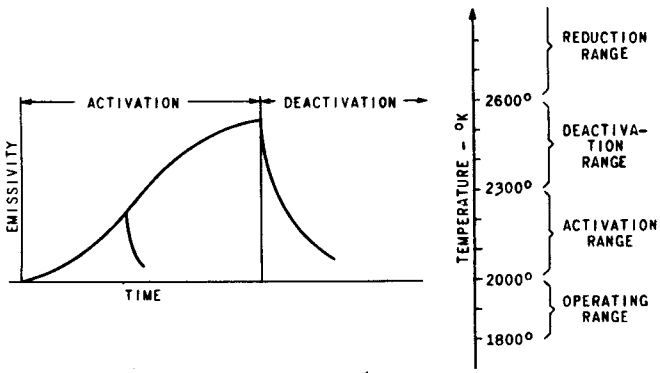


Fig. 6

b. Reduction of Thoria by Tungsten — The heat of formation of thorium oxide is greater than that of tungsten oxide. Nevertheless, thoria is reduced by tungsten in vacuum owing to the workings of the mass action law. The reaction occurs at an appreciable rate at 2600°K and over. The tungsten oxide is lost in the vacuum and deposits on the cooler parts of the device.

That the reduction actually occurs is supported by tests and analyses of Smithels, Geiss, and VanLiempt who examined the residue of flashed and untreated filament after dissolving the tungsten. The thorium metal formed is only slightly soluble in tungsten and the solubility increases with the temperature. The presence of thorium metal in the tungsten was confirmed by measurements of the temperature coefficient of wires by Fonda, Young, and Walker. That the formation of thorium metal is due purely to dissociation appears improbable because extremely high temperature would be required for such a process.

c. Nature of the Surface Film — From work by Langmuir, Brattain, Becker, and Nottingham, and from the suddenness at which activation and deactivation occur, it appears certain that the thoriated filament owes its activity to a monomolecular surface film. Thus, at the operating temperature, we have an equilibrium between the rate of arrival of thorium at the surface and the loss due to evaporation. Nevertheless, the rate of arrival of thorium is believed, and supported by evidence, to be greater than the evaporation of a monatomic film. It is merely the surplus of thorium which might form an incipient second layer which is lost at the operating temperature since the surface forces between tungsten and thorium are evidently much greater than the forces between the second thorium layer and the first. Thus, we assume that any surplus thorium over the amount necessary to form, maintain, and repair the first layer, is volatilized. That there is such a thing as the incipient second layer is evident from curves by Nottingham, Brattain, and Becker, who plotted emissivity versus time and found the former to attain a maximum and then to decrease to a slightly lower constant level. The formation of this second layer occurred more readily at lower temperatures as might be expected. Fig. 7 shows qualitatively the phenomenon described.

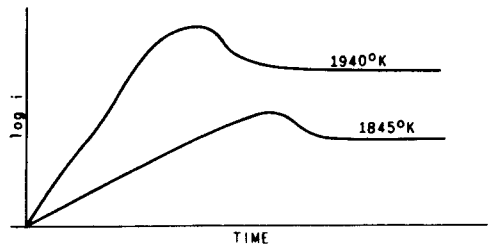


Fig. 7

d. Concentration of Thorium at the Surface; Surface Migration — Deposition of a contamination upon a base material has been observed to occur often in conformity with the underlying lattice. So, for example, if a polycrystalline tungsten wire is heated in an atmosphere of tungsten hexachloride, the tungsten so deposited upon the wire is polycrystalline. If a single-crystal wire is used, the deposit is monocrystalline. From this and other observations, Langmuir assumed that the concentration of thorium atoms must depend upon the geometry of the tungsten lattice. Tungsten crystallizes in body-centered-cubes, the side of a unit cube being 3.15 angstroms. Thorium, on the other hand, crystallizes

in face-centered cubes, the side of a unit cube being 5.04 angstroms. Evidently the size of thorium atoms is too large for a one-to-one relationship and thus Langmuir has calculated that the greatest number of atoms which can be packed into a monatomic layer is 64% of the number of tungsten atoms exposed on the surface. Unfortunately, we have not as yet a reliable method to check such speculation.

Notwithstanding the great attractive forces existing between tungsten and the monatomic layer of thorium, there are numerous manifestations that such adsorbed atoms may migrate along the surface. Also it has been held that the thorium reaches the surface of the wire by way of the grain boundaries. Both statements were recently verified by Nottingham and Johnson who experimented with a thoriated filament inside a cylindrical fluorescent screen. With this arrangement the magnification of the electron pattern in the radial direction was approximately equal to the ratio of the screen-to-filament radius, whereas the magnification along the filament axis was zero. From the experiments the following conclusions were obtained:

1) The thorium emerges at the grain boundaries and spreads over the entire surface by migration.

2) The points of emergence are the same even after several activations and deactivations.

3) The surface coverage depends on the orientation of the tungsten crystal, certain crystal faces showing a preferential adsorbtivity as compared with others.

It has long been known that polycrystalline wires give better all-round performance as emitters than single-crystalline wires. This has been recently confirmed by Clausing who investigated the emissivity of a single-crystal and a polycrystalline wire before and after depositing on each a shell of pure tungsten, grown thereon from the vapor phase. Clausing observed that the single-crystal wire which was encased in a single-crystal shell substantially exhibited tungsten emission although it was fully activated before the shell was deposited. The polycrystalline wire, however, activated with practically equal ease before and after the shell was grown on.

e. Carbonizing of Thoriated Filaments — If a tungsten filament is heated to about 1600°K in a hydrocarbon vapor, carbon diffuses into the tungsten. The rate of dissociation of the hydrocarbon increases, everything else being equal, with the gas pressure, whereas the rate of diffusion of the carbon into the interior of the filament decreases with decreasing temperature. Therefore, by selecting the gas pressure and filament temperature, it is possible to build up a shell of carbonized tungsten at the surface of the filament. There are two tungsten carbides, W_2C and WC. The former contains 3.16% carbon by weight, whereas the latter contains 6.12% carbon. Since the conductivity of tungsten carbide is vastly different from that of pure tungsten, the process of carbonization as it progresses can be followed by observing the resistance of the wire. The conductivity of tungsten (cold) as a function of its carbon content is shown in Fig. 8.

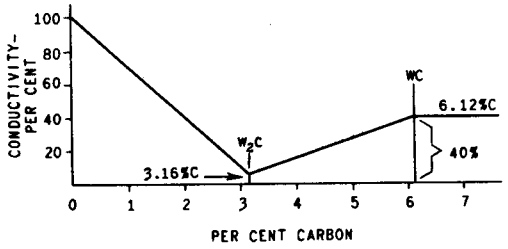


Fig. 8

Carbonized thoriated-tungsten filaments exhibit remarkable properties in that the rate of evaporation of thorium from such filaments is only about 1/6 at 2200°K of that from a tungsten surface. Hence, carbonized filaments can be operated at a higher temperature and, since the rate of thorium diffusion can thus be increased, the filament is more rugged toward adverse effects. Because tungsten carbide is very brittle, and in order to maintain as much of the mechanical strength as possible, it is customary to convert only a certain fraction of the cross-section of the filament to carbide and leave the core of the filament uncarbonized. Experience has shown that all-round satisfactory performance can be obtained if carbonization is carried out up to 80% of the original conductivity.

f. Aging Schedule and Poisoning Effect — Having surveyed qualitatively the salient properties of the thoriated-tungsten filament and bearing in mind that thorium is very active chemically, we can now design a general aging schedule for this type of cathode.

1) Remove residual gas left in the device after tip-off by using an active getter and aiding the clean-up by ionization if necessary with filament at temperature slightly less than 2600°K.

2) After removal of gas, flash filament at reduction temperature for not over two minutes.

3) Activate filament at temperature within the thorium diffusion range. If gas has not been cleaned up, the filament will not activate because of oxidation of the thorium on the wire surface.

Since the ratio of volume to surface for cylindrical filaments changes rapidly with increasing filament diameter, it is obvious that thick filaments contain more thorium reserve per unit area of surface to be covered with a monatomic layer. Hence, if a filament has been deactivated for some reason, it is usually possible to reactivate it again by applying the last two steps of our aging schedule. Filaments of small diameter contain little thorium reserve and once deactivated it is usually difficult to restore the initial emission by aging. Therefore, small-diameter filaments are particularly sensitive to

oxygen poisoning. One source of oxygen poisoning is due to the dissociation of oxide films under electron bombardment. Another source is the so-called "water cycle" known in the manufacture of incandescent lamps. The latter source is due to the presence in the tube of traces of hydrogen which, when ionized, reduces metal oxides as found on spot welds and forms water vapor. The water vapor attacks the thorium film on the filament or the filament proper and forms hydrogen with the result that the cycle begins anew.

Oxygen poisoning caused by dissociation of oxide films under electronic bombardment was first observed on the type 99. The life of this tube using a filament with a diameter of about 0.0007 inches had been notoriously bad. Changes in exhaust, aging, getter, and filament led to only temporary relief. Therefore, a detailed study was undertaken. In order that as many variables as possible might be removed, the emission versus time characteristic was observed for diodes. First, it was found that every lot of filament could be properly activated initially if the vacuum conditions were satisfactory. Second, it was found that all tubes had satisfactory life performance if the applied plate voltage did not exceed 13 volts. Increasing the plate voltage beyond this value resulted in a more or less rapid deactivation lasting from a few seconds to a few hours. The plate material in use was nickel. It was obtained from various sources of supply. Different lots affected the deactivation time. The next step was a test of other plate materials, and so iron, zinc, silver, and platinum were tried. Of these, the last three had no effect upon the filament, the emissivity remaining constant for several hundred hours with as much as 50 volts applied to the plate. At the same time it was observed on life test that tubes which contained a small amount of red phosphorus, in addition to the magnesium getter, exhibited a much better emission maintenance than regular production tubes.

From this study, it was evident that the deactivation observed was due to a minute oxide film on the nickel plates. This film formed as soon as the plates were exposed to the air after hydrogen firing. With an understanding of the phenomenon, it was a comparatively easy matter to devise a preventative. Since red phosphorus was too erratic with respect to results and since magnesium-nickel alloys containing several per cent of magnesium metal could not be had, a nickel-zinc alloy was finally adopted. Satisfactory life-test results were then obtained. The action of the zinc during high-frequency heating consisted in cleaning the surface of the plate material by diffusion, evaporation, and reduction.

REFERENCES

See end of Lecture 2

Lecture 2

FILAMENTS AND CATHODES — Part II

E. A. Lederer

OXIDE-COATED CATHODES

1. Introduction

In the first lecture we discussed the escape of electrons from clean and contaminated metals and we noticed the enormous difference in work function caused by monatomic layers of two different materials on a clean metal surface. Comparing for example thoriated and oxidized tungsten whose fork functions are 2.6 and 9.2 volts, respectively, we find that the emission of the former is several millionfold greater than the latter at the same operating temperature. The oxide-coated cathode is but a special case of a contaminated surface and because of its commercial importance deserves considerable discussion.

The first question suggesting itself is: What is the definition of an oxide-coated cathode? Because of the complexity of the article, both from the physical and chemical viewpoint, no such definition can be given at present. It seems logical to review at first methods of preparation and properties before attempting to define such a cathode. The definition will at best describe only certain salient points because our knowledge of this subject is far from complete.

The oxide-coated cathode was discovered in 1903 by A. Wehnelt, who was studying gas discharges at that time and observed a considerable decrease of the so-called cathode fall, when the cathode was accidentally coated with certain impurities. A careful study of the impurities revealed the presence of calcium compounds and after much experimentation Wehnelt found that the oxides of calcium, strontium, and barium were responsible for the observed effects. All three elements constitute the group of the so-called alkaline-earth metals. When speaking of an oxide-coated cathode, we mean a metal surface which can be heated and which is coated with any one of the oxides of calcium, strontium or barium; or with a mixture of two or three of these oxides. More recently the meaning of an oxide-coated cathode has been confined to one coated with barium and strontium oxides only since both exhibit electron emission in a greater degree than calcium oxide.

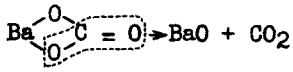
For ten years after its inception, the oxide-coated cathode remained in the laboratory of the physicist as the object of some study and of much controversy. Shortly before the war, the cathode was developed commercially to be used in telephone repeater tubes by the Bell Telephone Laboratories. The coating developed for this purpose was sintered in air upon a platinum alloy serving as the base material. Part of the base was dispersed in and combined with the coating with the result that the coating had a gray to black appearance. Hence, this coating was known as the combined coating and because of its dark color, required an input of 6 to 8 watts/cm^2 to maintain it at the proper operating temperature.

It was not until about 1925 that the technique of controlling a white uncombined coating was mastered to such an extent as to make use of it commercially. In the past 10 or 15 years, the technique has been greatly improved and with it the efficiency of the uncombined coating has been increased. Hence, the input required is smaller than 4 watts/cm^2 with a lower limit of approximately 1 watt/cm^2. The lower the cathode temperature, the more sensitive is the coating toward the effects of gaseous residues in the tube and, therefore, commercial cathodes are designed to operate at not less than 2.5 watts/cm^2. This high thermionic efficiency is the main reason why the uncombined coating is finding ever-increasing use in thermionic devices. Another very important advantage of the uncombined coating is its ease of application by dipping or spraying on a metal surface of any desired shape. Since the thermionic activity is confined strictly to the coated section and does not extend onto uncoated sections of the metallic base, it is possible to design tubes with well-defined coated-cathode areas. Such well-defined areas make it relatively easy to obtain definite values of cut-off, plate current, and transconductance.

In the following, the term "oxide-coated cathode" means an uncombined coating consisting of barium and strontium oxide on a metallic base.

2. Preparation of Oxide-Coated Cathodes

Since barium or strontium oxides are unstable in air and combine rapidly with moisture and carbon dioxide, it is commercially impossible to use these oxides as such for the preparation of the coating. Fortunately, there are a number of chemical compounds of the alkaline-earth metals which are stable in air and which when heated in vacuum decompose to the oxides. The best compounds are the carbonates because they have the required stability and can be obtained very pure. They are inexpensive and besides the oxides, the only decomposition product is carbon dioxide which, being a gas, can be removed by the vacuum pump. For example, barium carbonate can be written by its constitution formula



and decomposes as indicated by the dotted line and arrow into barium oxide and carbon dioxide at elevated temperature. The decomposition temperature is different for all three of the alkaline-earth metals. The relation between dissociation pressure in millimeters of mercury and temperature for magnesium, calcium, and barium carbon-

ates is shown in the following table.

Carbonate	800°K	1000°K	1300°K
$MgCO_3$	100	>750	-
$CaCO_3$	0.1	22	3000
$BaCO_3$	0	0	3.0

Taking the value 3.0 mm for barium carbonate at 1300°K means that if the pressure of the CO_2 atmosphere surrounding the $BaCO_3$ is kept smaller than 3 mm, eventually all of the barium carbonate is converted to the oxide at 1300°K. Our vacuum pumps are capable of maintaining a much lower pressure than 3 mm and thus the decomposition is made to proceed rapidly. In carrying out the experiment with pure barium carbonate, one finds, however, that the carbonate melts at about 1300°K forming a solid solution of carbonate and oxide which causes sputtering and gas inclusions in the melt. It has been found, however, that mixtures of about equal parts by weight of barium and strontium carbonates do not melt but decompose easily to the oxides. This is one reason why commercial oxide-coated cathodes are always coated with barium and strontium oxides.

The oxides of the alkaline-earth metals belong to the cubic system, rock-salt type. The dimensions of the unit cubes are:

CaO $A_w = 4.8\ A^o$
SrO $A_w = 5.15\ A^o$
BaO $A_w = 5.53\ A^o$

The difference between a barium-oxide crystal, for example, and a sodium-chloride crystal, is that in the former the lattice points are occupied by bivalent ions. The unit cube of barium oxide is sketched in Fig.9.

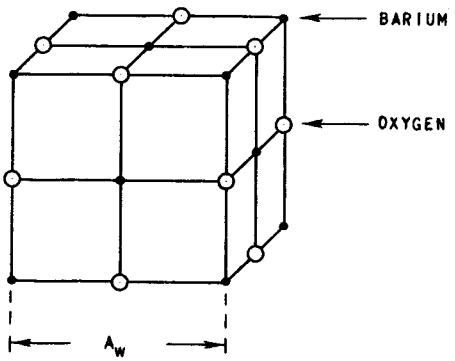


Fig. 9

Solid solutions of barium and strontium oxides of molecular proportions, such as we encounter in the cathode coatings, would have unit cubes in which barium and strontium alternate in the lattice points. Form X-ray analysis, we know that in this case, A_w of the mixed crystal would be the mean of the barium A_w and the strontium A_w. In fact from X-ray diffraction patterns, Benjamin and Rooksby determined the barium ratio of the solid solution which checked with that determined by chemical analysis to better than 10 per cent.

It has been said already that the coating contains about equal parts by weight of barium and strontium oxide. This mixture can be obtained by mixing barium and strontium carbonates in the desired proportions. Such mixtures are called single-mixed carbonates. The single carbonates, for example, pure barium carbonate, are made in our plant by precipitation. Barium nitrate, $Ba(NO_3)_2$, is first dissolved in distilled water. Then it is mixed with a solution of sodium carbonate, Na_2CO_3. The insoluble barium carbonate precipitates. In chemical symbols, we have:

$$Ba(NO_3)_2 + Na_2CO_3 = BaCO_3 + 2NaNO_3$$

It remains only to filter the resulting liquid to retain the precipitate, to wash it repeatedly in hot distilled water, and to dry it. Strontium or calcium carbonate can be obtained in exactly the same manner.

Mixing of the single carbonates is a purely physical process, depending on the particle size, the method of mixing, the time, and other factors. No matter how well the mixing process has been carried out, the final product consists of particles of two different components. At best, the final particles are aggregates of several thousand molecules and physical means do not permit better mixing.

When we dissolve in water the nitrates of barium and strontium simultaneously, we obtain mixtures, down to molecular dimensions, in an easy and inexpensive manner. Precipitating this mixture with sodium carbonate, we obtain barium and strontium carbonate mixed more intimately than has been possible with physical means. The carbonates prepared in this manner are called coprecipitated double carbonates. Sometimes it is of advantage to add to the barium and strontium carbonates a certain amount of calcium carbonate and if prepared by coprecipitation, we then speak of coprecipitated triple carbonates.

Examining the carbonates as precipitated under the microscope, we observe either distinct crystals or aggregates thereof. The temperature and the method of precipitation controls the crystalline structure of the precipitate, and this in turn affects, within certain limits, the results obtained with the coating.

Precipitating pure barium nitrate with sodium carbonate yields very fine submicroscopic crystals which agglomerate to round clusters, commonly called spherulites. Increasing the temperature during the precipitation up to the boiling point increases the size of the crystals but has little effect upon the size of the aggregates. Precipitating barium and strontium nitrates in approximately molecular proportions with sodium nitrate yields large needles (5 to 15 μ long) from which a fluffy coating can be prepared. If ammonium carbonate instead of sodium carbonate is used, a dense spherulitic precipitate is obtained.

In general, dense carbonates of spherulitic structure are useful and desirable in close-spaced rectifiers. Fluffy carbonates are used for coating of cathodes for low-power receiving tubes. Single-mixed carbonates are used for coating filaments of all sizes and also make a good coating for certain close-spaced rectifiers. Experience has shown that fluffy coatings are generally easier to activate than dense coatings. The latter type of coating, especially carbonates with spherulitic structure, sputters less readily when used in rectifiers.

The carbonates can be coated onto a metallic base by many methods. Two methods, because of their simplicity and ease of control are widely used. They are: spraying, and coating by the drag process. Spraying is used for coating indirectly heated cathodes and in rare cases for coating filaments. The drag process is suitable for filaments only.

The first method involving the use of a spray mixture consists in applying the carbonates suspended in a suitable binder by means of a spray gun. The spray mixture is prepared by ball-milling the carbonates together with an organic vehicle for several hours (18 to 72 hours). In order to improve the adherence of the carbonates, it is customary to employ nitrocellulose as a binder; hence, the organic vehicle used should be a solvent of nitrocellulose. By the proper selection of the vehicle or the use of mixtures of solvents, the drying properties of the coating can be varied within wide limits. Slow-drying sprays give dense, smooth coatings. Solvents having a high vapor pressure at room temperature are used to obtain fluffy coatings. The amount of nitrocellulose seldom exceeds 2 per cent by weight of the solid carbonates. Nitrocellulose is to date the best binder we have because it decomposes at a low temperature (approximately 200°C) without leaving any perceptible residue. This is because nitrocellulose does not depend solely on the supply of oxygen from the atmosphere for combustion, but is partially self-combusting.

The second method, or drag process, is carried out by applying the carbonate coating in successive thin layers onto the filament and baking it between applications in carbon dioxide to about 700°C. In this case the coating mixture is a suspension of the carbonates in water, and in order to improve the adherence, barium and/or strontium nitrate are added (up to 10 per cent of the carbonates) to the solution. In the baking operation in CO_2, the nitrates are converted to carbonates and in this manner the coating particles are cemented together. The coating process consists of dragging the filament through the coating mixture which by capillary forces collects in a groove of a slowly rotating small wheel whose lower section is immersed in the mixture. In this manner the mixture is agitated and a definite amount of coating, controlled by viscosity and surface tension, is transferred to the filament.

There is another method of coating which should be mentioned. It is known as coating by cataphoresis, and is applicable to filaments and indirectly heated cathodes. This method yields coatings of high density and is less flexible than the two methods previously described. Furthermore, with cataphoretically coated cathodes, it appears more difficult to meet the present commercial emission limits.

In describing the amount of coating on cathodes, it is convenient to specify the weight of the coating in milligrams per cm^2 of coated surface. Experience has shown that with a coating weight of from 3 to 6 mg/cm^2 satisfactory life can be obtained. Coatings which are heavier than the upper limit given above are difficult to degas.

It is of great importance that the coating adhere well to the base material. If the coating mixture does not wet the core during decomposition, blisters and air pockets are produced and cause poor electron and heat transfer in the overlying coating.

Theoretically, it makes no difference how the carbonate coating is applied onto the base material. Practically, however, in the manufacture of radio receiving tubes with a multitude of variables, the task of reducing initial shrinkage and obtaining good quality is often facilitated by selecting the proper coating and binder for the particular purpose. This is the main reason why a multitude of spray mixtures[1] have been devised and standardized in the course of time.

3. Activation

Having reviewed methods for coating cathodes, we have already indicated that it is the final goal of all the chemical processes to produce, while maintaining the vacuum, an uncontaminated coating of barium and strontium oxides on the surface of the cathode. Experience has shown that such a coating exhibits the property of electron emission only after a more or less complicated activation schedule. In its activated state, the coating is indistinguishable microscopically (and chemically after removal from the bulb) from its unactivated form, and yet the electron emission may differ by a factor of several millions. Our immediate interest is focused now on what goes on during activation.

As early as 1920, Arnold found that the "active material" could be transferred from a well-activated filament to a cleaned-tungsten filament arranged parallel and in close proximity to the former. During this migration process which required some time depending on the temperature, no thermionic activity of the donor filament was lost but the receptive tungsten filament built up thermionic activity evidently from nothing.

In 1924, W. Espe measured the activation of

[1] Ref. 2-1

oxide-coated platinum filaments under very good vacuum conditions, as a function of time, filament temperature, and plate potential. Espe measured the emission at suitable time intervals by applying a definite anode potential for a very short time only. The interdependence of filament temperature, anode potential and time is shown graphically in Fig. 10. In the first test period

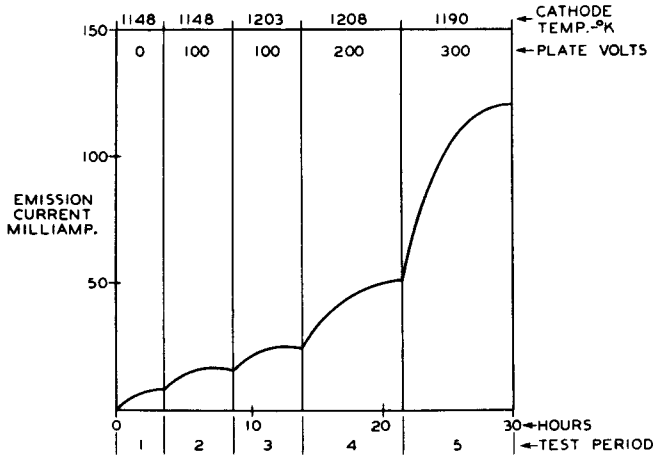


Fig. 10

during which no anode potential was applied, the emission increased but little. During successive test periods with anode potential applied, it can be seen from Fig. 10 that the emission increased more with increases in anode potential that it did with increases in temperature. The activation is, therefore, a function of the space current which flows through the oxide coating. Repeating the experiment at higher filament temperature merely changed the time within which the steady state was reached.

In 1926, L. R. Koller found that the activation of an oxide-coated filament can be destroyed by flashing it at 1600°K and over, but that it can be restored by prolonged operation at about 1100°K, especially when space current is drawn from the filament. Koller found that activated oxide-coated filaments were extremely sensitive to traces of oxygen and water vapor, both of which destroyed the activation completely. He then proposed the theory that the thermionic activity of an activated oxide-coated cathode is due to the presence of minute amounts of free barium metal. We see that the facts presented by Arnold, Espe, and Koller fit this theory very well. Arnold observed the evaporation and deposition of barium metal and Espe observed its production by electrolysis.

If we assume that Koller's theory is correct, in what manner is barium produced in or on the coating? This question is particularly interesting since only a few years ago the preparation of barium metal was considered a difficult task by the chemist. Barium metal is chemically one of the most active metals. A layer of barium metal about 10000 molecules thick on glass disappears instantaneously after exposure to moist air. Barium has a density of 3.8, a melting point of 850°C, and a considerable vapor pressure at 800°C. Barium metal is produced in the chemical industry by electrolysis of complex fluorides, and it is stored and shipped either in evacuated containers or immersed in hydrocarbons, for example, kerosene.

Speculating on all possible methods of its formation in the oxide coating, we have three possibilities which are:

1) Thermal dissociation of the oxide
2) Reduction
3) Electrolysis.

Dissociation is very improbable. By means of Nernst's approximation formula,* we can compute the oxygen dissociation pressure at any desired temperature and find that at 1000°K the pressure is about 10^{-30} microns. In order to produce barium by dissociation, the oxygen pressure must exceed that of the surrounding space. Because the best vacuum we can produce is about 10^{-9} mm, it follows that the temperature required for dissociation will be far in excess of 2000°K. However, such temperature is never required during the activation process.

Experimental evidence indicates that barium metal is produced by reduction. Comparing the activity of an oxide-coated cathode with chemically pure platinum, electrolytic nickel, and commercial nickel, everything else being equal, there is no doubt that best results can be obtained (using one and the same schedule) with the last named material. It has been known for many years that chemically pure platinum or nickel (electrolytic) filaments do not activate as readily as those made from commercial metals. If they activate at all, it is necessary to submit them to a prolonged aging schedule during which barium metal is produced in the coating by electrolysis. This is the most logical interpretation to which we can submit Espe's findings.

Thus, we are lead to believe that impurities in the nickel (or platinum) are capable of reducing the coating of barium and strontium oxides. In order that a slug of cast nickel can be rolled and drawn, it has been found necessary to eliminate nickel oxide in the metallic mass. The most effective and inexpensive expedient consists in adding certain reducing agents to the molten nickel. Such reducing agents are commonly called scavengers. As scavengers for nickel, silicon, titanium, manganese, magnesium, carbon, and calcium are used. When these materials react with nickel oxide, the respective metal oxides are formed. The oxides (with the exception of carbon

* $\log p = -\frac{Q}{4.571T} + 1.75 \log T + c$, where p = pressure in atmospheres, Q = heat of formation, T = temperature in °C, and c = a constant = 2.8 for oxides.

dioxide) are expelled by the molten nickel and form a slag on top of the melt. Adding the scavengers, of which only one or two of the named materials are used, is an art and is left to the experience of the operator. Adding too little scavenger does not produce the desired results, while too much renders the nickel unworkable again. Thus, it happens that commercial-worked nickel always contains a small surplus of scavenger which is of great importance for the process of activation.

Let us consider now the process of chemical reduction, for example, between barium oxide and metallic manganese. Expressing the process by a chemical equation, we have:

$$2BaO + Mn \longrightarrow MnO_2 + 2Ba$$

$$250000 \text{ cal} \qquad 126000 \text{ cal}$$

As is well known, the formula not only indicates the nature of the reagents but also their quantities. By inserting the atomic weights into the equation, we can compute the quantities in any desired units of weight. But since all scientific data are based upon the gram as the unit of weight, BaO represents one gram-molecule in the language of the chemist, and its weight is therefore 153.36 grams (Ba = 137.36, O = 16.00) Combining 137.36 grams of barium metal with 16 grams of oxygen, we obtain energy in the form of heat. Expressing this heat in calories, the scientific unit of heat, we have 125000 calories per gram-molecule. If we wish to split one gram-molecule of BaO into its constituents, we must expend an equal amount of work, namely 125000 calories. The heat of combustion for BaO and MnO_2 has been written below the equation. From a glance at the figures, it is obvious that the heat we obtain when burning up a gram-atom of manganese with the oxygen taken from the barium oxide is only a trifle more than one half the work required for complete reduction of the barium oxide. Obviously, the reaction does not proceed under ordinary conditions. However, in vacuum at elevated temperature the reaction proceeds in the direction indicated by the arrow because, firstly, heat is supplied externally (supplying the deficiency in heat of combustion) and, secondly, the reaction product, barium metal, because volatile under these circumstances, is removed from the sphere of reaction. The general rule, for which this reaction is but a specific example, is that chemical reactions carried out in vacuum run in the direction of the most volatile products. This is precisely the reason why thorium metal is produced when subjecting a thoriated filament to a "hot shot." Thus, we see that a so-called "mild reducing agent" can under proper conditions reduce a highly refractory compound such as barium oxide.

It seems justifiable to assume that the interaction between the reducing impurity and the oxide of the coating takes place at the interface between coating and base metal. Once the impurity near the interface is used up it is replenished by diffusion from greater depths. The interface is usually distinguished by a gray to dark gray discoloration caused by the accumulation of reaction products and by finely dispersed base material. Electrolytic nickel or very pure grades of nickel do not cause the formation of a dark interface. Konel, an alloy containing approximately 3 per cent titanium, is known to form a very dark interface. The color of the interface affects the cathode temperature perceptibly. Dissolving the coating from a well-activated cathode in very dilute acid permits visual inspection of the interface. Since the interface does not dissolve in acid as readily as the white coating, it may be taken as an indication that the former is chemically different from the latter. Dissolving the coating is a simple test for undecomposed carbonates. However, the removal of the cathode from the tube and the immersion of it in acid must be done without exposing it to the air for but a few minutes. A well-decomposed cathode coating exposed to the air for about 20 minutes may be completely converted to the carbonates, depending on atmospheric conditions.

4. Electrolysis

At elevated temperature, the oxide coating on a cathode becomes a conductor whose conductivity C can be expressed by the formula

$$C = \alpha \, \varepsilon^{-\frac{\beta}{T}}$$

in which α and β are two constants. However, this formula expresses the temperature dependence of solid electrolytes and of semiconductors as well. Summarizing the experimental evidence, we are apparently justified in assuming that the conductivity of the pure alkaline-earth oxides (in the unactivated stage) is ionic while that of the oxides in the activated state is predominantly electronic. The impurity causing the activation is alkaline-earth metal and it seems permissible to classify the activated oxide coating as an impurity conductor. There are two types of impurity conductors. One type includes those conductors in which free metal is responsible for the conductivity. In these, more metal is present than is indicated by the chemical formula M_xO_y. Examples are BaO, SrO, CaO, ZnO. The other type of impurity conductors contains more oxygen than is indicated by the chemical formula. In this case, oxygen assumes the role of the impurity. An example is copper oxide. Only such metals forming two or more oxides like CuO and Cu_2O can become impurity conductors of this type. In measuring the conductivity of the coating, it has been observed that a remarkable parallelism exists between conductivity and thermionic activity. The conductivity increases with progressing activation. On the other hand, good conductivity

does not necessarily mean high thermionic activity because, as we know, impurities other than alkaline-earth metals may affect the conductivity but may poison the emission. Furthermore, both conductivity and electron emission may be poisoned by the same agent, such as oxygen sulfur, or chlorine, all of which react rapidly with free barium metal.

Since ions migrate through the coating during electrolysis, barium metal deposits at the interface (cathode) while oxygen appears at the surface (anode) of the coating. Due to its far greater volatility (at the temperature of operation), the oxygen combines with the getter. The process of electrolysis goes on enriching the coating with barium metal until finally the conductivity has changed from ionic to electronic. The evolution of oxygen has been ascertained by means of mass spectrographic determinations, while the presence of barium metal in the coating has been also measured quantitatively. Several investigators have presented evidence indicating sustained electrolysis even in a well-activated coating. Their findings indicate that the ratio of ion to electron current is one to several hundred. Summarizing the evidence, we are led to believe that the conductivity of the oxide coating at the operating temperature is both ionic and electronic. If the coating is well-activated and therefore contains free metallic barium, the conductivity is predominantly electronic. However, if the free barium metal in the coating reserve is depleted, for some reason or other, the conductivity becomes ionic. In this latter condition, barium metal is produced by electrolysis, and may accumulate in the coating, until finally the conduction becomes electronic again.

5. The Amount of Barium Metal in the Coating

The simplest and most reliable determination of the amount of alkaline-earth metal produced by the oxide coating is due to T. P. Beredenikowa. The method is based upon the well-known reactions:

$$Ba + H_2O = BaO + H_2$$

or

$$Ba + 2H_2O = Ba(OH)_2 + H_2$$

Since pressure measurements can be carried out with great accuracy down to pressures of 10^{-6} mm, and by working with a vacuum system of smallest possible volume, we can measure 10^{-8} grams of alkaline-earth metal with certainty. The method consists of causing the alkaline-earth metal to react with a liberal surplus of water vapor, which is admitted to the vacuum system after proceeding with the activation to a desired degree, and then freezing out the excess water vapor with liquid air and measuring the pressure of the remaining hydrogen. From the known volume of the system and the pressure, the amount of barium can be calculated.

If it is desired to find the amount of free barium metal dispersed in the cathode coating, one has to bear in mind that some of the barium metal evaporates during activation and deposits on the glass wall or other parts of the tube. Carrying out the measurement in the manner described, we would find the total amount of barium metal consisting of that in the cathode coating and that on the walls and other parts of the tube. In order to measure the amount of barium in the coating only, it is necessary to transfer the cathode, in vacuum, from the tube in which it has been activated to another vessel and to disconnect the tube from the vacuum system before beginning with the measurement. In this manner the writer has carried out many measurements of the free barium metal in the coating.

A summary of what is known at present of the methods of production of free, alkaline-earth metals in the coating and their relative amounts are given in the following:

1) Testing BaO on pure platinum up to 1200°C showed no thermal dissociation within the limits of sensitivity of the equipment.

2) The yield of alkaline-earth metal due to electrolysis varies from 1×10^{-8} to 9×10^{-8} per coulomb depending on chemical and physical properties of the oxide layer.

3) The ionic conductivity is but a small fraction of the total conductivity which is mainly electronic.

The following results, although accurately determined, require verification on a still larger number of samples.

4) Coprecipitated double carbonates on commercial nickel yield approximately 0.015 mg barium metal per cm^2 per minute at the ordinary "hot shot" temperature (approximately 1350°C brightness temperature, corresponding to 14 volts on the heater in the cathode of a 6K7 tube).

5) Only barium (no strontium metal) is produced by the reduction due to impurities in the nickel.

6) About 25 to 50 per cent of the barium metal produced by the reduction evaporates from the coating and deposits on the anode, bulb, and other tube parts. The remaining quantity of barium metal stays in the coating.

7) A well-activated cathode contains adsorbed in the coating from 3 to 100 times the amount of barium metal which would be needed to coat the macroscopic surface with a monatomic layer.

8) Various well-activated "vapor-process cathodes" (see the following paragraphs) contained from 1-1/2 to 2 times the amount of barium metal which would be needed for coating the macroscopic surface with a monatomic layer.

6. Adsorption

So far we have considered cathode coating whose thickness measured tens of thousands of molecular layers. We also touched upon the findings of Arnold who observed that a clean tungsten fila-

ment, arranged near an oxide-coated cathode, becomes emissive although no visible change of the tungsten surface can be detected. We know now that the thermionic activity of Arnold's tungsten filament is due to the adsorption of barium metal and since such effects are of great importance, both technically and commercially, a short discussion of the fundamentals may be in order.

We are accustomed to regard an atom as composed of a nucleus surrounded by a number of electrons revolving about the nucleus in definite orbits. If we remove a negative quantity, an electron, from the atom, which as such is uncharged, a positively charged body, called a positive ion, remains. We know that we can remove only the outermost electrons which are loosely bound to the atom. If an atom becomes adsorbed on a metallic surface and no change in the number of its electrons takes place, we speak of an adsorbed atom or adatom. If one of the electrons becomes associated with the atoms of the metal, the atom is changed to an ion and since it is adsorbed we speak of an adion. It is also conceivable that the adatom shares one electron with the metal in which case it may change its identity from adatom to adion and back about 10^{16} times per second. We believe this happens in view of recent computations. Metal atoms which have a low work function (or low ionizing potential in the vapor state) lose their valency electrons in relatively weak fields. If such metal atoms, for example, caesium or barium, become adsorbed on a metal with high work function, such as tungsten, the adsorbed film consists of adatoms and adions. Thus, the adsorbed film which must be not more than one atom deep, possesses properties which are not common to the adsorbed metal in bulk.

Let us consider a condenser whose plates C and A are made of tungsten, and short-circuited by a tungsten wire, as shown in Fig. 11. If we introduce a positively charged grid G between

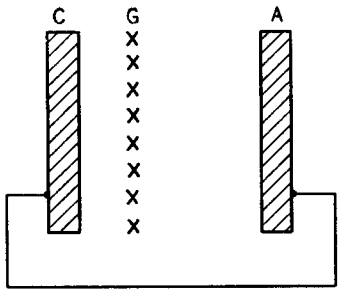


Fig. 11

the two condenser plates, fields are produced between C and G, and between G and A. Also, there will be a potential difference between C and G and an equal potential difference between G and A. Moving G closer to C causes the field between C and G to grow stronger. If we imagine the grid to consist of atoms and positive ions one atomic distance away from C, we have a picture of adatoms and adions and the fields they produce. Opening the electrical connection between both plates causes the adsorbed film to manifest itself by causing a contact potential difference and, because of the strong field between C and G, the film aids the egress of electrons from C, i.e., it lowers the work function of C.

The reduction in work function due to adsorbed ions has been computed by J. A. Becker. He has made certain simplifying assumptions and has supposed that the physical dimensions of adions and adatoms are equal to their dimensions derived from X-ray data. Comparing computation and experimental data, Becker concluded that on the average 10 per cent of the adsorbed particles are ionized. We must bear in mind that due to strong assymetric forces adatoms and adions must be distorted and are, therefore, different from those in free space. The distortion must have some effect upon the work function. A superficial correlation between work function and atomic volume has been pointed out by Schottky, Langmuir, and Dushman. While we know that monatomic films arrange themselves in conformity with the underlying lattice, it must be admitted that we do not possess sufficient information to compute the decrease of the work function due to adsorbed particles. Finally, when we deal with the magnitude of surfaces in evaluating experimental results, great errors may occur. Bowden and Rideal measured the ratio between actual and apparent surfaces. If this ratio is unity for polished nickel, it is about 10 for sandblasted nickel, and 45 for oxidized and reduced nickel. The measurements of the writer pertaining to the barium content of vapor-process cathodes were carried out with polished surfaces.

The difference in thermionic activity between monatomic layers of barium and barium in bulk can be seen qualitatively from a diagram in Fig. 12 by Ryde and Harries. They measured emission, as a function of time, from a tungsten filament

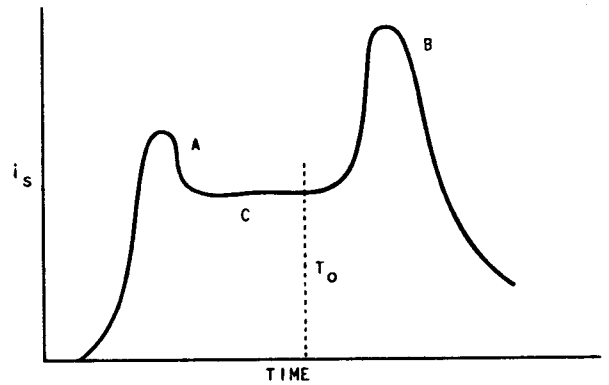


Fig. 12

on which barium metal was allowed to condense. The evaporation of barium onto the filament could be stopped at will.

With the tungsten filament at a given constant testing temperature and barium condensing thereon, the emission rose to a first maximum A. The emission then decreased due to the deposition of more barium, and finally attained a constant level C. At time T_o indicated by the dotted line, the barium source was cut off. Then, due to evaporation from the filament, the emission rose to a maximum B at which time the barium layer had been reduced to a monatomic film. The decrease after B depends upon the filament temperature and is less rapid the lower the temperature.

7. The Barium-Vapor Process

It has been found that the depletion of an adsorbed monatomic layer of barium due to evaporation can be greatly retarded, and concurrently the emission can be enhanced, by interposing a monatomic layer of oxygen between the base metal and the barium layer. This can be done, for example, by flashing a tungsten filament at about 2200°K in vacuum in order to clean its surface and exposing it for a few seconds to oxygen after cooling. If barium is then allowed to condense upon this filament, an emissive layer is formed with thermionic properties vastly superior to the pure barium film.

This process has been used commercially in Europe and is usually referred to as the "barium-vapor process." It has been carried out either by oxidizing a tungsten filament in air till it shows a blue color or by coating a tungsten filament with a thin film of copper oxide, and then mounting such a filament in a tube whose anode or grid has been painted liberally with barium azide (BaN_6). During the bakeout, the barium azide decomposes to barium metal and nitrogen, the latter being removed by the pump. After the tube is sealed off, the filament is heated to about 1100°K. A small positive potential (about 50 volts) is applied to the anode while it is heated gradually by induction to vaporize the barium metal. The sudden appearance of a green glow in the tube indicates the presence of barium vapor which reacts with the oxide of tungsten or with the copper oxide on the filament. The surface of such a filament consists of a mixture of barium oxide, and barium metal interspersed with the reduced metal on the surface (W or Cu). While the process is carried out with an abundance of material, i.e., more than required to form monatomic layers of oxygen and barium, it appears that most of the excess material is bombarded off during aging. In fact, a filament activated by this process exhibits a clean metallic surface. Such tubes give good life performance of several thousand hours. The process, however, is inherently messy because the surplus barium metal causes leakage on the insulating structural members. Furthermore, barium azide is highly explosive and poisonous.

8. Emission Constants

While in the foregoing we have indicated that among the alkaline-earth oxides mixtures of barium and strontium oxide exhibit the largest electron emission, a comparison of the emission constants of other substances may further elucidate this. In order to facilitate the comparison in practical units, we have included in the following table computed values of emission/cm^2 at 1000°K.

Substance	i_s/cm^2 at 1000°K milliamp.	φ	A amp./$cm^2 deg^2$
CaO	0.003	1.95	.0002 to 0.02
SrO	0.6	1.42	.0002 to 0.02
BaO and SrO	300	1.1	.0013 to .000013
Ni	1×10^{-14}	5.0	60.0
W	5×10^{-6}	4.5	60.0
W - Ba		1.56	1.5
W - O - Ba	>200, <300	1.1 to 1.34	0.18
Ba in bulk		2.11	60.0

While the values given in the table have been derived from rather recent publications, they cannot but merely convey a relation of the order of magnitude. For it is well known that the thermionic activity of cathodes with composite surfaces depends on the previous treatment and the physical condition of the oxides. It is also well known that the technique of the preparation of oxide-coated cathodes has progressed remarkably in the past decade.

From a practical point of view the efficiency of emission (milliamperes per watt) as a function of power input (watts per cm^2) is of great importance. In Fig. 13 is shown a special cross-section paper by means of which comparison between various emitters can be made easily. The figure shows a comparison between pure and thoriated tungsten, while curves 1, 2, 3, and 4 indicate emission of oxide-coated cathodes made in 1923, 1926, and 1935, respectively.

9. The Barium-Strontium Ratio and Impurities in the Coating

During the past 15 years, the oxide mixture which has been found to give the most reliable results in practice, is composed of equal molecular proportions of barium oxide and strontium oxide. Emission measurements of mechanically mixed and coprecipitated carbonates with barium content ranging from 0 to 100 per cent were carried out by Benjamin and Rooksby and by Widell and Panofsky. The results are given in Fig. 14. As indicated by the curve of Widell and Panofsky, barium content between 25 to 75 molecular per cent exhibits the same emission efficiency.

Measurements on pure barium oxide are difficult to carry out and are, therefore, uncertain since pure barium carbonate melts on the cathode and often contains gas inclusions. Pure strontium oxide has not been obtained, for even the best preparations contain a fraction of 1 per cent barium oxide which cannot be eliminated chemically. Thus, the end points of both curves in Fig. 14 have to be regarded as uncertain.

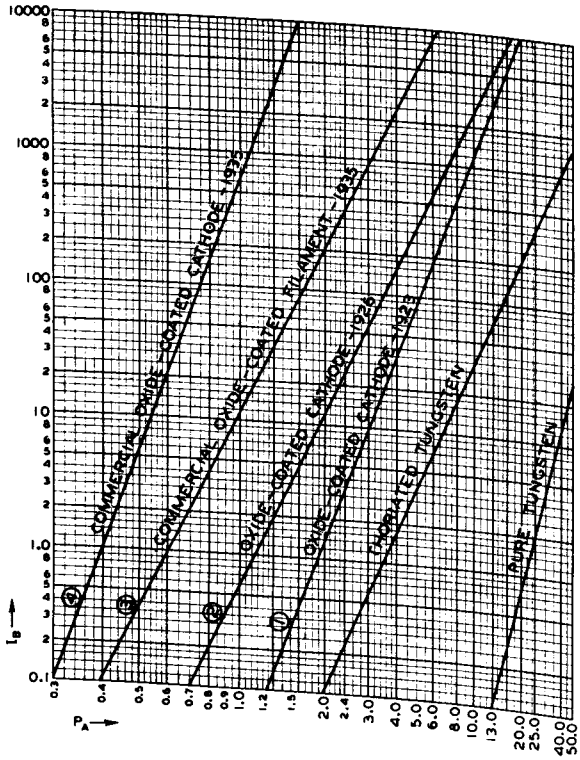


Fig. 13

About 10 years ago the writer carried out a series of experiments testing the effects of various impurities admixed to a standard coating material. It was found that metal oxides of an acidic nature such as WO_3, MoO_3, Ta_2O_5, etc., when present in amounts less than 0.1 per cent by weight practically render the coating worthless. Other oxides which were expected to be indifferent to the coating, such as MgO, ThO_2, and CeO_2, also reduced the emission but probably only due to changes in texture, since such additions even if present in amounts smaller than 0.5 per cent by weight would affect the melting point of the mixture. The effects of MoO_3 upon the emission are sometimes encountered in the manufacture of radio tubes in which the control grid is made from molybdenum wire. When such tubes are aged using grid bombardment, a small amount of MoO_3 may be formed on the surface of the grid. This oxide when vaporized or sputtered off may reach the cathode and will then poison the emission. While impurities in the coating inhibit the activation, a well-activated oxide-coated cathode can be poisoned by exposing it to gases, such as oxygen, water vapor, chlorine, etc. Considering that the maximum amount of free metallic barium observed in an activated coating is 1×10^{16} atoms per cm^2 or approximately 2 micrograms, we need only 0.2 liter micron of oxygen to deactivate the cathode completely. Thus, gases liberated from the electrodes or glass wall of the tube by electron bombardment may be sufficient to impair the electron emission. A deactivated cathode may be temporarily restored by (1) a hot shot, (2) prolonged aging (electrolysis), (3) condensing barium from another source upon its surface, or (4) exposing it to bombardment of positive ions of an indifferent gas (CO). When a cathode is deactivated, the texture of the coating is generally impaired, so that the above remedies do not give practical satisfactory results. Much has been written and claimed about the advantages of certain gases upon

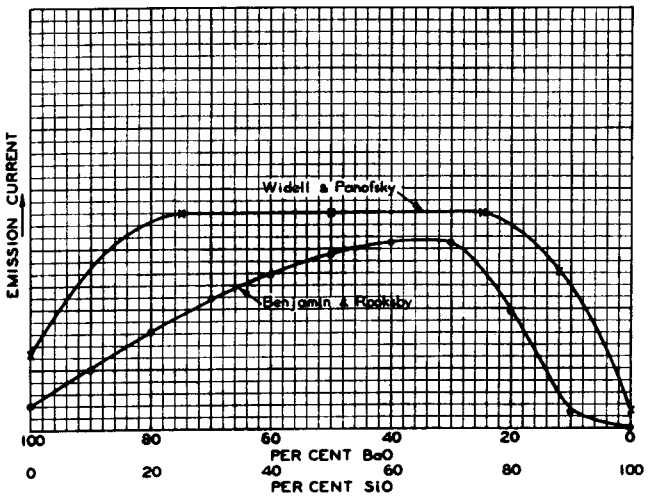


Fig. 14

the progress of activation. Such procedures, except perhaps in isolated cases, have never found general application. The best rule is to activate the cathode in the best obtainable vacuum.

10. The Mechanism of Emission: Seat of Emission

In view of our discussion of oxide-coated cathodes, we can now form a qualitative picture as follows: At elevated temperature, electrons diffuse to the surface of the metallic core. Without a coating on the core, all but a few very fast electrons are driven back into the core by the image force. When the core is provided with a coating of selected properties, barium metal is produced by reduction and by electrolysis. The adsorbed atoms become ionized due to thermal ionization and the adions compensate the image force. Thus, a large number of electrons can pass the boundary between metal and oxide. Since barium readily diffuses outwardly into the coating, the coating contains throughout its entire mass a large number of adions and adatoms. The electrons emerging from the metal to the oxide

move through the semiconducting layer of oxide to its surface where they finally escape into free space.

The coating must have certain properties which are:

1) A certain conductivity both ionic and electronic.

2) Chemical indifference toward free barium metal.

3) Ability to adsorb and hold barium atoms and ions.

The first two requirements are fully met by barium-strontium oxide as discussed in a previous paragraph. The last requirement appears to be a function of the crystalline shape of the unit cube and of its dimensions. Only when duplicating the crystalline structure and order of magnitude of the unit cube, but with chemically different oxides, has the writer been able to produce artificial oxide-coated cathodes having fair emission and good life.

From the foregoing it appears superfluous to discuss the seat of emission. However, since this topic has been the subject of much controversy, an experiment by Widell should be mentioned. He measured the perveance of cathodes with thick and thin coatings in diodes. From these measurements, the origin of the emission was computed and found to coincide precisely with the surface of the coating. In interpreting these results, one must remember that the conductivity of the coating follows a different law than that controlling the perveance. Besides, since the coating is a conductor, the field of the anode most probably ends on the outermost layers of the coating. Hence, the measurements indicate clearly the point of emergence of the electrons which no doubt is the surface of the coating.

11. Reverse Emission and Grid Emission

In the preceding paragraphs, we have pointed out that barium metal and oxygen can evaporate from an oxide-coated cathode during life and during activation. We have also discussed the methods of preparation of barium-vapor-process cathodes. Under the conditions of preparation, it is indeed surprising that grid and reverse emission are not a common property of every tube which is manufactured. Still more favorable for the occurrence of grid and reverse emission is the fact that every tube electrode is coated by at least a monatomic layer of oxygen or metal oxide. This is so because, after firing the electrodes in a reducing atmosphere, they are exposed to the air during mounting. Recent experimental evidence proves the presence of such oxide films under such conditions. Fortunately, save for a few exceptions, most tubes are so designed that under normal operating conditions neither plates nor grids attain temperatures at which grid and reverse emission manifests itself. In the light of our present knowledge, it seems that this critical grid temperature is about 300°C. Attempts to combat grid emission with special alloys or coatings on grids have thus far been only partially successful, because any substance which poisons grid emission on the grid sputters sufficiently so as to poison the forward emission of the cathode in time. The only remedy available at present is the design of the electrodes in such a manner that their operating temperature never exceeds 300°C. Of course, copper side-rods, grid wires of high heat conductivity, grid radiators, large grid-cathode spacing, and efficient blackening of the anodes are expedients to maintain low electrode temperature.

12. The Poisoning Effect

In activating a cathode near a bright nickel anode, one always observes a discoloration of the nickel surface facing the oxide coating. If after the hot shot a very low plate voltage (less than 5 volts) is applied, the plate current is generally stable. With gradual increase in plate voltage (E_b), the plate current (I_b) stops to increase or even begins to decrease at a definite potential. Fig. 15 elucidates this behavior but is somewhat exaggerated.

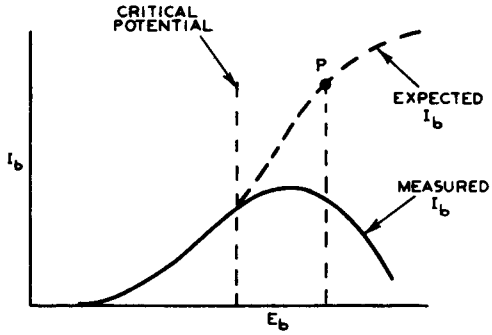


Fig. 15

Testing again under the original conditions with low E_b, we find that I_b has decreased very little but often shows a tendency to increase. Repeating the testing procedure at the higher E_b values, we find that the phenomenon repeats.

Headrick and the writer have investigated this phenomenon and have built many tubes in which the anode could be moved in such a manner as to expose a clean nickel surface facing the cathode. The discoloration on the anode caused by the activation of the cathode has been called the activation spot. The electrode arrangement with the anode rotated 90° after activation of the cathode is shown in Fig. 16. By moving the activation spot away from the cathode and measuring I_b again as a function of E_b, the dashed curve of Fig. 15 was obtained. Maintaining the plate voltage corresponding to point P (Fig. 15) and rotating the anode so that the activation spot faced the cathode again, caused a rapid drop

of the plate current. The test was repeated several times with the same results. The anode was carefully centered in order to maintain constant cathode-anode distance at any point.

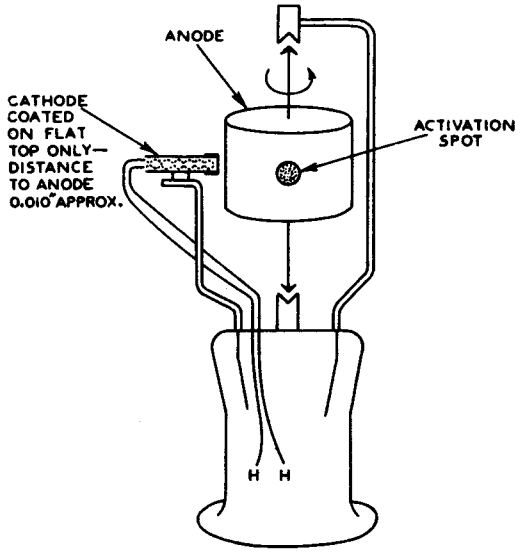


Fig. 16

It is not difficult to explain this phenomenon. The emission of any oxide-coated cathode depends on the maintenance of a certain barium-metal reserve. If more barium is produced by electrolysis and by reduction than is lost by evaporation and gas poisoning, the emission remains stable. If the two last-named influences predominate, the emission drops. We know also that the barium production due to reduction and electrolysis is greatly enhanced with increasing temperature.

The fact that the activation spot appears as a discoloration indicates the presence of foreign material. Firstly, the activation spot contains more nickel oxide than any section of the anode of equal size, because the spot receives heat by radiation from the cathode and is exposed while hot to the gaseous decomposition products of the cathode. During the activation of the cathode, this spot becomes contaminated with barium metal which subsequently becomes oxidized at least in part, due to gaseous residues in the bulb (although a batalum getter was used) and due to oxygen given off from the cathode (electrolysis). The final result is that if the activation spot is bombarded with electrons, the nickel oxide and barium oxide thereon decompose and oxygen diffuses back to the cathode combining with the free barium in the coating. At a certain critical voltage, the oxygen evolution is so fast that it exceeds the rate of barium evolution in the cathode and thus the emission drops due to oxygen poisoning.

Computing the $mv^2/2$ of the electrons as a function of anode voltage and comparing this energy with the heat of formation per molecule of nickel oxide, barium oxide, and aluminum oxide, and presupposing a complete energy transfer between impinging electron and molecules of metal oxide, one obtains the following critical potentials, which have been checked by experiment.

NiO	3.9 Volts
BaO	9.5 Volts
Al_2O_3	25 Volts

In general, the critical poisoning anode potential is lower the smaller the cathode-anode distance, the lower the cathode temperature, and the lower the barium reserve of the cathode. The cathode-anode distance merely affects the concentration of oxygen per unit volume which diffuses back to the cathode.

A similar poisoning effect in the type 6K7 has been investigated recently by Wamsley and the writer. The effect manifested itself as slumping transconductance (g_m) when the tube was tested at 5 volts on the heater. The graphs in Fig. 17 were made to show the behavior of g_m with time.

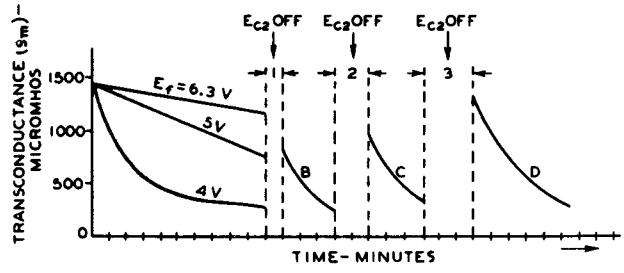


Fig. 17

The curves show clearly that the lower the cathode temperature, and hence the lower the rate of barium production, the more rapid the drop in g_m. Curve B shows the drop in g_m at 4 volts on the heater after the screen-grid potential (E_{c_2}) had been interrupted for one minute. Curve C shows the drop in g_m after removal of screen potential for two minutes. The recovery in g_m indicates that during the off period, barium was supplied to the coating by diffusion from the interface. Next, by measuring contact potentials of grid No.1, the barium production, migration, and subsequent oxidation during the operation of the tube was determined. Thus, it was found that a layer of barium oxide on grid No.2, bombardment of this oxide by electrons, and poisoning of the cathode due to oxygen derived from grid No.2, were responsible for the slumping g_m. Heating grid No.2 by electron bombardment for about one minute to a temperature of 900°C removed the oxide layer thereon and with it the causes for slumping g_m.

13. Aging

Aging has a twofold purpose. The residual gas must be removed from the tube and the cath-

ode must be activated so as to produce the maximum thermionic efficiency. We know from the foregoing discussion that activation can only be attained if the gas pressure in the device is low. One of the preferred expedients in reducing gas pressure is to ionize the gas and cause it to combine with the getter in the tube. Since under these conditions the activation of the cathode is not developed, one has to resort to higher than normal cathode temperature so that sufficient ionization takes place. Therefore, any aging schedule for oxide-coated cathodes will consist of two steps:

1) Ionization of the gas residue in the tube, heating the cathode above normal temperature, and drawing a comparatively large space current. Removal of the ionized gas by the getter.

2) Activation by reduction (hot shot), or by electrolysis, or by a combination of both.

14. Temperatures of Oxide-Coated Cathodes

Experience has shown that oxide-coated cathodes should operate between 1000° and 1100°K brightness temperature to give best life performance. At temperatures above 1100°K, the rate of evaporation of barium metal becomes very high and the useful life is shortened. Below 1000°K, the rate of barium production may become so low that adverse effects (poisoning, for example) may cause falling emission; furthermore, the proper balance between the rate of production and the rate of removal of barium is disturbed. In line with these facts, it has been known for years that cathodes having low operating temperature require much better vacuum conditions than those having high cathode temperature.

Measuring the cathode temperature by means of an optical pyrometer of the disappearing filament type enables one to obtain the brightness temperature. Since the coating is a poor heat conductor and since the coating is being heated by the core, it follows that the temperature of the coating is lower than that of the core. There is a temperature gradient from the interface to the outer surface. It has been found that the coating is partially transparent for the radiation of the core and interface so that in measuring brightness temperature of the coating it is not the surface temperature which we measure but that of some intermediate layer. The relation between true core temperature, thickness of the coating, and brightness temperature is given by Widell and Perkins[2] in the curves of Fig. 18.

That the brightness temperature depends on the amount and color of the interface and, therefore, also on the kind of core material, can be seen from the curves of Fig. 19. Hence, it is obvious that specifying the cathode temperature in watts per cm^2 is somewhat ambiguous. Watts per cm^2 should be used only if we wish to make comparisons or computations, and even then only when one kind of core material of general established properties is employed.

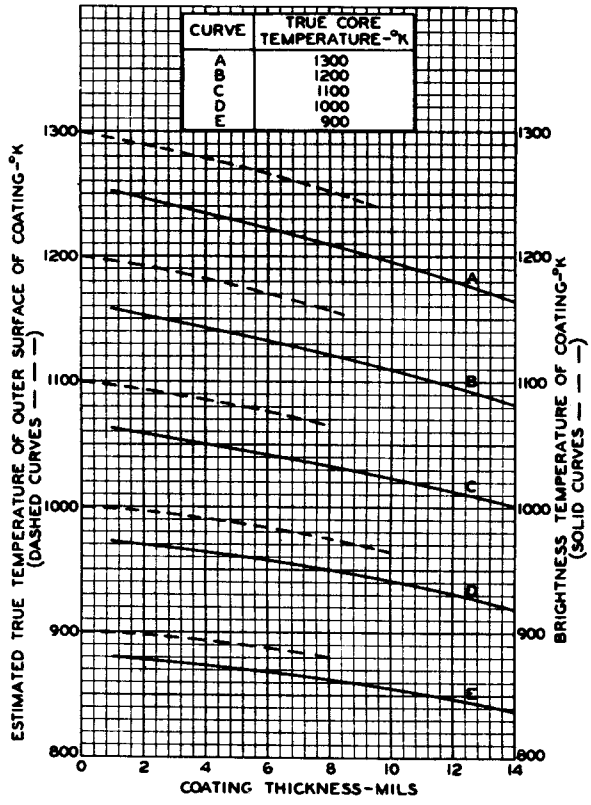


Fig. 18

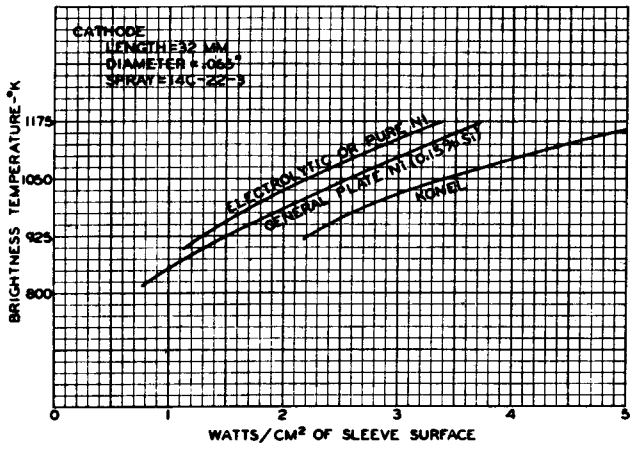


Fig. 19

Another factor influencing the temperature of the cathode is the temperature of its surroundings. The effects of "back heating" are well known and may be computed roughly by applying Stefan-Boltzmann's radiation law.

Furthermore, the temperature of the cathode is not uniform along its length, and is affected by end losses. The heat conductivity of cath-

[2] Ref. 2-2

ode tabs, the thermal contact between cathode and mica or the filament supports, as well as the design of the heaters in indirectly heated cathodes play an important role.

For the design of indirectly heated cathodes, the ratio of the thermal emissivity of the coated to bare-core material is of importance. With the improvement of the technique in the manufacture of oxide-coated cathodes, this ratio must be redetermined frequently. A recent determination by Widell shows this ratio to be 1.6 for the present stage of technique. The spectral emissivity of an oxide-coated cathode (for average conditions) has also been determined by Widell and found to be about 0.56 at $\lambda = 0.665\mu$.

15. Calculation of Oxide-Coated Filaments

Generally, the electrical constants of a tube are set by convention. For example, the filament voltage and current have to meet certain specifications. From the desired value of g_m the filament-grid spacing, and other mechanical considerations, we can determine the length of the filament. Thus, we have

Filament length = L
Filament voltage = E_f
Filament current = I_f

We also know the desired operating temperature to maintain good emission life which is approximately 1050°K or 3.2 watts per cm^2 = D, and the specific resistance ρ of the core material at the operating temperature.

We wish to know the width (w) and the thickness (t) of a filament which having a length (L) and operating at $E_f \times I_f$ watts input attains a temperature of 1050°K. To solve for the two unknowns, w and t, we need two equations. These are:

$$\frac{E_f I_f}{2(w + t)\, L} = D$$

$$R = \frac{\rho\, L}{w\, t} = \frac{E_f}{I_f}$$

On substituting we get:

$$w = \frac{E_f I_f}{4LD} + \sqrt{\left(\frac{E_f I_f}{4LD}\right)^2 - \frac{\rho\, L}{R}}$$

$$t = \frac{E_f I_f}{4LD} - \sqrt{\left(\frac{E_f I_f}{4LD}\right)^2 - \frac{\rho\, L}{R}}$$

In these equations, dimensions are to be expressed in centimeters and electrical values in volts, amperes, and ohms.

REFERENCES

O. W. Richardson, "Emission of Electricity from Hot Bodies." (Second Edition) Longmans, Green & Company, London, 1921.

L. A. Reiman, "Thermionic Emission." John Wiley and Sons, New York, 1934.

S. Dushman, "Thermionic Emission," Review of Modern Physics, Vol. 2, No.4; 1930.

K. T. Compton and I. Langmuir, "Electrical Discharges of Gases," Part I, Review of Modern Physics, Vol. 2, No.2; 1930.

V. K. Darrow, "Statistical Theories of Matter, Radiation and Electricity." Bell Telephone Laboratories Reprint B-435; 1929.

P. W. Bridgeman, "Thermodynamics of Electrical Phenomena in Metals." The Macmillan Company, New York, 1934.

J. H. DeBoer, "Electron Emission and Adsorption Phenomena." The Macmillan Company, New York, 1934.

L. R. Koller, "The Physics of Electron Tubes." McGraw-Hill Book Co., New York, 1934.

Lecture 3

HEATERS AND HEATER-CATHODE INSULATION

G. R. Shaw and L. R. Shardlow

INTRODUCTION

With the development of heater-type tubes for a-c operation, which first became commercially active in the early part of 1927, the problem of insulation between heater and cathode began. The first material to be used was a twin-holed procelain rod fitting snugly into the cathode sleeve. The heater was a hairpin of tungsten fitting into the holes. In continuous service, the porcelain was quite satisfactory, but with intermittent operation, it melted onto the heater and caused heater failure by pulling the heater wire apart due to the difference in coefficient of expansion. This was corrected by using, instead of the porcelain, sintered magnesia insulation with a small amount of talc, which had a much higher melting point and consequently left the heater free to move. Soon, however, the demand for a reduction in the heating time of from 30 to 45 seconds to a more reasonable period of from 5 to 10 seconds made it imperative to make a change. The large amount of insulation which had to be heated before the operating temperature at the external surface of the cathode was reached was, of course, the chief obstacle to be overcome. However, the application of only a thin coating of insulation and a reduction in the weight of nickel in the sleeve met the requirements. Involved in the change was the design of the heater, because the decrease in the amount of insulation permitted operation at a lower temperature. To compensate for the decrease in specific resistance resulting from the lowered temperature, it was necessary to increase the length of the heater.

Numerous heater designs have been proposed at various times. These designs can be classified by types:

1. Hairpin, either single or in multiple as in the folded heater.
2. Unidirectional-wound coil, which may be varied by bending it back on itself.
3. Reverse-wound helix. This is the type most used in production-type tubes.

Each of these designs has its own advantages and disadvantages which influence its use in particular applications.

Fundamentally, a complete unipotential-cathode assembly consists of:

1. A heater for the generation of heat by the passage of current.
2. A conducting sleeve surrounding the heater to serve as the base of the operative cathode.
3. A means of insulating electrically the heater from the sleeve.
4. An active coating on the sleeve to serve as a source of electrons.

It is necessary in this section to deal only with the problems arising under items 1 and 3.

Due to the fact that the heater wire will attain a temperature of about 1100° to 1200°C in operation and a temperature of possibly 1600°C in the course of processing during tube manufacture, the choice of heater materials is rather limited. Obviously, only materials with a melting point greater than 1600°C will be possible and, in addition, only those with a vapor pressure so low at that temperature as to cause negligible loss by vaporization in the time required for manufacture, will be useful. Boron, carbon, thorium, titanium, zirconium, tantalum, uranium, vanadium, the platinum group of metals, tungsten, and molybdenum are the only ones having a possibility of fulfilling the requirements. The first six are too reactive with insulating coatings, the next two are not available as wire, and the platinum metals are too expensive. So, tungsten and molybdenum are the only materials available for consideration. By making suitable alloys of these two elements, metallurgists have provided a further field in which the engineer has a choice of materials for his use. The hot strength of molybdenum has been found to be so low under temperatures to which a heater is subjected during tube manufacturing as to cause it to sag under its own weight. This low strength results in a distorted heater and a distorted distribution of the energy supplied to the cathode from it. Molybdenum, therefore, can be eliminated from present consideration.

Only tungsten and some of its aloys with molybdenum are left as sources of filament wire producing satisfactory heaters. The two metals alloy in all proportions, and from all those alloys possible, two have been selected as representative, namely, the 80% molybdenum alloy with tungsten and the alloy containing equal parts of the two metals. The former is known as "H" wire and the latter as "J" wire. The hot strength decreases with an increasing amount of molybdenum, but the ductility and degree of brittleness improve. These factors must be weighed against each other in the selection of a material for heater wire. The difficulty of producing wire also decreases with increasing molybdenum content, so that this factor enters into the problem of heater selection as a matter of material cost. In Fig. 1 is shown the relationship between resistivity and absolute operating temperature of pure tungsten wire, "J" wire, "H" wire, and pure molybdenum wire. While the cold resistance of the various materials is widely different, it is interesting that, at the operating temperature of heaters used in radio receiving tubes, i.e., in the neighborhood of 1300° to 1500°K, there is very little difference in the resistivity. For

approximate calculations of new filaments, it is even possible to assume a round value of 40 microhms per cm. Table I lists some of the useful properties of these four materials. From these data, it is possible to design a heater to operate under the desired conditions of voltage, current, temperature, and size.

Table I

Material	Density	Diameter (mils) $\sqrt{mg/200\ mm}$	ρ (microhm/cm) 300°K	1000°K	1400°K
Tungsten[1]	19.2	.716	5.65	24.9	37.2
"J" Wire	13.25	.863	8.9	27.1	(38.7)*
"H" Wire	11.25	.963	7.0	25.25	(36.7)*
Molybdenum	10.1	.987	5.65	23.9	35.2

* Extrapolated

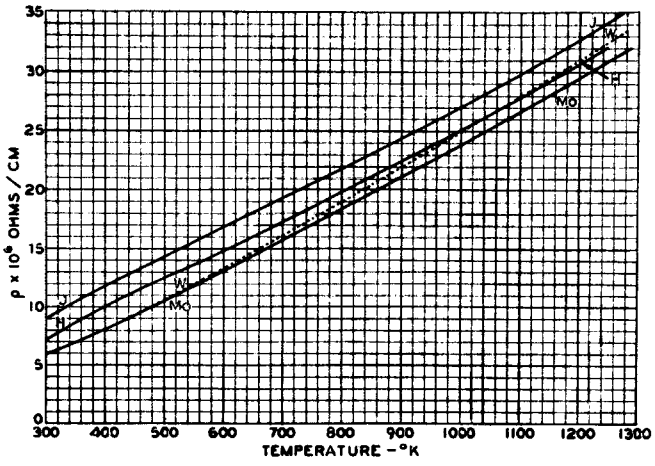


Fig. 1

HEATER DESIGN

It has so happened that in the design of heaters for vacuum tubes, widely variant temperatures have been attained, as illustrated by the values in Table II for different commercial tubes. It will be noted that while the range in operating temperature of 6.3-volt, quick-heating tubes with reverse-helical filaments is only from 1167° to 1235°C, folded filaments vary much more widely.

Table II

Tube Type	E_f Volts	I_f Amp.	Avg. Operating Temp.★ °C	Type of Heater
6D6	6.3	0.3	1187	Reverse helix
6K7	6.3	0.3	1167	Reverse helix
6L7	6.3	0.3	1202	Reverse helix
6N7	6.3	0.8	1167	Reverse helix
6L5G	6.3	0.15	1050	Folded
6J7	6.3	0.3	1060	Folded
6L5	6.3	0.15	1147	Folded
6J5G	6.3	0.3	1235	Reverse helix
6X5	6.3	0.6	1127	Folded
6C5	6.3	0.3	1210	Reverse helix
6F5	6.3	0.3	1205	Reverse helix
956	6.3	0.15	1347	Folded
955	6.3	0.15	1286	Folded
954	6.3	0.15	1197	Folded
6S7G	6.3	0.15	1047	Folded
56	2.5	1.0	1067	Reverse helix
25L6	25.0	0.3	1272	Single helix

★ As measured by hot and cold resistance method.

Table III shows how heater temperature varies with voltage for the type 6K7.

Table III

E_f Volts	I_f Amperes	°C Temperature
3.0	0.188	817
4.5	0.243	987
5.5	0.275	1092
6.3	0.298	1167
7.5	0.330	1287
14.0	-	1725

When a heater is designed, several factors must be considered. Since the value of the volt-

[1] A summary of the properties of tungsten is given by Forsythe and Worthing, Astrophysical Jour. 61, p. 146; 1925.

age E and the current I is set, together with the operating temperature T, which determines the specific resistance ρ, we have

$$\frac{E}{I} = \frac{4 \rho S}{\pi d^2}$$

where S = length of wire for heater.

Unfortunately, we are dealing with energy transfer by both radiation and conduction and no well-defined formula can be given for the rate of transfer of energy from heater to cathode sleeve. This is further complicated by the fact that the geometrical shape of the heater is not a simple one in practice, and so calculation is difficult. In addition, neither the radiating characteristics of the coated heater, as affected by thickness of coating, type of spray, and shape, nor the reflecting properties of the inner surface of the nickel sleeve are known. Consequently, this leads to the method of designing the heater by successive approximations to obtain the desired current, voltage, and temperature. Much stress has been placed on the importance of correctly designed heaters. The reason for this will be given later when we discuss the insulation phases of the subject.

Probably the easiest way to explain heater design[2] is by example. The following calculations apply to a theoretical heater for the 6K7, and are based on the following assumed values: an operating temperature of 1360°K, which has been determined by experience to fall within the temperature "safety zone" of our insulating materials, a dissipation per sq cm of wire surface of 5.5 watts, a current of 0.3 ampere, and a specific resistance of 36×10^{-6} ohms/cm at the operating temperature.

The diameter of wire is given by the equation

$$d = \sqrt[3]{\frac{4 I^2 \rho}{\pi^2 n}}$$

where d is the diameter of the wire in centimeters, I is the current in amperes, ρ is the specific resistance in ohms/cm, and n is the dissipation in watts/cm^2. Substituting numerical values in the equation we have

$$d = \sqrt[3]{\frac{4 \times .09 \times .000036}{\pi^2 \times 5.5}}$$

$$d = .0062 \text{ cm} = .00244 \text{ inches}$$

For purposes of illustration, we will suppose that this heater is to be made of "J" wire. Then, from Table I, we have

$$d_{(mils)} = .863 \sqrt{mg/200 \text{ mm}}$$

$$\left(\frac{2.44}{.863}\right)^2 = mg/200 \text{ mm}$$

$$8.02 = mg/200 \text{ mm}$$

The cut length of wire for the heater is found from the equation

$$S = \frac{E I}{\pi d n} = \frac{6.3 \times .3}{\pi \times .0062 \times 5.5} = 17.6 \text{ cm}$$

To determine the turns per inch (TPI) in each half of the double helix of a heater made from wire of the above specifications, we have the equation

$$TPI = \frac{\sqrt{\left(\frac{S}{2L}\right)^2 - 1}}{\pi (D + d)}$$

where S is the length of wire for the coil in mm, L is the coil length in mm, D is the mandrel diameter in inches, and d is the wire diameter in inches.

Assume that we are constructing a coil for a 45-mil cathode, and that the wound length of the coil shall be 23 mm, using a 17-mil mandrel, and that the leg length on the coil is 5 mm. Substituting numerical values in the above equation, we have

$$TPI = \frac{\sqrt{\left(\frac{166}{46}\right)^2 - 1}}{\pi (.017 + .00244)} = 57$$

As mentioned before, the radiation characteristics of the coating, as affected by thickness, type of spray, and reflecting properties of the inner face of the sleeve, all affect the design of a heater, but they are to all intents and purposes unknown. Consequently, the design of a heater in its final analysis becomes a series of successive approximations. However, let us compare the theoretical heater we have made for the 6K7 with the actual heater now being used.

Table IV

6K7 Heater	Theoretical	Actual
Wire	J	J
Wire diameter (mils)	2.44	2.46
Wire weight (mg)	8.02	7.94 to 8.26
Wire length (mm)	176	165
TPI	57	58
Operating temp. (°K)	1360	1440

[2] Ref: 3-2

The next step is to take the theoretical heater and test it in a tube to determine its current and operating temperature. It is at this point that the "series of successive approximations" begins. These calculations are based on Ohm's law, which is E = IR.

Remembering that resistance is inversely proportional to cross-section (which can be expressed as wire weight), other factors remaining constant, and that resistance is directly proportional to length, we can correct the theoretical heater. For instance, it will be noticed in Table IV that the cut lengths of the theoretical and actual heater do not agree by approximately 6.5%, and that the operating temperatures are approximately 80° apart. The correction in cut length was made by the use of Ohm's law as shown above.

INSULATING MATERIALS

The problem of insulation of heaters is a difficult one. In the automobile industry, where ceramic bodies are used as a means of insulation in spark plugs, the requirement of low conductivity at high temperatures is recognized as one of the most severe to be met with in the commercial field. But translated into terms of design practice of modern radio tubes, the insulation would not be satisfactory by a factor varying between 100 and 100000, depending on the type of tube, the temperature at which it operates, and somewhat on the method of using the tube.

It is evident that the choice of suitable materials for insulating purposes is limited to a marked degree. The better grades of porcelain were used at first, but their low melting points (as compared with the conditions of use), their low electrical resistivity, and their reactivity with the heater wire rendered them obsolete. The coefficient of thermal expansion of porcelain is only about half that of the pure oxides now generally used and matches that of tungsten heaters more closely, but, in spite of this, only the pure oxides offer the properties which are needed for modern tubes. Of the oxides available, only alumina, magnesia, beryllia, zirconia, and thoria have a sufficiently high melting point to offer the possibility of satisfactory service. Some of the properties of these oxides are given in Table V.[3]

Since the insulation must necessarily operate at high temperatures, its conductivity as related to temperature is perhaps its most important property to be considered. Werner[4] gives an idea of the relationship between the resistance of the oxide and the temperature at which it must operate. The actual value of conductivity is dependent on several factors, one of the most important of which is the porosity of the oxides.

Table V
PROPERTIES OF REFRACTORY OXIDES

Fused Oxide	Melting Point (°C)	Specific Gravity	Heat Capacity (cal/g 50°C)	Linear Coef. of Expansion (25°- 800°C)
ThO_2	3050	9.3 - 9.7	0.059	9.3×10^{-6}
MgO	2800	3.6 - 3.7	0.232	13.4×10^{-6}
ZrO_2	2687	5.7 - 6.1	0.106	6.6×10^{-6}
BeO	2570	3.0	0.261	7.5×10^{-6}
Al_2O_3	2050	3.9 - 4.0	0.198	7.9×10^{-6}
*SiO_2	1713	2.2	0.185	0.5×10^{-6}

* Vitreous

This is more or less obvious, since the conductivity of the air or vacuum spaces is less than that of the oxides themselves, especially at high temperatures. The resistance increases roughly according to the percentage of porosity of the body under test. Unfortunately, no data similar to that quoted above are available as regards the resistance of beryllia, but from the results available, the data should be approximately the same as for alumina.

Navias[3] has reported on the reactivity of the oxides with the heater wire, an important factor in the choice of insulation. Zirconia shows signs of reaction, as indicated by a darkening of the coating, even under the conditions of preparation, i.e., firing in hydrogen at about 1700°C. Thoria is reduced by a hot, tungsten filament so that it gets dark and gives distinct evidence of the presence of metallic thorium by the very high electron emission available from such an assembly. At 1700°C, magnesium oxide will react with tungsten to the extent of giving a bright, mirror-like deposit of the metal on the cold walls of the containing vessel. Only alumina and beryllia will withstand high heating in contact with tungsten without evidence of vaporization, dissociation, or reaction, and they will remain white up to a temperature of at least 1800°C. It is natural to expect that the conductivity at these temperatures can be rated in terms of the reactivity between them and the heater.

Since it appears that only alumina and beryllia offer the properties to make their use for heater insulation possible, it is of some interest to examine a few of their properties. Bragg[5] has shown that their atomic structures have some important similarities. In each case, the oxygen atoms are in approximately the closest

[3] Navias, Jour. Amer. Cer. Soc., Vol. 15, p. 248; 1932.

[4] Werner, Sprechsaal, Vol. 63, pp. 437, 557, 581, 599, 619; 1930.

[3] Loc. cit.

[5] W. L. Bragg, "Atomic Structure of Crystals," Cornell University Press, p. 94.

hexagonal packing possible. In the case of alumina, two thirds of the oxygen spaces are filled and an aluminum atom occupies the space between six oxygen atoms. In the case of beryllia, the beryllium atom also lies between six oxygen atoms. The close correspondence between the distance between layers is to be noted. Magnesia, on the other hand, crystallizes in a face-centered cubic structure, as do the other alkaline earths.

Before discussing further the behavior of specific insulating materials suitable for use in tubes, it might be well to examine what is known about conductivity in crystals in general. Joffe[6] found widely divergent values for the specific conductivity of different crystals of such diverse materials as optically pure quartz, calcite, and chemically pure alums. However, by repeated purification of the ammonium alum, he was able to arrive at a fairly constant and much lower value of conductivity, differing from that of the material with which he had started by a factor of as high as 1:100. Conversely, by contaminating potassium alum with ammonium alum, he was able to raise the conductivity enormously. When potential was applied to the crystal, the conductivity decreased at a very rapid rate until after some hours it approached that of the pure material, despite the fact that the current carried could not account for the complete removal of more than a small portion of the impurity. However, there was a detectable and appreciably higher concentration of ammonia on the cathode side of the crystal than on the anode side. Similarly, optically pure quartz, on electrolysis, gave a deposit of sodium which did not belong in the lattice structure, and only at higher temperatures could silicon and oxygen be produced. It is at once evident, then, that current can be carried in a crystal by the migration of material as ions, in other words, by electrolysis. In the case of substances like rock salt and calcite, more than 99% of the current is carried by the positive ion.

But conductivity in the crystal can also occur electronically. For instance, the conductivity of rock salt is ordinarily the same in both light and darkness. However, after it is subjected to the action of X-rays or cathode rays, it becomes discolored, and then if it is exposed to sunlight, its conductivity increases so that it is 10^6 times that in darkness. This effect is not peculiar to rock salt but is characteristic of a large number of materials. The conductivity is not decreased by reduction of the temperature to that of liquid air. With a given intensity of light, the current through the crystal can be deflected by a powerful magnetic field in the same way as a photo-electron beam, and is proportional to the voltage applied, up to field intensities of the order of 10^5 volts per cm. Similarly, the behavior of oxide cathodes indicates electronic conductivity for thermionic current through the coating from the base metal to the surface of the coating. Without a doubt, similar conditions prevail in the insulating coating on a heater.

The fact that thermionic emission and electrolytic conduction as functions of temperature both follow laws which are exponential and quite similar has made it difficult to distinguish between the two methods by which conductivity can take place. Some help in distinguishing between them may be obtained by an examination of the curves in Fig. 2. Fig. 2 shows the leakage current between heater and cathode in two different tubes

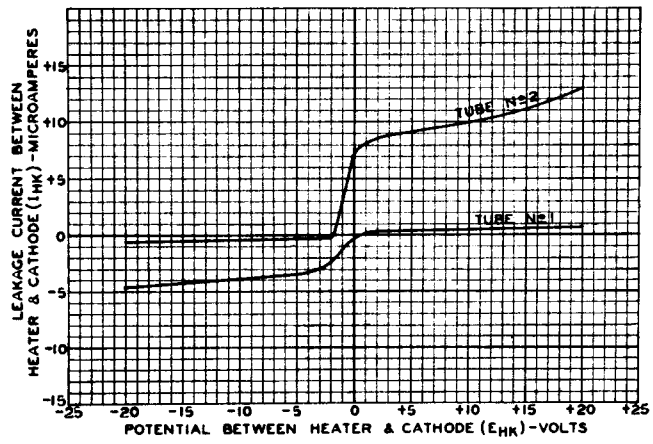


Fig. 2

when the heater potential is made positive and negative with respect to the cathode. The cathode is the point of reference. One of the two tubes which differ quite widely in type has high negative leakage and the other has high positive leakage. It will be noted that as the voltage goes through zero from positive to negative or vice versa, the greatest and most rapid change in the leakage current takes place. The rate at which the change occurs, however, is controlled somewhat by the maximum value of leakage attained. This would lead to the conclusion that the conduction is, at least in major part, electronic in nature, but the fact that the current does not saturate on the high-current side, but keeps rising slowly with an increase in voltage, indicates that either electrolytic or semi-conductor effects are also present. Some tubes have been seen in which the current-voltage relation substantially follows Ohm's law, and in these cases it is thought that conduction is electrolytic. When the temperature of the heater is raised sufficiently with a potential applied to the cathode, a point is reached at which occurs a breakdown that has all the appearance of electrolytic failure. Fig. 3 is of interest in that it shows how the leakage current is affected by temperature as measured in terms of the power input. It will be seen that in each case there is a straight-line relation on the power-emission chart exactly like the straight-line relation for thermionic

[6] A. F. Joffe, "The Physics of Crystals," McGraw-Hill Book Co., Inc., p. 81.

emission. This chart, unfortunately, does not permit distinguishing between electrolytic and electronic conduction because of the similarity in their temperature relationships.

It is obvious from the shape of the curves in Fig. 2, which show typical changes in leakage with changing heater-cathode voltages, that when the applied potential is alternating, very large swings in leakage current will occur as the voltage passes through zero. The impedance in this region will be low. On the other hand, if there is a fixed bias at some point on a nearly horizontal part of the curve, small applied a-c potentials will cause only small changes in current and the impedance will be high. That this

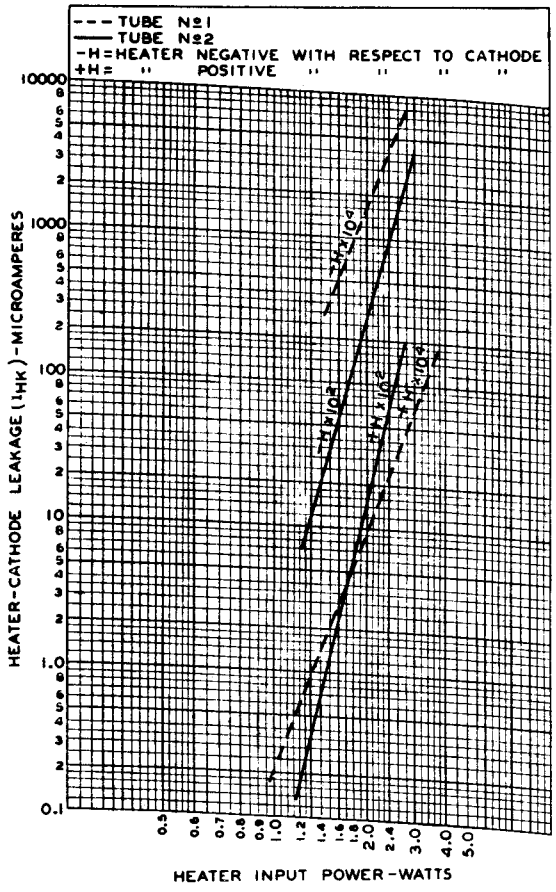


Fig. 3

is so can be seen from a typical curve for heater-cathode impedance in relation to the fixed bias applied between heater and cathode, as shown in Fig. 4. It is obvious that the absolute value of the impedance as measured in such a test will depend on:

1. The amount of a-c voltage applied. In general, the smaller the potential, the lower will be the impedance, since there will be a greater chance for all the potential swing to be most effective in the region where the slope of the current-voltage current is the greatest.

2. The frequency of the applied signal. Effects similar to those of polarization occur in the heater when d-c potential is applied to it just the same as occur in glass and all electrolytic conductors. Only by the use of an alternating potential can these effects be minimized. Since the net effect of any polarization is to raise the impedance, it follows that the higher the frequency, the lower will be the impedance as shown in Fig. 4.

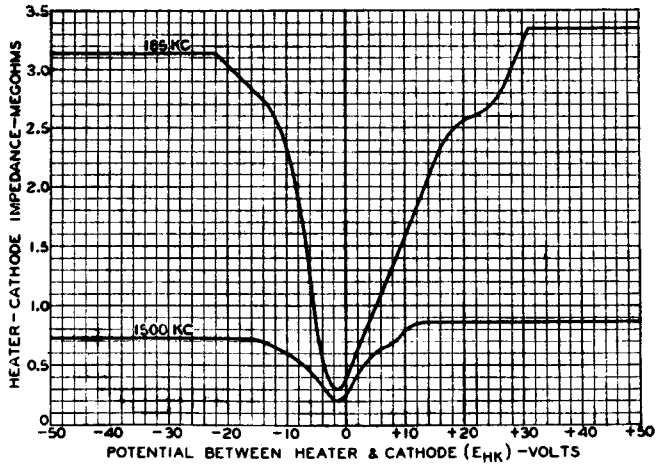


Fig. 4

It is evident that there will also be some relation between the a-c impedance which exists between heater and cathode, and the hum which will occur in a circuit. The actual amount of hum will be dependent on the conditions assumed. A curve by Bucklin (Fig. 5) illustrates a typical case.

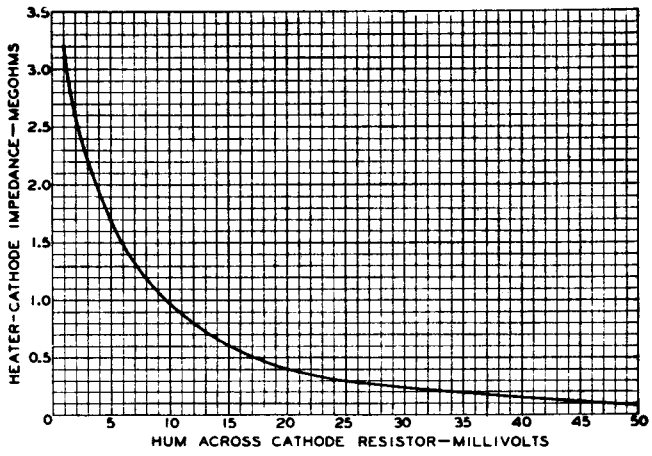


Fig. 5

Returning now to the discussion of materials suitable for heater insulation, we have already seen that alumina offers possibilities of satisfactory service. The source of alumina commercially is almost always bauxite ore. The ore is digested with sodium hydroxide under pressure or fused with a suitable sodium compound to form

sodium aluminate which is soluble in water and can be separated by filtration from impurities such as iron and silicon which are rendered insoluble by this treatment. By seeding the cooled filtrate properly, it is possible to precipitate hydrated alumina and separate it by filtration. After washing, the product becomes the base material for the aluminum industry as well as for our insulation. Due to the fact that commercial aluminum is prepared from a bath of sodium aluminum fluoride by electrolysis, not much attention need be paid to the sodium content. For our purpose, however, the alkali content is very important, and it is necessary to make further efforts to remove it. This can be done by acid washing of the alumina, which has been fired to a temperature of about 900°C to convert it partially to α alumina and render it insoluble. In this way, the alkali content, calculated as Na_2O, can be reduced to less than 0.2%. Another method effective in the removal of the alkali consists in fusing the material in an electric-arc furnace, wholly converting it to α alumina, or corundum, and volatilizing much of the alkali by virtue of the high temperature attained. The fused mass, however, is very hard (9 on Moh's scale) and presents great difficulty in grinding. For our work, grinding is done in an iron mill, and the ground material is treated with acid to eliminate the iron introduced. The acid washing also serves to eliminate further any alkali left. Neither of the two processes does much towards a reduction in the silica content.

When the materials are received, the acid-washed alumina is designated as "acid-washed bauxite-ore concentrates," and the acid-washed fused material is known as "alundum." The bauxite-ore concentrates require further treatment in the factory before they are ready for use, while the alundum does not. The precipitated alumina requires calcining at 1600°C to complete the conversion of the material to α alumina and thereby reduce the firing shrinkage, and subsequent grinding to attain the desired particle size. Milling is carried out in rubber-lined mills with either mullite or flint balls for grinding in order to keep to a minimum the impurities introduced at this operation.

Alumina is mixed with a spraying binder containing nitrocellulose by ball-milling the materials together in a porcelain ball mill. The alundum does not require ball-milling with the binder, and can be mixed by rolling the two together in a bottle. This difference in mixing procedure is necessary because of the differences in particle shape of alundum and alumina. The alundum being a crushed, fused product has angular faces on the particles while the calcined, precipitated material has particles approaching spheroids. These round grains aggregate readily and need a milling or crushing action to separate them so that each grain will be wet by the binder to give a smooth spraying mixture.

There are, in general, four methods of insulating a heater from the cathode: (1) the use of an extruded sleeve between the heater wire and the cathode; (2) application of insulating material to the heater by dipping the heater wire in a suspension of the insulating mixture; (3) spraying a mixture similar to a dip mixture on the heater; and (4) application of insulating material by electrodeposition or cataphoresis.

The extruded sleeve method is nearly obsolete. The chief disadvantages of insulating a heater by this method were slow heating time and excessively high operating temperatures of the heater wire.

The method of coating by dipping has several disadvantages when applied to helical heaters but may be used to advantage for folded heaters. In the latter case, the wire is coated by the drag-coating process (which is a modification of the older dipping process) before being formed into heaters. The folded type of heater shows some cost advantages.

At the present time the great bulk of our heaters are of the double-helical type using a sprayed-insulation coating. The cost of manufacturing such a heating element is high compared with other types but this type has electrical and magnetic advantages.

The last method of applying coating to heater wires is by cataphoresis. In this method, the suspended particles of insulating material are applied to the wire by making the heater an electrode in the suspension and using a sufficient potential of the proper polarity to build up the desired thickness of coating on the wire. At the present time this method is still in development, and little can be said about it except that it holds some promise.

Regardless of the method of application, the type of coating is dependent on: (1) the fineness of grain of the insulating material; (2) the shape of the grain; (3) freedom from aggregates of the material in the binder; (4) the viscosity of the coating mixture; and (5) the speed of application.

Spray mixes should be made from material of fine grain size (90% or more of the particles should be smaller than 4 microns and no particles greater than 30 microns), free of aggregates, and of a viscosity sufficient to keep the particles suspended. Rapid, wet spraying results in a smooth, dense coating which has a tendency to crack during firing. Slow, dry spraying results in a fluffy, granular coating which is loosely bonded to the wire and tends to chip when the heater is inserted in the cathode. The fine grain is necessary to insure a compact, relatively smooth coating without spraying "wet", since the particles strike the wire at random, and fine grains give a more compact coating with greater mechanical strength than do larger particles.

Drag-coating mixtures can be made with material of larger grain size than that used for spray mixes and still yield a smoothly coated

wire. The use of larger grain sizes for drag coating has another advantage in that it results in a coating which does not crack during firing. The reason for this is that fewer particles and consequently fewer interfaces are required for coating a given length of wire. Since the interfaces are filled with a film of the suspending liquid, there is less liquid to remove and consequently less linear shrinkage in the case of larger particles. Fine grains can be used for spraying, because a large portion of the suspending liquid is removed by evaporation between the spray gun and the filament.

The shape of the grain is important in both spraying and drag coating, since fragmental and rough-shaped particles pack with a larger number of interface contact points than do spherical particles. As a result, greater mechanical strength of coating after firing is obtained. This is due to the fact that the bond between particles is a function of incipient fusion which takes place between particles at the points of contact—the smaller the point of contact the lower the temperature at which the fusion will take place.

If the insulating materials have been correctly processed and applied, the heaters need only be fired at a temperature sufficiently high and for a time long enough to sinter the grains together and to the wire. Longer firing periods or temperatures higher than this are detrimental to the coating. They give a dense, glassy matrix which results in high heater-cathode leakage and hum, or both.

The presence of impurities in the aluminum oxide has an important effect upon the characteristics of the resulting insulation. If the alumina were truly chemically pure, a temperature of 2050°C would be required to melt it; and approximately the same temperature would be required to sinter the material so that it would remain on a heater. Even as short a time as five years ago, when 1600°C was about the limit to be attained in hydrogen firing furnaces, it was necessary to add silica to the purest grades of alumina or alundum in order to obtain a product with adequate hardness to permit handling. In some tests carried out at that time, the effect of the addition of silica on the heater-cathode leakage was investigated. The results are shown in Fig. 6.[7] The material used as a base was alundum containing 0.7% SiO_2 and 0.3% Na_2O. It was inferior to the grade of product available today. Other lots contained up to three times as much alkali and five times as much silica as this sample.

On the other hand, the use of addition materials has advantages when they are properly selected. Assuming that the presence of alkali in the alumina is the most effective cause for the reduction of its insulating properties, it is most certainly present in the form of an aluminate of some type. Assuming that it is present in the form of metaaluminate, according to the classical Arrhenius theory, there will then be present sodium ions and metaaluminate ions in equal quantities, so that the whole will be electrically neutral. Under the influence of a potential gradient, the sodium ion will tend to migrate to the negative electrode or cathode and the aluminate ion to the anode. From the studies on the conductivity of glass, it is likely that the major portion of the current will be carried by the more mobile sodium ion, and a comparatively small portion by the complex, large,

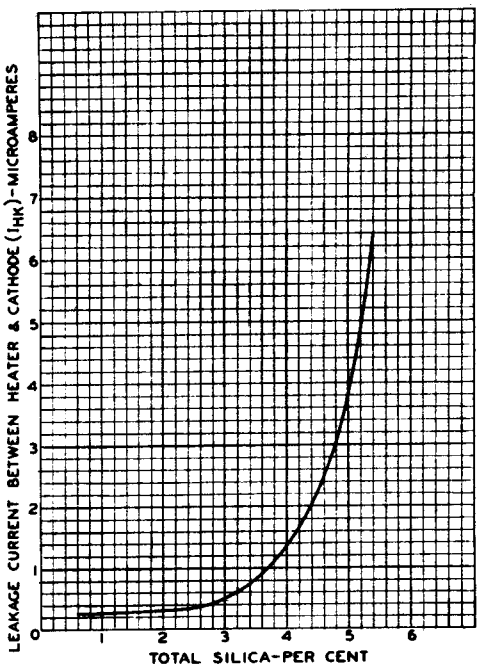


Fig. 6

metaaluminate ion. If the sodium ion migrates to the negative electrode, on reaching there it loses its charge and becomes a neutral atom. If the temperature is sufficiently high, it is volatilized, but if it is not, it remains and then becomes a very effective emitter of electrons. According to Arrhenius' theory, it should be possible to hinder this whole process to a considerable degree since the ionization is an equilibrium process:

$$Na^+ + AlO_2^- \rightleftharpoons NaAlO_2$$

While for simplicity, the phenomena have been discussed from the standpoint of the classical Arrhenius theory, they can be equally well explained by the use of the Debye-Hückel theory or the more recent theory involving ion "holes"

[7] Ref. 3-7.

as discussed by Schottky[8] and deBoer.[9] By increasing the concentration of the aluminate ion, the reaction can be driven to the right, thus reducing the concentration of the sodium ions since the product of the two concentrations is a constant for any temperature. This can be accomplished by adding metaaluminates in which the positive ion is one which does not migrate with the facility of the sodium ion, e.g., the alkaline earth metals. A series of experiments in which tests were made on insulation of alumina and of the same alumina with the addition of 1% of the various alkaline-earth oxides very well illustrates this behavior as shown in Figs. 7 and 8. Here the impedance between heater and cathode at various values of fixed bias was measured when a signal of 0.5 volt was impressed at a frequency of 100000 c.p.s. It was found that the addition of each of these materials was effective in reducing the conductivity of the untreated alumina to various degrees. A large amount of practical experience indicates that a barium-oxide addition is the most effective of those things tried. The strontium oxide used for the test was of commercial grade, and it would be expected that a purer grade would lead to superior results. It was found from a study of the curves that whereas the tops are flat in the case of alumina even at high temperature and high bias, in each of the other cases there is a tendency for the insulation to decrease under these conditions. This can be explained by the fact that at high temperatures the larger barium ions have a greater mobility and can move under the influence of the higher potentials. That this is so is borne out by the fact that intermittent life tests at 8.5 volts on 6.3-volt heaters with 180 volts bias between heater and cathode show more failures in the cases where additions have been made than in the case of alumina alone. In the case of 6.3-volt life tests, there appears to be no similar difference. Work by Benjamin[10] has indicated that beryllia additions are even more effective than baria, but experiments in this laboratory have not been conclusive on this point.

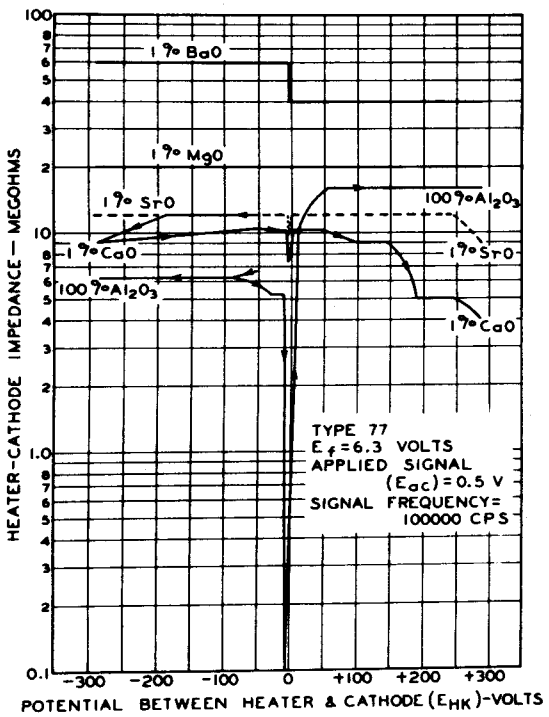


Fig. 7

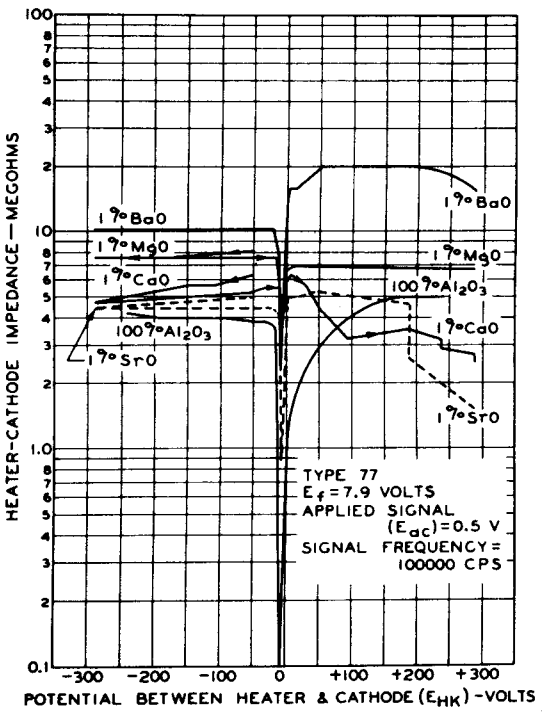


Fig. 8

From some studies of the use of electrolytic as compared with metallurgical tubing, there is available very good evidence of how the conditions of use affect the insulation between heater and cathode. For reasons connected with tube characteristics, it was found to be more desirable to use electrolytic tubing for the cathode sleeve than to use the metallurgical nickel standard at that time. It was found that the operating temperature was substantially higher

[8] Schottky, Naturwissenschaften, 23, 656; 1935. Z. physik. Chem. (b) 29, 335; 1935.

[9] DeBoer, Recueil Des Travaux Climiques Des. Pays-Bas, 56, 301-390; March, 1937.

[10] Private communication to author.

with the electrolytic tubing and a constant input wattage, and consequently it was necessary for the heating element to function at a correspondingly higher temperature to attain the necessary heat transfer and the proper emission. This is illustrated in Fig. 9.[11] From this, it can be seen that there is a difference of about 50°K between the two cathodes and this has the pronounced effect on the heater-cathode leakage as shown in Fig. 10. It will be seen that at the operating point, the temperature difference accounts for a change in the leakage by a factor of about 5:1. At the same time, the difference in temperature of the heaters as operated is shown by a small but consistent difference in the heater current, indicating a higher resistance and a correspondingly higher temperature for the one enclosed by electrolytic nickel.

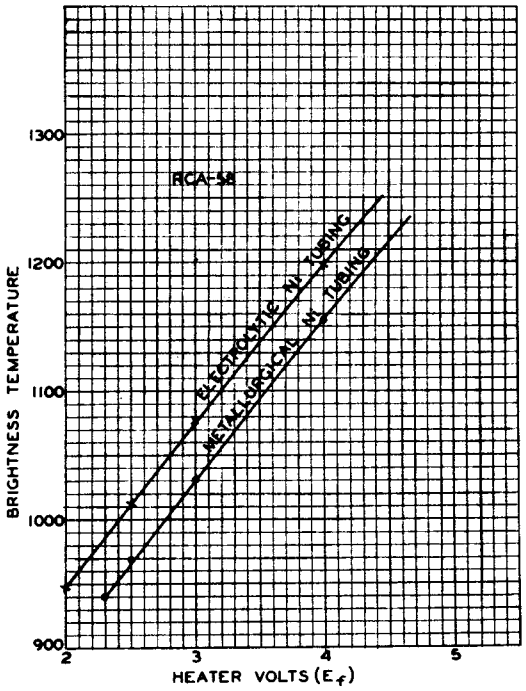


Fig. 9

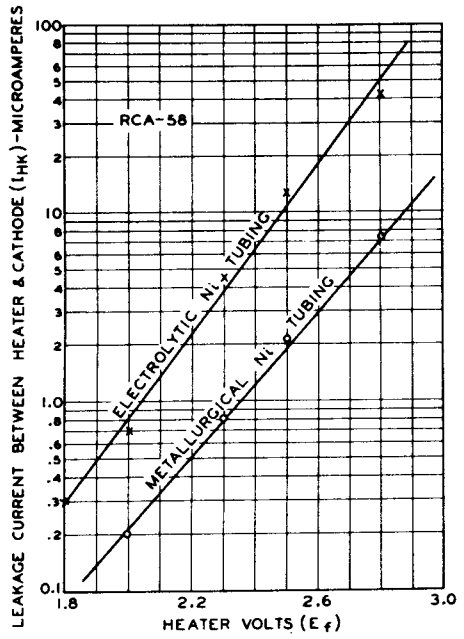


Fig. 10

[11] Ref. 3-11.

Lecture 4

PHOTOELECTRIC AND SECONDARY EMISSION

L. B. Headrick

Part I - Photoelectric Emission

INTRODUCTION

The discussion will be limited to the external photoelectric effect or the emission of electrons from the surface of a conductor or semiconductor upon the absorption of visible and near-visible light by the first few hundred atomic layers near the surface. Characteristics of clean metals, thin films of one metal on another metal, metals with a thin oxide layer on the surface, and composite surface layers of a metal-metal oxide semiconductor will be considered. Photoconductivity in crystals and the barrier-layer photoelectric phenomena, such as selenium and copper-oxide cells, will not be considered, except with respect to one item, namely, the efficiency of light-sensitive devices.

Einstein in 1905 made the first theoretical advance in an attempt toward an explanation of photoelectric phenomena when he postulated that light must be adsorbed, as well as emitted, in quanta of energy $h\nu$, and if one quantum is adsorbed by one electron, the following relation must hold:

$$h\nu - \varepsilon_o = 1/2\ mv^2_{max} = eV_{max}$$

$$h\nu = eV_{max} + \varepsilon_o$$

where

- ν = frequency of incident light = 1/wavelength λ
- h = Planck's constant
- ε_o = minimum energy required to remove an electron from the metal to free space or vacuum
- m = electron mass
- v_{max} = velocity of the fastest emitted electrons
- V_{max} = the retarding potential required to stop the fastest electrons

The frequency of light which will just eject electrons with zero velocity is called the threshold frequency (ν_o) and is defined by

$$h\nu_o = \varepsilon_o$$

The corresponding threshold wavelength or long wavelength limit (λ_o) may be defined by

$$h/\lambda_o = \varepsilon_o$$

After Einstein, the next important theoretical advance in the field of photoelectric emission was made by Sommerfeld in 1928 with the introduction of the electron theory of metals. This theory failed to explain all of the observed photoelectric phenomena because the electrons and atoms were assumed to be in thermal equilibrium. Later Sommerfeld applied the Fermi-Dirac statistics to electrons in metals where the electron energy distribution extends over a range from zero to a definite maximum value at 0°K. Since photoelectric emission involves the interaction between a quantum of light and the electrons near the surface of a metal, it is essential to have a satisfactory theory of electrons in metals in order to have a sound basis for a theory of photoelectric emission. The Fermi-Dirac statistics applied to the electron theory of metals has been used quite successfully to explain a number of the observed characteristics of photoelectric emission with no important discrepancies and, therefore, appears to satisfy the above requirement.

A. PHOTOELECTRIC EMISSION FROM CLEAN METALS AND THIN METAL FILMS

1. Energy Distribution of Photoelectrons

The total energy distribution characteristic of photoelectrons for different temperatures is shown in the upper part of Fig. 1 by the solid and dashed curves, as calculated from the Fermi-Dirac theory. These curves show that only at a temperature of 0°K is there a definite, sharp limiting value of electron energy in a metal and that at higher temperatures there are increasing numbers of electrons with higher energies as the temperature increases. Therefore, the theory indicates that the energy distribution of electrons emitted by monochromatic light of a given wavelength will only have a sharp limit at 0°K and will tail off at higher temperatures. The theory agrees with observations. However, until very recently, experimental observations did not fit the calculated curves closely at the lower values of energy. The departure of the older experimental data is shown by the dotted curve. Using thin films of potassium, C. L. Henshaw[1] has

[1] C. L. Henshaw, "Normal Energy Distribution of Photoelectrons from Thin Potassium Films as a Function of Temperature," Physical Review, Vol. 52, No.8, p. 854; October 15, 1937.

recently shown by very careful experimental work that the theoretical curves can be obtained experimentally.

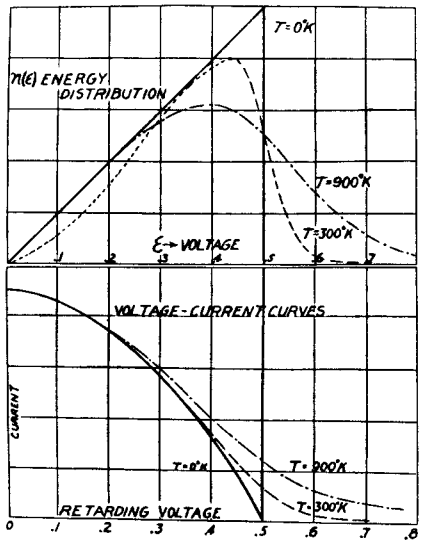


Fig. 1 - Theoretical total-energy distribution and current-voltage curve.
(L. A. DuBridge, see Bibliography)

The curves in the lower part of Fig. 1 represent the theoretical form of the retarding potential vs current curves which have now been verified by the experimental work of Henshaw. The retarding potential vs current curves are obtained directly from data taken to determine the energy distribution of emitted photoelectrons.

Nottingham has made two attempts and Mitchell one separate attempt to modify the theory to conform to the older experimental data. The results of their calculations are shown in Fig. 2, and are compared with the DuBridge-Fowler calculation based on the Fermi-Dirac theory of metals and the older experimental data shown by the dots. None of these calculations of Nottingham or Mitchell fit the experimental data at all well, probably due to the fact that both the modified theory and data are in error.

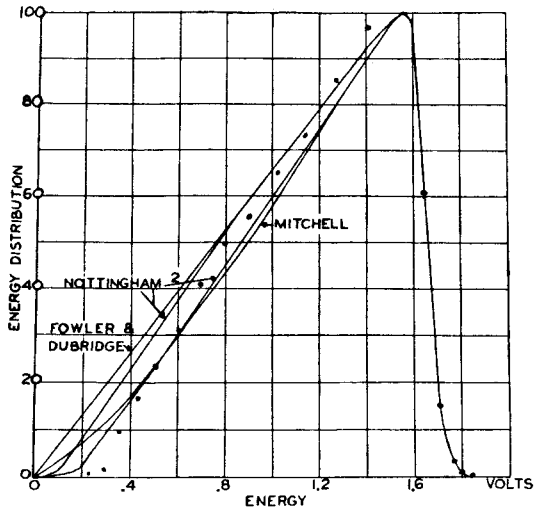


Fig. 2
(Israel Liben, Phys. Rev., Vol. 51, p. 647, 1937)

2. Spectral Distribution of Photoelectrons

The curve in Fig. 3 shows the form of a typical spectral distribution curve which approaches the frequency axis asymptotically, and an extrapolation of the curve to determine the apparent threshold. DuBridge and Fowler have worked out a method of extrapolation of spectral distribution

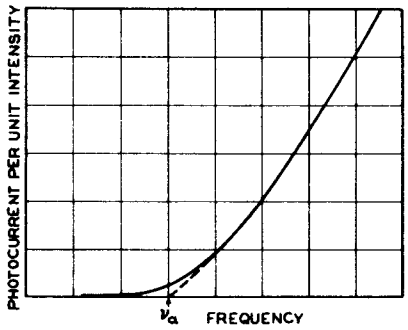


Fig. 3 - Typical spectral distribution curve, showing extrapolation to determine ν_a (exaggerated).
(L. A. DuBridge, see Bibliography)

curves based upon Fowler's theory of spectral distribution which give values of the threshold frequency and corresponding values of the surface work function that agree well with thermionic data.

Ives and Briggs have postulated that photoemission should follow the rate of absorption of energy at the surface of the metal and have, therefore, calculated the rate of absorption of energy at the surface of thin films of potassium on platinum. Fig. 4 shows the results of their

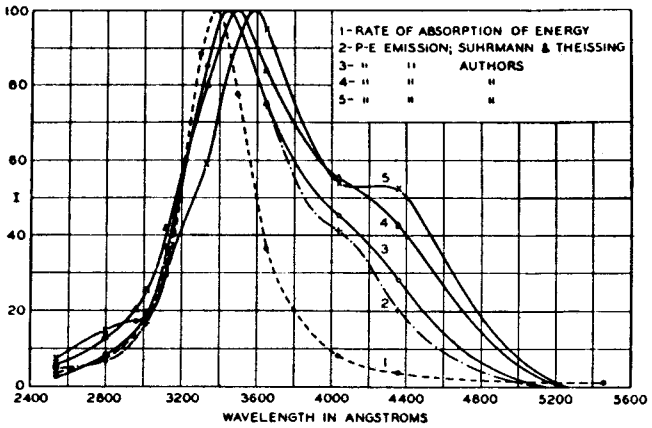


Fig. 4 - Calculated and observed photoelectric emission from thin film of potassium on platinum.
(Ives & Briggs, J.O.S.A., Vol. 26, p. 249, 1936).

calculations compared with their photoelectric data as well as with the data of Suhrmann and Theissing. The agreement between theory and experiment is exceptionally good on the short wavelength side but is not at all good on the long wavelength side of the spectral distribution curve. The lack of correlation on the long wave side may be due to the effect of surface gas adsorption on the spectral distribution characteristic.

3. Threshold Frequency or Wavelength and Work Function

The relation between threshold wavelength λ_o in Angstrom units and the work function in volts at 0°K is given by $\theta = 12340/\lambda_o$. For tungsten, $\lambda_o = 2720$ A°. Therefore, $\theta = 12340/2720 = 4.54$ volts.

Fig. 5 shows the variation of the photoelectric work function throughout the periodic table of elements. It will be noted that the alkali and alkaline-earth metals have the lowest values of work function, as is well known from thermionic data. Since low work function means long wavelength response, the alkali metals are used in phototubes having high sensitivity to visible light.

The photoelectric threshold and work function of typical surfaces of different types are shown in Table I. With the exception of the data for the clean metals, the values of threshold and work function are very rough experimental values.

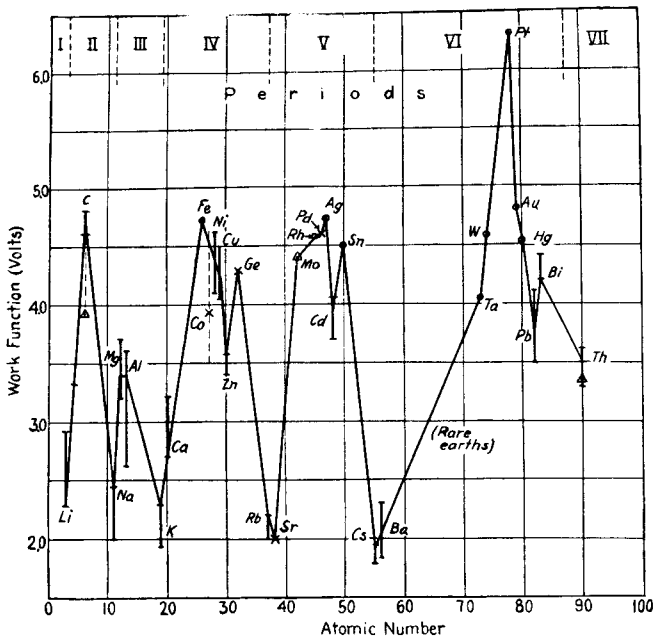


Fig. 5 – Variation of photoelectric work function throughout the periodic table. ⊙, values for clean surfaces. ×, single values of uncertain precision. △, reliable thermionic value (indicated only when photoelectric value is uncertain. I, vertical lines indicate approximate range covered by various reported values.
(Hughes & Dubridge, see Bibliography)
Courtesy of McGraw-Hill Book Co., Inc.

Table I

TYPE OF SURFACE	TEMPERATURE °K	THRESHOLD WAVELENGTH A°	WORK FUNCTION θ Photoelectric (Volts)	WORK FUNCTION θ Thermionic (Volts)
Clean Metals—				
Ta	293	3010	4.13	4.13
Mo	303	2992	4.14	4.15
W	295	2720	4.54	4.54
Pd	305	2490	4.96	4.99
Thin Metal Films—				
Na on Pt		5900	2.08	
K on Pt		7700	1.60	
Rb on Pt		7950	1.56	
Cs on Pt		8900	1.38	
Th on W		4900	2.52	
Ba on Ag		7900	1.56	
Metals and Metal Oxides—				
Th		3600	3.43	
Oxidized Th		4100	3.01	
Ur		3400	3.63	
Oxidized Ur		3800	3.25	
Ca		4000	3.09	
Oxidized Ca		5200	2.37	
Ba		5000	2.47	
Oxidized Ba		6900	1.79	
Composite Oxide Surfaces—				
Ag: Na_2O, Ag - Na		6800	1.85	
Ag: K_2O, Ag - K		8000	1.55	
Ag: Rb_2O, Ag - Rb		10000	1.20	
Ag: Cs_2O, Ag - Cs		12000	1.00	

4. Space Distribution of Photoelectrons

The curve in Fig. 6 shows the space distribution of photoelectrons from rubidium with polarized light. There is a slight difference between the two curves, one for parallel (∥) and the other for perpendicular (⊥) polarized light, but the fact that both curves follow close to a cosine distribution when the light is incident at 70° from the normal shows that there is no tendency for the emitted electrons to follow the electric vector of the incident light.

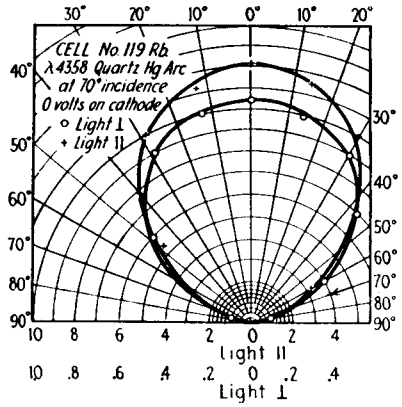


Fig. 6 - Symmetrical angular distribution of photoelectrons from Rb, showing no tendency to follow the electric vector. Arrow shows direction of incident light.
(Hughes & DuBridge, see Bibliography)
Courtesy of McGraw-Hill Book Co., Inc.

5. Relative Spectral Response of the Alkali Metals

The spectral distribution curves for four of the alkali metals, Na, K, Rb, and Cs, are compared with the spectral response curve of the human eye in Fig. 7. All of the curves are drawn with their peaks at the same level to bring out the uniform shift toward longer wavelengths and the similarity of the form of the curves as we go from Na to Cs. It was found by Miss Seiler[2] that

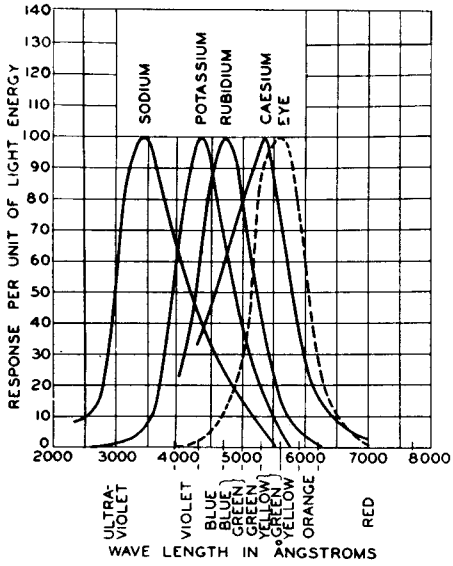


Fig. 7 - Photoelectric cells using the pure alkali metals respond differently to light of various colors.
(A. R. Olpin, Bell Labs. Record, Vol. 9, p. 311, 1931)

[2] E. C. Seiler, "Color Sensitivity of Photoelectric Cells," Astrophys. Jour., Vol. 52, p. 129; Oct. 1920.

for the alkali metals, the peak broadened and shifted toward longer wavelengths, and the maximum sensitivity decreased, with an increase in atomic weight from Li to Cs.

6. Variation of Photoelectron Emission with Crystal Planes

In Fig. 8 are shown electron image pictures of a barium-covered nickel surface. Picture (a) shows the non-uniform distribution of photoelectrons from a polished surface before heating. Picture (b) is the image of thermionic electrons after the surface has been etched by evaporation. In the center portion the crystal structure shows up clearly. Picture (c) is the image of photoelectrons after the surface was heated to pro-

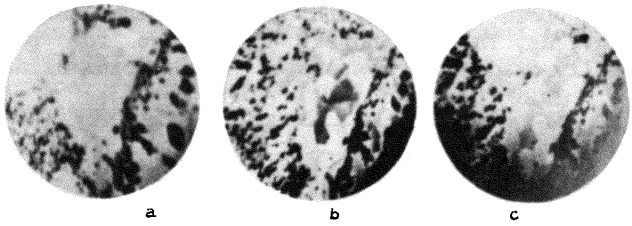


Fig. 8 - Electron emission pictures of a barium coated nickel surface. (a) Photoelectric before heating. (b) Thermionic. (c) Photoelectric after heating.
(Gross & Seitz, Zeits. für Phys., Vol. 105, p. 735, 1937)

duce the thermionic-emission picture. Here the crystal structure also shows up well but with not quite the contrast shown by the thermionic-emission picture. It is of interest to note the good correlation between the areas showing high emission for both photoelectric and thermionic emission. The other factor of interest is that different crystal faces have different values of work function even when coated with a layer of another metal.

7. The Selective Photoelectric Effect

The selective photoelectric effect is so named because of the form of the spectral distribution curve, which has a rather sharp selective maximum usually at a wavelength of several hundred Angstrom units below the threshold. It has been found experimentally that this selective maximum observed for thin films and composite smooth surfaces with unpolarized light is due entirely to the component of the light polarized with its electric vector parallel to the plane of incidence or perpendicular to the plane of the photosurface. When light polarized with its electric vector perpendicular to the plane of incidence, or parallel to the plane of the photosurface, is incident on a smooth surface, no selective maximum is found in the spectral distribution curve. Rough surfaces show the selective

effect with light polarized in any plane, because there is always part of the surface perpendicular to the electric vector. K. Mitchell[3] gives a theoretical interpretation of the phenomena of selective photoelectric emission.

B. PHOTOELECTRIC EMISSION FROM THIN OXIDE FILMS AND COMPOSITE OXIDE SURFACES

1. Change in Spectral Distribution by Oxidation

The curves in Fig. 9 show the effect of oxidizing the surface of calcium. The curve for the oxide surface has a higher peak shifted toward longer wavelengths with respect to the curve for the pure metal. The oxidation of the surface also extends the long wavelength threshold by more than a thousand Angstroms. Similar effects have been observed with several other metals, e.g., Mg, Th, Al, and Ba. The curve for partially oxidized Ba evaporated from a barium oxide-coated cathode is compared with that for pure Ba in Fig. 10. The relative sensitivities of Ba and Ba + BaO are not similar to the case of Ca and an oxidized calcium surface, probably because the surface is of different structure, being formed by entirely different methods. However, the long wavelength threshold is shifted by several hundred Angstroms by the partial oxidation.

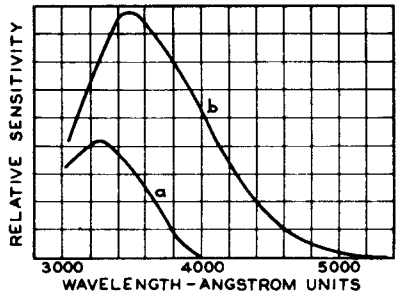


Fig. 9 - (a) Sensitivity curve for a calcium cell in Corex D glass. (b) Sensitivity curve for the same cell treated with the oxygen from a tube at 40 mm pressure.
(Rentschler & Henry, J.O.S.A., Vol. 26, p. 31, 1936)

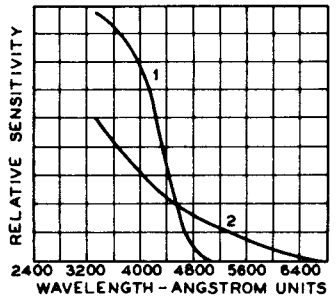


Fig. 10 - (1) Sensitivity curve for barium metal in Corex D glass. (2) Sensitivity curve for the vaporized deposit from a barium oxide coated cathode.
(Rentschler & Henry, J.O.S.A., Vol. 26, p. 31, 1936)

2. Quantum Efficiency of Light-Sensitive Devices

The curves of Fig. 11 show that the caesium-oxide type of phototube has an extremely low energy efficiency in terms of amperes per watt or in percentage of theoretical maximum yield of electrons per quantum of light as compared with other light-sensitive devices. The two Photox cells for which data are represented in Fig. 11 have sensitivities of 200 and 150 μa/lumen while the selenium cells have sensitivities of 450 and 120 μa/lumen, respectively. Commercial selenium cells can now be obtained with a sensitivity of 480 μa/lumen and caesium phototubes with a sensitivity of 35 to 55 μa/lumen. But even the best experimental caesium-oxide phototube with a sensitivity reported to be about 100 μa/lumen or quantum efficiency of the order of 3% is far below the theoretical maximum. For commercial phototubes the quantum efficiency ranges from about 0.01% to 1.0%. The lower values of quantum efficiency are obtained with the pure metals such as Ca, Mg, Cd, Th, and their oxides. These surfaces are used for measurements of ultra-violet radiation. The higher values of quantum efficiency are obtained with the composite surfaces, such as Ag: Cs_2O, Ag-Cs or Ag: Rb_2O, Ag-Rb which are used in the visible and infra-red region. This extremely low efficiency for the external photoelectric effect is due mainly to the low probability of reaction between a quantum of light and the bound surface electrons. There is practically no reaction between light quanta and the conduction electrons, or nearly free electrons, in a metal. Therefore, many quanta of light pass through the surface layer of electrons of a metal without ejecting an electron and lose their energy to lower-lying bound electrons associated with the atoms by exciting the electrons to higher energy states. The energy of the photon is not sufficient to eject an electron from these bound states in the metal. However, the case of secondary emission is entirely different because the primary electron is equivalent to a quantum of much higher energy and, therefore, may eject several low-lying bound electrons.

This theory also points out why the quantum efficiency of photoconductive-type photocells, such as those with a barrier layer, copper oxide or selenium, may have much higher quantum efficiencies than phototubes which use the external photoelectric effect. The reason is because with photoconductivity we are dealing with semi-

[3] K. Mitchell, "The Theory of the Surface Photoelectric Effect in Metals," Proc. Royal Society, Sec. A, (Part I) Vol. 146, p. 442; 1934; (Part II) Vol. 153, p. 513; 1935.

conductors which have plenty of relatively high energy-bound electron states throughout the body of the crystal so that the excitation of an electron to an unfilled conduction level is the result of a volume reaction of high probability between light quanta and bound electrons. The caesium-oxide phototube has considerably higher quantum efficiency than clean metal surfaces because the Cs_2O-AgCs complex structure is of the nature of a semiconductor. Therefore, it is possible that electrons may be ejected from its surface by reaction with light quanta at different depths in the layer. But the layer is either too thick or has a value of conductivity too high to give an efficiency approaching that of the barrier layer photocell.

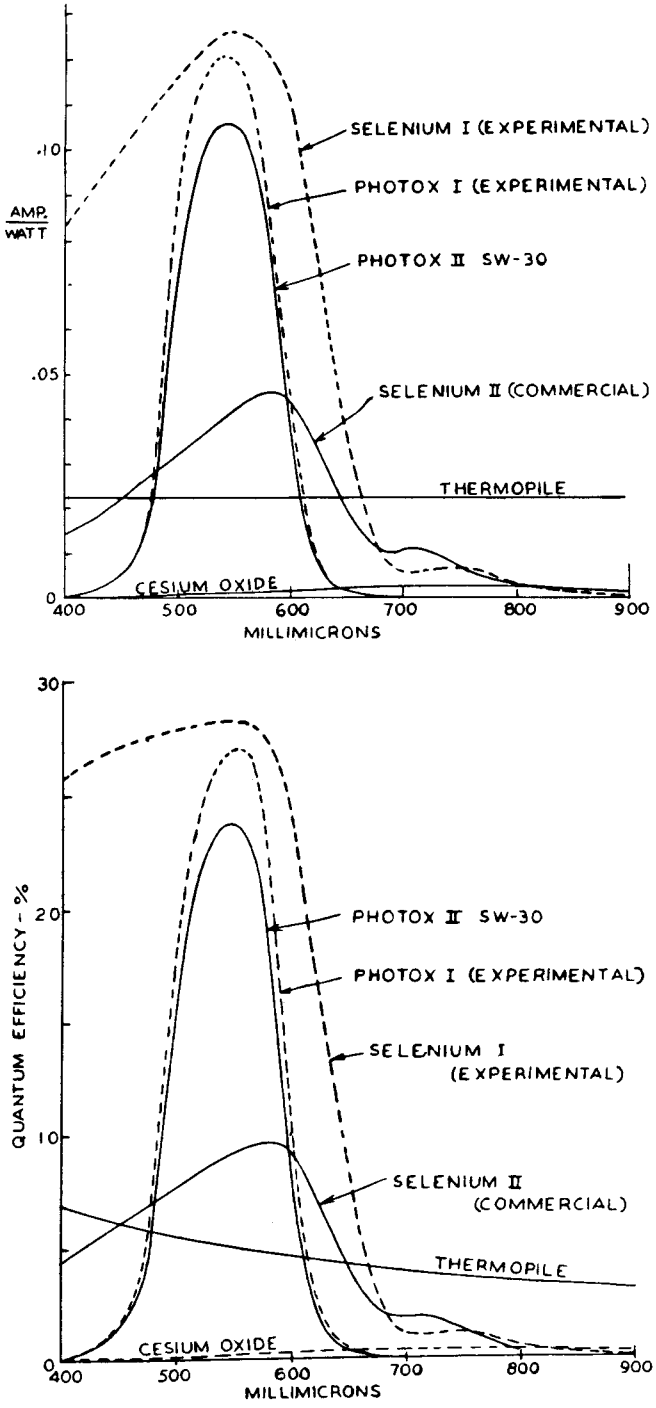


Fig. 11 - (top) Absolute spectral response curves for various light sensitive devices. (bottom) Quantum efficiency curves for various light sensitive devices.
(C. C. Hein, J.O.S.A., Vol. 25, p. 205, 1935)

3. Fatigue of Photosensitive Surfaces

High sensitivity photosurfaces of the type represented by Ag: Cs_2O, Ag-Cs are subject to fatigue or decrease of sensitivity on exposure to light. The curves in Fig. 12 show that the fatigue increases as the wavelength of the light decreases. Fatigue is assumed to be due to photochemical reaction which causes changes in the surface composition of these thick films of complex structure.

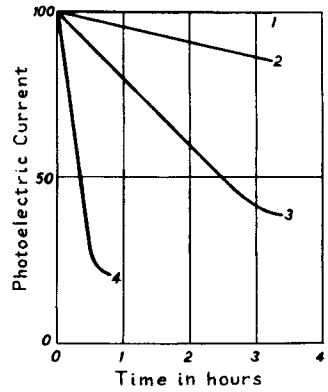


Fig. 12 - Decrease of current for a given Ag:Cs_2O, Ag-Cs cathode: 1, in infra-red light; 2, in red light; 3, in green light; 4, in blue light.
(J. H. DeBoer, see Bibliography)
Courtesy of The Macmillan Company

The curves in Fig. 13 show the change in the spectral distribution characteristic produced by the fatigue reaction. The result of fatigue is similar to increasing the amount of caesium on the surface, as will be shown later. Therefore, fatigue appears to be due to a photochemical reducing reaction whereby caesium is liberated from the caesium-oxygen structure just below the surface and diffuses to the surface.

4. Time Lag in Gas-Filled Phototubes

In order to increase the output of phototubes, it has been customary to introduce into the tube an inert gas such as argon to the order of 100 microns pressure. The primary photocurrent is

then amplified by ionization in the gas. The large mass of the ions produces a lag in the response to high frequencies.

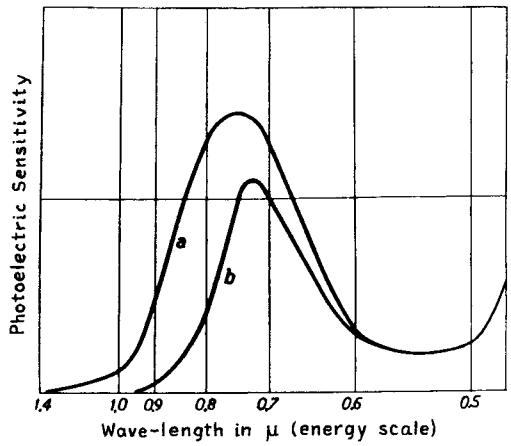


Fig. 13 - Spectral distribution curve of a Ag:Cs_2O, Ag-Cs cathode: (a) in a fresh state; (b) after fatigue has set in.
(J. H. DeBoer, see Bibliography)
Courtesy of The Macmillan Company

The curves in Fig. 14 were taken on a gas-filled phototube of the 868 type with a gas pressure slightly below normal after the tube had been exposed to illumination for 694 microseconds followed by a dark period of 7630 microseconds. The different curves are for different voltages on the tube with the steady-light value adjusted to give a photocurrent of 10 microamperes. At 20 volts the curve is indistinguishable from that

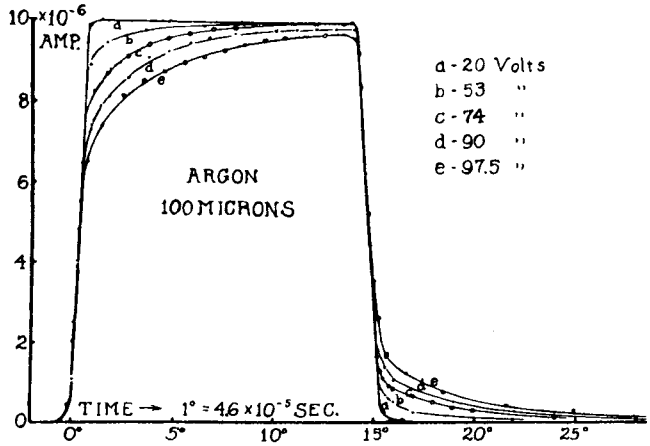


Fig. 14 - Time-lag plots for an argon cell at 100 microns of pressure.
(Huxford & Engstrom, Review of Scientific Instruments, Vol. 8, p. 389, 1937)

of a vacuum tube. The gas amplification ratio for the tube was about 4 at 90 volts. The fact that between 60% and 80% of the total current flows in less than 50 microseconds indicates that a large part of the gas amplification arises from the flow to the anode of electrons ejected from the atom, through primary ionization by the photoelectron. The slower-moving ions and metastable atoms produced by electron impact reach the cathode later and release more electrons which contribute to the main portion of the current which lags by 50 to 750 microseconds. The persistence of the current after the light beam is removed shows similar characteristics to those during the rise. The electrons, ions, and metastable atoms left in the space after the light is cut off continue to bombard the cathode for from a few to several hundred microseconds, and cause electron emission which accounts for the current flow dying off slowly after the light beam is cut off.

5. Energy Distribution of Photoelectrons at Different Wavelengths

The curves in Fig. 15 show the energy distribution of electrons from a thick film of potassium for the wavelengths 4350 A° and 3650 A°. Comparing the curves with those of Fig. 1, it can be seen that they approach the shape of the theoretical curves except for the slight curvature on the low-energy side of the maximum. As was pointed out earlier, this curvature can probably

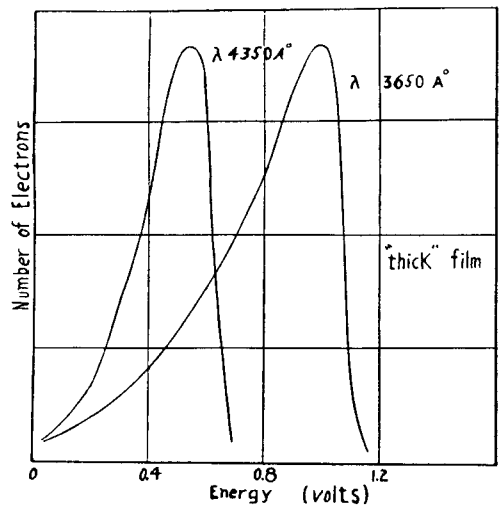


Fig. 15 - Energy distribution of photoelectrons from a 30 molecular layer film.
(J. J. Brady, Phys. Rev., Vol. 46, p. 772, 1934)

be attributed to experimental error, because of the difficulty of collecting low-velocity electrons. The shift in the peak toward higher-energy electrons for the shorter-wavelength light is to be expected, because the shorter-wavelength light quanta have higher values of energy.

The curves in Fig. 16 are similar to those in Fig. 15, except that the ones in Fig. 16 are for a caesium-oxide surface of the type Ag: Cs_2O, Ag-C_s and are all plotted so that the point at

which the number of high-energy electrons becomes very low is the same point for each curve. Thus, the horizontal axis of electron velocity may be considered to be in per cent of an arbitrary maximum velocity. The curves bring out the following interesting points. None of them approach the shape of the theoretical curve for a

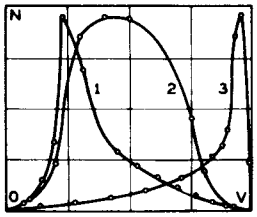


Fig. 16 – (1) $\lambda = 4358$ A°; (2) $\lambda = 5460.7$ A°; (3) infra-red.
(Pjatinitzki & Timofeew, Phys. Zeits. Sowjetunion, Vol. 9 p. 195, 1936)

clean metal surface as in Fig. 1 and the long-wavelength infra-red radiation gives the highest proportion of high-velocity electrons. The first fact shows that either the absorption characteristics or the excitation probability of complex films is quite different from that of clean metals. This last fact indicates that most of the electrons emitted by infra-red radiation come from the surface layer of atoms and, therefore, little energy is lost by the electrons in escape from the metal. This evidence is in support of DeBoer's picture of the infra-red peak in the caesium-oxide surface being due to a type of photoionization of the surface caesium atoms.

6. Spectral Distribution Characteristics of the Caesium-Oxide Type (Ag; Cs_2O, Ag-Cs) Photosurface

The spectral distribution curve for an 868-type caesium-oxide phototube is shown in Fig. 17.

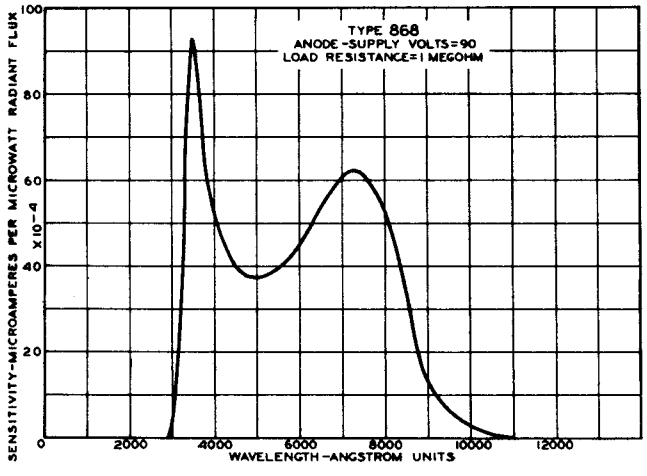
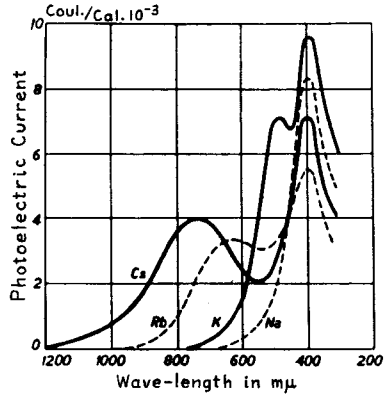


Fig. 17 – Spectral distribution curve for Type 868 caesium-oxide phototube.

It will be seen that the infra-red peak occurs at about 7300 A° and that the long-wavelength limit is reached at nearly 11000 A°.

The curves in Fig. 18 show the relative spectral sensitivities of the oxide surfaces of the alkalies Na, K, Rb, and Cs. From these curves it can be seen that Cs and Rb are the only alkali oxides with a long-wavelength maximum. It is also interesting to note that K and Na with no long-wavelength peak have higher peaks, near 4000 A°, than Cs or Rb.



Spectral distribution of Ag:Me_2O, Ag–Me cathodes (Me = Cs, Rb, K or Na).
(W. Kluge, Phys. Z., Vol. 34, p. 115, 1933)

Data relating to Ag:Me_2O, Ag–Me cathodes (W. Kluge, *Phys. Z.* 34, 125 (1933)).

Cathode	Short-wave max. mμ	Long-wave max. mμ	Long-wave limit mμ	Energy of the $h\nu$ of the long-wave limit electron volts
Ag:Na_2O, Ag–Na	410–420	*	~ 680	~1.85
Ag:K_2O, Ag–K	400–420	460–520	~ 800	~1.55
Ag:Rb_2O, Ag–Rb	400–420	620–680	~1000	~1.2
Ag:Cs_2O, Ag–Cs	410–430	730–800	~1200	~1.0

Fig. 18

Prescott and Kelly of Bell Telephone Laboratories did a considerable amount of research on the influence of the caesium-to-oxygen ratio on the characteristics of the caesium-oxide-type phototube. The importance of the caesium-to-oxygen ratio in determining the behavior of the phototube is well illustrated by the spectral distribution curves of Fig. 19. The curve No.1 is characteristic of a surface with a low caesium-oxygen ratio; the peak sensitivity occurs at about 7500 A° but is low; and the threshold wavelength is slightly beyond 11000 A°. Curve No.2 shows the effect of adding more caesium. The increase in caesium-to-oxygen ratio increases the peak sensitivity, moves the peak toward longer wavelengths to a maximum of about 8000 A°, and extents the threshold to about 12000 A°. Curve No.3 shows the effect of a still higher caesium-to-oxygen ratio. The peak sensitivity still in-

creases but the peak moves back to 7500 A°, and the threshold wavelength decreases to just beyond 10000 A°. The effect of a still higher caesium-to-oxygen ratio is shown by curve No.4 where the peak-sensitivity wavelength and the threshold wavelength are considerably reduced. The spectral distribution characteristic offers a very sensitive check on the caesium-to-oxygen ratio for a caesium-oxide photosurface, but unfortunately it is not a characteristic suitable for practical control of processing. Rather, it is one of the results which must be held to somewhat close limits by other less-sensitive methods of test.

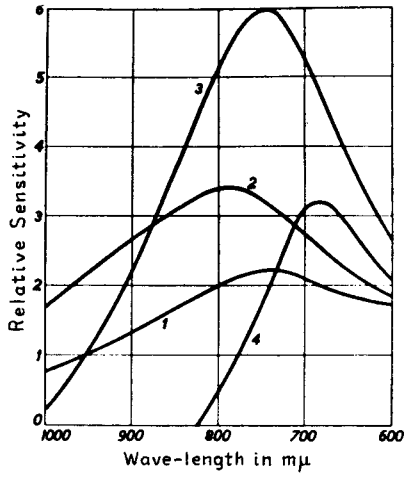


Fig. 19 - Several spectral distribution curves of Ag: Cs_2O, Ag-Cs photocathodes with different amounts of adsorbed Cs atoms. The amount of Cs increases from curve 1 to curve 4. C. H. Prescott, Jr. and M. J. Kelly, Bell System Tech. Jour. 11, 334 (1932). Courtesy of American Telephone & Telegraph Company.

7. Silver Sensitization

The Japanese physicists, Asao and Suzuki,[4] discovered that the sensitivity of the caesium-oxide type photosurface could be increased from about 20 to 100% by the evaporation of a metal such as silver after the optimum photosensitivity has been obtained by baking-in the caesium. Usually the evaporation of a very small amount of Ag on the sensitive surface causes a slight rise in sensitivity, about 5%, followed by a rapid decrease in sensitivity with further silver evaporation. However, if the Ag is evaporated until the photosensitivity has decreased to about 20% of its original value and baked for a few minutes between 170 and 190°C, the sensitivity will usually rise to about 50% or more above the original value.

The spectral distribution curve shown in Fig. 20 is typical for a caesium-oxide-type surface sensitized by further evaporation of Ag after caesiation. The infra-red peak sensitivity is high and occurs at about 8500 A°. The threshold wavelength is near 12000 A°. The ultra-violet peak is not greatly changed by this process.

8. Semitransparent Photosurfaces

It has been found that thin films may show rather high photoelectric emission when light is incident on one side of the film and the electrons are taken off the opposite side. Films have been produced with white-light sensitivities from 15 to 40 μa/lumen. However, the films

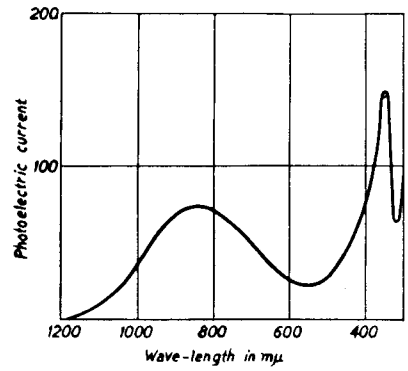


Fig. 20 - Representative spectral distribution curve of a Ag: Cs_2O, Ag-Cs cathode with a surplus of Ag in the layer. S. Asao, Physics, 2, p. 19 (1932). Courtesy of American Physical Society.

with the higher sensitivities have very high values of resistance (in the order of several megohms). Films with sensitivities in the order of 15 to 20 μa/lumen have sufficiently low values of resistance (in the order of 100000 ohms) to be useful in television pickup tubes. In Fig. 21 is shown the spectral distribution curve for a semitransparent photosurface having a white-light sensitivity of 16 to 17 μa/lumen.

Besides being useful for television pickup tubes of a special type, the semitransparent photosurface can be used for a microscope, an infra-red image tube, or phototubes for special applications. The form of and results obtained with an image tube using a semitransparent photosurface are shown in Fig. 22. At the upper right is a diagram of the image tube. The lens O focuses an image on the semitransparent photosurface K. The electron image is focused on the luminescent screen S by the electrostatic lens E, formed by the potential difference between the electrodes Z_1 and Z_2 and the magnetic lens M. The mesh grid at the left was photographed on the luminescent screen from an image on the semitransparent photocathode. The pictures of the

[4] S. Asao and M. Suzuki, "Improvement of Thin Film Caesium Photoelectric Tube," Proc. Phys. Math. Soc. Japan, Ser. 3, Vol. 12, pp. 247-250; 1930.

boy at the lower right are the light image on the semitransparent photocathode and the electron image on the luminescent screen, respectively.

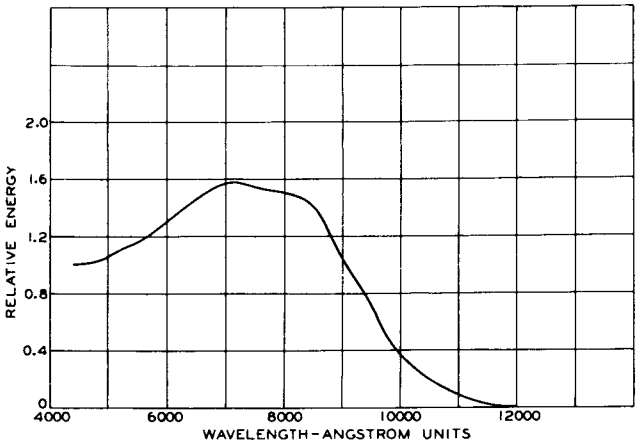


Fig. 21 - Spectral distribution curve for semitransparent photosurface having white-light sensitivity of 16-17 μa/lumen.

9. Sources of Alkali Metals

The alkali metals may be obtained from their compounds by a reducing reaction with a metal, in a manner similar to the thermit reaction between iron oxide and aluminum. A reaction of this type

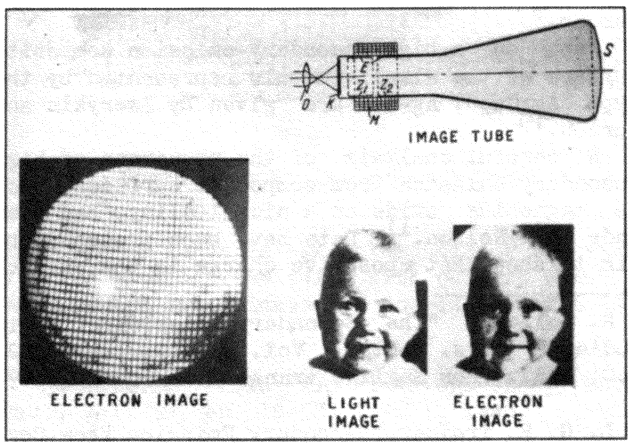


Fig. 22
(W. Schaffernicht, Zeits für Phys., Vol. 93, pp. 765, 767, 1935)

is exothermic and liberates considerable heat after being started at a temperature of 500° to 1000°C depending on the materials. The heat of reaction raises the temperature of the reacting constituents and helps to carry the reaction to completion.

The following metals and compounds of the alkali metals can be used as suitable sources of the alkali metals for phototubes: alkali chromate, dichromate, or permanganate used with silicon, aluminum, tantalum, columbium, titanium, or vanadium. These materials are generally used in the form of finely divided powders. The alkali compound is mixed with the reducing metal and the mixture pressed into a small pellet. This pellet is enclosed in a tight-fitting metal box and flashed by high-frequency heating. The reducing reaction for some mixtures begins at a relatively low temperature and proceeds rather slowly, while for other mixtures the reaction may start at a much higher temperature and proceed with almost explosive violence. Different mixtures are used for different purposes. In phototubes, a mixture of caesium dichromate and silicon is quite widely used as a source of caesium.

BIBLIOGRAPHY

Text Books

Hughes and Dubridge, "Photoelectric Phenomena." McGraw-Hill Book Co., Inc., 1932.

L. A. DuBridge, "New Theories of the Photoelectric Effect." Herman and Cie., Paris, 1935 (in English).

Zworykin and Wilson, "Phototubes and Their Application." John Wiley and Sons, 1930.

J. H. DeBoer, "Electron Emission and Adsorption Phenomena." Cambridge University Press, The Macmillan Company, 1935. Chapters I, II, III, IV, V, VI, VIII, IX, XIII.

Periodicals

W. V. Houston, "The Surface Photoelectric Effect," Phys. Rev., Vol. 52, p. 1047; Nov. 15, 1937.

A. L. Hughes, "Fundamental Laws of Photoelectricity," Journal A.I.E.E., p. 1149; Aug. 1934.

W. S. Huxford, "Townsend Ionization Coefficients In Cs - Ag - O Phototubes Filled with Argon," Phys. Rev., Vol. 55, p. 754; April 15, 1939.

H. E. Ives and H. B. Briggs, "Calculated and Experimental Photoelectric Emission from Thin Films of Potassium," Jour. O.S.A., Vol. 26, p. 247; June, 1936.

H. E. Ives and A. R. Olpin, "Optical Factors in Caesium Silver Oxide Phototubes," Jour. O.S.A., Vol. 24, p. 198; August, 1934.

L. B. Linford, "Recent Developments in the Study of the External Photoelectric Effect," Rev. of Mod. Physics, Vol. 5, No.1, p. 34; January, 1934

Carl F. Overhage, "The Normal Energy Distribution of Photoelectrons from Sodium," Phys. Rev., Vol. 52, p. 1039; Nov. 15, 1937.

C. H. Prescott, Jr. and M. J. Kelly, "Caesium-Oxygen Silver Photoelectric Cell," Bell System Tech. Jour., Vol. 11, p. 334; July, 1932.

H. C. Rentschler and D. E. Henry, "Effect of Oxygen upon the Photoelectric Threshold of Metals," Jour. O.S.A., Vol. 26, No.1, p. 30; Jan.,1936.

F. Seits and R. P. Johnson, "Modern Theory of Solids," Jour. of App. Physics, Vol. 8, 1937. (Part I) February, p. 84; (Part II) March, p. 186; (Part III) April, p. 246.

C. Zener, "Theories of the Spectral Selective Photoelectric Effect," Phys. Rev., Vol. 47, p. 15; Jan. 1, 1935.

Part II - Secondary Emission

INTRODUCTION

Thermionic and photoelectric emission from a metal surface occur when the conduction electrons (free electrons) of a metal absorb a sufficient amount of thermal energy or energy in the form of incident light to overcome the potential barrier represented by the work function of the metal surface. Therefore, metals of low work function readily emit electrons when subjected to either thermal or light energy. The number of emitted electrons is found experimentally to agree well with the number calculated on the basis of the general electron theory of metallic conduction. However, considerably less is known about the mechanism of secondary emission (the emission of electrons from a solid bombarded by electrons). For example, until the last few years it has been impossible to determine from the literature whether metals of low work function give high or low values of secondary emission. From measurements by Copeland[5] on beryllium, aluminum, and calcium, the impression is gained that the secondary emission from these metals is large compared with that from metals of higher work function, such as nickel and tungsten. However, Farnsworth,[6] who investigated several metals in high vacuum, found that freshly evaporated surfaces of magnesium had a lower value of secondary emission than nickel, iron, or tungsten. The conflicting results in the large amount of published data on secondary emission are due to the great influence of small surface contaminations on the secondary-emission characteristics of solids and the poor vacuum conditions often used. Most of the older data, with the exception of the data by Farnsworth as well as other data taken under good vacuum conditions with well-degassed surfaces of metals which can be heated to very high temperature such as tungsten, molybdenum, tantalum, and columbium, are not only inaccurate but are not even good enough to show general trends. Therefore, very little of the early data will be used in this treatment of the subject. However, for the purpose of historical interest and completeness, reference is given to a rather complete summary of the entire field of secondary emission and a comprehensive bibliography up to 1937 published by Kollath.[7]

Beginning with 1937, Bruining and DeBoer have published a series of papers on secondary emission which give data on a large number of metals, composite layers and compounds, taken under good vacuum and very carefully controlled conditions. These data are used to formulate a consistent and logical theory of secondary emission which is an important addition to our knowledge of this phenomenon. Reference will be made to these papers as the various subjects are discussed.

Some interesting results on the relation between secondary emission and work function for very thin films of metals of low work function, for example, barium on a metal of high work function such as tungsten, are given in a paper by Treloar.[8] The results of Treloar are in agreement with similar data given by Bruining and DeBoer.[9]

Data on a high secondary-emission composite surface of the alkali metals represented by the type Ag:Cs_2O, Ag-Cs are given by Zworykin and Morton.[10]

A careful analysis of the phenomena of high secondary emission from composite surfaces, such as magnesium oxide on a nickel alloy, has been made by Nelson.[11] Data have been presented by him to show that a positive charge in the surface

---

[5] P. L. Copeland, "Correlation between Variation of Secondary Electron Emission and Atomic Number," Phys. Rev., Vol. 46, pp. 167-168; Aug. 1, 1934.; "Secondary Emission of Complex Targets," Phys. Rev., Vol. 48, pp. 96-98; July 1, 1935.

[6] H. E. Farnsworth, "Electronic Bombardment of Metal Surfaces," Phys. Rev., Vol. 25, pp. 41-57, July, 1925.

[7] R. Kollath, "The Secondary Emission of Solid Bodies," Phys. Zeits., Vol. 38, No.7, pp. 202-240; 1937. (For English translation, see Ref.4-7)

[8] L. G. R. Treloar, "Secondary Emission from Complex Surfaces," Proc. Roy. Soc., Sec. A, Vol.49, p. 392; 1937.

[9] H. Bruining and J. H. DeBoer, "Secondary Electron Emission," Part VI—The Influence of Externally Adsorbed Ions and Atoms on the Secondary Electron Emission of Metals, Physica VI, No.9, pp. 941-950; October, 1939.

[10] V. K. Zworykin and G. A. Morton, "Television," p. 30, John Wiley & Sons; 1940.

[11] Herbert Nelson, "Phenomena of Secondary Electron Emission," Phys. Rev., Vol. 55, p. 985; May 15, 1939.

film is an important factor contributing to high secondary emission. He also has pointed out the relation between high secondary emission from composite surfaces and thin-film field emission described by Malter.[12]

Data on the secondary-emission characteristics of insulators, such as glasses and luminescent materials, are far less numerous than that on metals but some results may be found in the following references.[13,14,15]

SECONDARY EMISSION OF CLEAN METALS

The secondary-emission ratio (δ = the number of secondary electrons emitted per bombarding primary electron) increases to a maximum value, then decreases when the voltage of the primary electrons is increased, as shown in Fig. 23 which gives data on a number of clean metals. A similar type of curve is characteristic of the variation of secondary emission with primary electron voltage for contaminated or composite surfaces.

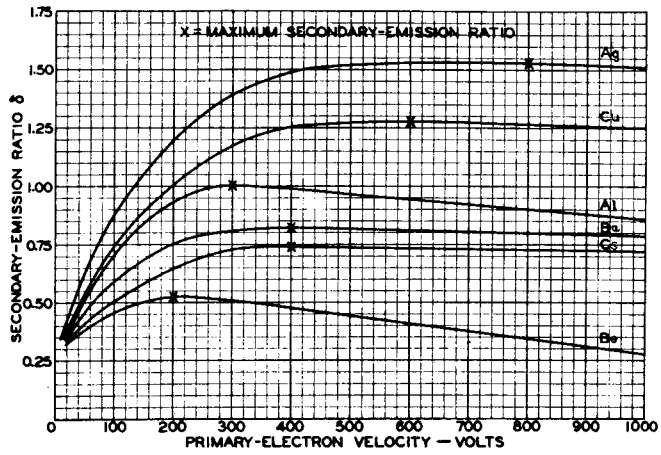


Fig. 23

(Data from Bruining and DeBoer)

Data from Bruining and DeBoer[16,17] on the maximum value of δ for several metals and the primary electron voltage at which it occurs are given in the following table along with data on work function, atomic weight, and atomic volume for comparison. It will be noted that, taken as a whole, the data in this table show a lack of correlation of the maximum value of δ with work function, atomic weight, or atomic volume.

Table I

SECONDARY EMISSION FROM CLEAN METALS

Metal	Maximum Secondary Emission Ratio δ	Primary Voltage at Maximum Emission (approx. Volts)	Work Function (Volts)	Atomic Weight	Atomic Volume
Be	0.53	200	3.16	9.0	4.9
Li	0.56	83	2.28	6.9	13.0
Cs	0.72	400	1.81	132.8	71.0
Ba	0.83	400	2.11	137.4	36.1
Mg	0.95	200	2.42	24.3	14.1
Al	0.97	300	2.26	27.0	10.0
Th	1.14	750	3.38	232.2	21.1
Ni	1.25	500	5.03	58.7	6.7
Mo	1.25	375	4.15	96.0	9.3
Cu	1.29	600	4.30	63.6	7.1
W	1.35	630	4.54	184.0	9.8
Ag	1.56	800	4.74	107.9	10.2

However, upon more critical examination, one finds that the following generalizations with regard to the secondary-emission characteristics of different metals are illustrated by the curves of Fig. 23 and the data in Table I.

1) The maximum value of δ occurs at lower voltages for the light metals than for the heavy metals.

2) The rate of decrease of δ with increasing primary-electron voltage beyond the maximum is higher for the light metals than for the heavy ones.

3) The maximum value of δ occurs at higher values of primary-electron voltage for metals of high work function than for those of low work function among metals of similar atomic weight.

---

[12] L. Malter, "Thin-Film Field Emission," Phys. Rev., Vol. 50, pp. 48-58; July 1, 1936.

[13] W. B. Nottingham, "Electrical and Luminescent Properties of Willemite under Electron Bombardment," Jour. of App. Phys., Vol. 8, pp. 762-778; November, 1937.

[14] Von C. Hagen and H. Bey, "Charge Potential of Insulators Bombarded by Electrons," Zeits. f. Physik, 104, pp. 681-684; 1937.

[15] S. T. Martin and L. B. Headrick, "Light Output and Secondary-Emission Characteristics of Luminescent Materials," Jour. of App. Phys., Vol. 10, pp. 116-127; February, 1939.

[16] H. Bruining and J. H. DeBoer — Preliminary Communication: "Secondary Electron Emission of Metals of Low Work Function," Physica IV, No.6, pp. 473-477; June, 1937.

[17] H. Bruining and J. H. DeBoer, "Secondary Electron Emission," Part I—Secondary Emission of Metals, Physica V, No.1, pp. 17-30; January, 1938; Part II—Absorption of Secondary Electrons, Physica V, No.10, pp. 901-912; December, 1938.

4) The maximum value of δ is higher for metals of high work function than for metals of low work function among metals of similar atomic weight.

5) The maximum value of δ is higher for metals of low atomic volume than for those of high atomic volume except in the case of a few very light metals with low atomic volume.

Apart from the velocity of bombarding electrons and other factors already mentioned, the secondary-emission ratio depends upon the nature of the surface (rough or smooth) and the angle of incidence. The value of δ is a minimum for electrons arriving perpendicular to the surface (angle of incidence = 0^o) and for smooth surfaces increases from 2 to 5 times for different metals at large angles of incidence approaching 90^o. The values of δ are considerably less for rough surfaces than for smooth ones. Bruining and DeBoer[16] have found, upon analysis of data showing the relation between δ and angle of incidence for different metals and different voltages, that the low value of δ at zero angle of incidence is principally due to absorption of secondary electrons before they reach the surface. The decrease of δ with increased primary-electron voltage above the maximum value may also be attributed to secondary-electron absorption.

Accurate measurements of the velocity distribution of secondary electrons are difficult to make because of scattering and contact potential difference; therefore, little data are available. Haworth[18] gives data on the energy distribution of secondary electrons from well-outgassed molybdenum. However, it has been found that most of the secondary electrons have velocities less than 30 volts and more than half of them less than 10 volts. The velocity distribution of secondary electrons is practically independent of the velocity of the primary beam.

SURFACES OF LOW SECONDARY EMISSION

In the operation of radio tubes it is often found that high secondary emission from certain elements, for example, triode plates, and tetrode screen grids and plates, produces undesirable electrical characteristics. The suppressor grid of the pentode was introduced to eliminate the effect of secondary emission from the plate. It is not always practical to introduce a suppressor element to avoid the unwanted effects of secondary emission, nor is it always possible to design electrodes of such a shape that low-potential regions created by space charge will effectively suppress secondary emission. Therefore, surfaces which are inherently low secondary emitters have been of considerable importance in radio-tube design and a few words regarding their preparation will not be out of place.

It has been generally found that rough surfaces have low secondary emission and that carbon and finely divided graphite have very low secondary emission. Carbonized metals have been used in radio tubes because of their low secondary emission and other desirable properties for many years. Bruining, DeBoer, and Burgers[19] have published data on the secondary-emission characteristics of specially prepared carbon surfaces of very low secondary emission. Their data are shown in Fig. 24. Microphotographs of the surfaces of the two types of carbon films are shown in Fig. 25.

It appears from these pictures that the sprayed surface which gave the higher secondary emission is the rougher. However, the very finely divided precipitated soot film (Fig. 25a) probably appears rougher and has a higher percentage of open space presented to the electron than the agglomerated sprayed film (Fig. 25b). It is also interesting to note that the precipitated soot film could absorb Ba and BaO evaporated from the cathode surface without an appreciable change in secondary emission although the sprayed carbon film showed an appreciable increase in secondary emission under similar conditions.

SECONDARY EMISSION OF GLASS AND LUMINESCENT MATERIALS

The wide use of glass in radio tubes and the more recent extensive use of cathode-ray tubes with their luminescent screens lends interest to the secondary-emission characteristics of these materials. The prevalent use in such tubes of the alkaline-earth metals as getters and of colloidal graphite bulb-wall coatings, both of which are likely sources of contamination, has given rise to a wide range of secondary-emission characteristics of glass and luminescent materials as measured in various tubes.

Data on the electron bombardment of glasses which are relatively free of contaminations show that with an anode potential up to approximately 200 volts, glasses stay at cathode potential, but at slightly above 200 volts, they suddenly assume a positive potential approaching that of the anode. The glass remains near the anode potential up to about 1600 volts. On further increase in anode potential, the glass potential begins to depart from the anode potential and reaches a maximum of about 2000 volts when the anode potential is several hundred volts higher. The glass potential remains at a nearly constant value of 2000

---

[16] Loc. cit.

[18] L. J. Haworth, "The Energy Distribution of Secondary Electrons from Molybdenum," Phys.Rev., Vol. 48, pp. 88-95; July 1, 1935.

[19] H. Bruining, J. H. DeBoer, and W. G. Burgers, "Secondary Electron Emission of Soot in Valves with Oxide Cathodes," Physica, Vol. 4, pp. 267-275; 1937.

volts independent of further increase of anode potential. When the anode potential is decreased below 2000 volts, the same curve is followed downward to about 200 volts, where the glass instead of going suddenly to cathode potential as in the case of increasing voltage, continues to remain near anode potential down to about 100 volts, at which point the secondary-emission ratio again becomes less than 1, and the glass goes suddenly to cathode potential. This type of curve is characteristic of glasses, but an increase of alkali oxide on the surface will increase somewhat the voltage range over which the glass floats near anode potential. Clean Nonex glass gives a secondary-emission ratio equal to or greater than 1 over a small range of anode potential from about 150 volts up to about 1200 volts. Lime glass with an excess of alkali oxide on the surface gives a secondary-emission ratio equal to or greater than 1 for a larger range of anode potential from about 70 to several thousand volts.

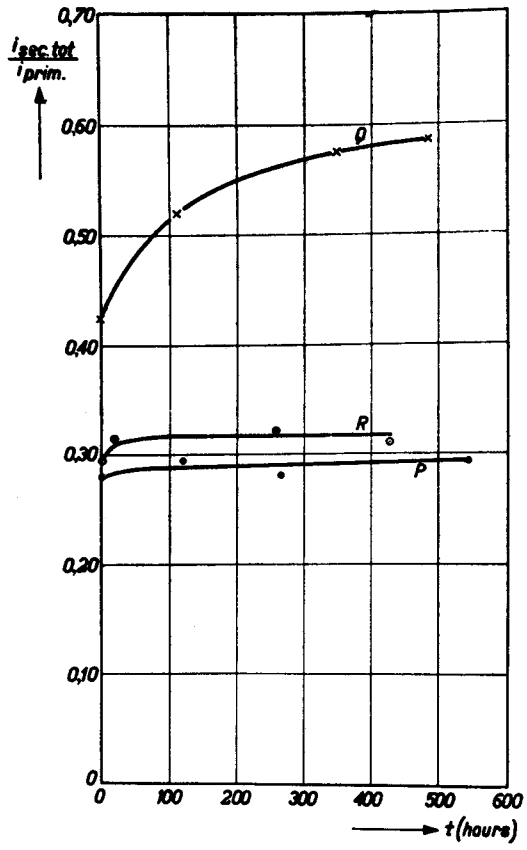


Fig. 24 - Change of capability for secondary emission of different soot surfaces face to face with an oxide cathode as a function of time.
Curve P: soot precipitated from a flame; V_a = +250 V, V_g = +11 V. (mean value for 5 tubes)
Curve Q: sprayed soot; V_a = 250 V, V_g = +11 V. (mean value for 9 tubes)
Curve R: soot precipitated from a flame; $V_a = V_g$ = 100 V. (mean value for 5 tubes)

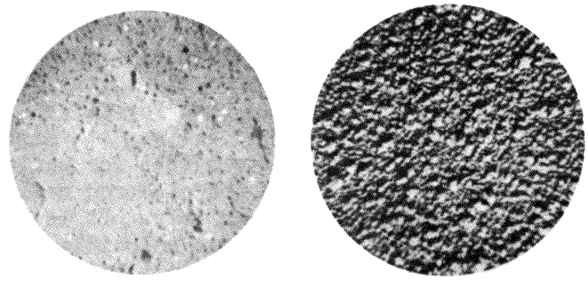


Fig. 25 - Microphotographs of two kinds of carbon soot. (a) Precipitated directly from a flame of burning turpentine: particle size is smaller than the resolving power of the microscope. (b) Obtained by spraying an alcoholic suspension of soot: agglomerated particles of ~100 μ diameter. (DeBoer, Bruining, & Burgers, reference 19)

There are three classes of luminescent materials generally used in cathode-ray tubes, i.e., silicates, sulphides, and tungstates. Examples of these are: zinc silicate (green willemite), zinc beryllium silicate (yellow to orange), zinc sulphide (blue to green), zinc cadmium sulphide (green to red), calcium tungstate, and magnesium tungstate (blue). Values of the primary-electron beam voltage for which δ is greater than 1 with luminescent screens on glass are given in Table II.

Table II

Luminescent Material Class	Electron Voltage for δ > 1 (Volts)
Silicates	300 to 15000 *
Sulphides	400 to 8000 *
Tungstates	500 to 4000 *

* The higher potentials are often referred to as the limiting potential of the material in question.

These values are only rough approximations as they may be appreciably influenced by BaO or other contaminations and also by the fact that the glass surface area is probably not more than 75 per cent covered by luminescent material. It has

been shown by experiment that all values of the limiting potential, between those given in Table II and that for glass, can be obtained by variations in glass coverage with luminescent material.

RELATION BETWEEN SECONDARY EMISSION AND WORK FUNCTION AT LOW ELECTRON VELOCITY

It has been shown by Bruining and DeBoer[20] in experiments on barium and silver using electrons with very low velocities (below 100 volts), that δ is greater for barium than for silver for electron velocities below 30 volts. At higher voltages δ is greater for silver than for barium. The results were corrected for electron reflection which was measured separately. It is concluded that at very low electron velocities, absorption of secondary electrons is not important because the primary electrons penetrate only a very few atomic layers. Therefore, the secondary electrons for the most part have only the potential barrier or work function at the surface of the metal to overcome in order to be emitted. Under these conditions, metals of low work function may be expected to show higher values of δ than those of high work function, while at higher values of electron velocity where the primary electrons penetrate a large number of atomic layers, the absorption of the secondary electrons is of greater importance than the surface work function.

Further data on the relation between secondary emission and work function at low electron velocities are given by Treloar[8] with experiments on barium-coated tungsten and by Bruining and DeBoer[9] with experiments on barium-coated molybdenum. In both of these cases it is found that as barium is evaporated onto a metal of high work function, in less than a monatomic layer, δ increases until the work function reaches a minimum value, and then slowly decreases as the work function increases with more barium on the surface. Bruining and DeBoer[9] also show that as barium is evaporated the photoelectric emission rises with an increase in δ but drops much more rapidly than δ beyond the minimum value of the work function as the work function increases. Another interesting point brought out by Bruining and DeBoer[9] is that the δ increases linearly with an increased percent coverage by barium of a molybdenum surface while the photoemission increases exponentially with the surface coverage for values of surface coverage below that corresponding to the minimum value of work function. This difference between the behavior of secondary and photoelectric emission is due mainly to the difference in electron velocity range in the two cases. In the case of the photoelectric emission excited by visible light, the electrons as they approach the surface from within the metal have a velocity distribution of about 3 volts, while for secondary electrons the velocity distribution is about 30 volts. By using ultraviolet radiation (λ = 2537 A°) where the electron velocity distribution is increased to nearly 5 volts, the photoemission increases with the surface coverage nearly linearly and thus more closely approaches the linear relation between surface coverage and δ.

SECONDARY EMISSION OF COMPOUNDS AND COMPOSITE SURFACES

Compounds and composite surfaces of the alkali and alkaline-earth metals have high values of δ. They are much higher than those of the pure metals of which the compounds are composed, as illustrated by the data on the left side of Table III and the curves in Fig. 26. The data given in the right side of Table III show that δ for other compounds is of the same order as δ for the metals of which they are composed.

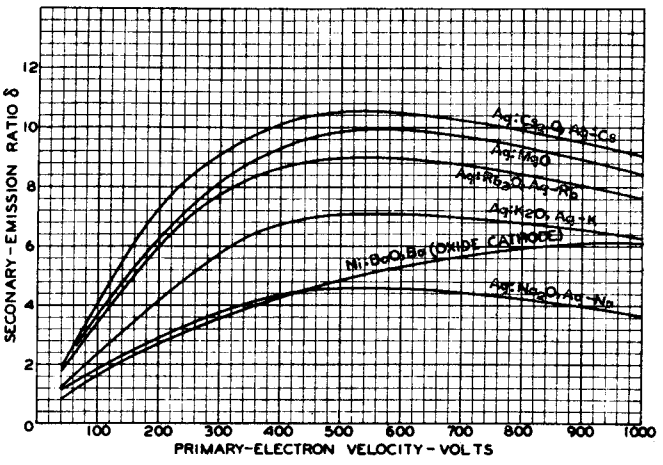


Fig. 26

(Data for MgO from Nelson, reference 11; data for alkali metals from Zworykin and Morton, reference 10; data for Ba oxide cathode from Bruining and DeBoer, reference 21.)

Layers of the compounds in the left column have not been found to make suitable surfaces of high secondary emission for practical tubes because electron bombardment of these materials causes δ to decrease with time rather rapidly if any appreciable current per unit area is drawn from the electrode. Quite satisfactory results have been obtained in electron multipliers which use the Ag:Cs_2O, Ag-Cs type composite surface

---

[8] Loc. cit.

[9] Loc. cit.

[20] H. Bruining and J. H. DeBoer, "Secondary Electron Emission," Part III — Secondary Electron Emission Caused by Bombardment with Slow Primary Electrons, Physica V, No.10, pp. 913-917; Dec., 1938.

similar to that used for photoelectric emission. Nelson,[11] as well as Bruining and DeBoer,[21] give data showing high values of δ for a thin film of magnesium oxide obtained by evaporation of magnesium oxide or evaporation of magnesium in a low pressure of oxygen. These films of magnesium oxide have also been found to be suitable for use as high secondary-emission surfaces in practical tubes.

Table III

Compound	δ_{max}	Compound	δ_{max}
CaF_2	3.2	WS_2	0.96 - 1.04
BaF_2	4.5	AgO_2	0.90 - 1.18
NaI	5.5	MoS_2	1.10
KI	5.6	MoO_2	1.09 - 1.33
LiF	5.6	Cu_2O	1.19 - 1.25
NaF	5.7		
RbCl	5.8		
CsCl	6.5		
NaCl	6.8		
KCl	7.5		

Compounds that have high values of δ are those in which the unfilled electron band has such an energy position that the secondary electrons brought in to this band by the action of primary electrons are able to escape from the surface without taking up additional energy. Thus they are, in general, insulators where the ion of an electro-positive element has a closed electron shell (similar to the rare gas structure) and are not good semiconductors. However, to be of practical value as a high secondary-emission surface in tubes where current must be drawn through the surface, the compound or composite layer must have some appreciable electron conductivity. Therefore, for practical high secondary-emission surfaces, we look for materials which are poor semiconductors, or good insulators which have been modified by their method of preparation so that they become poor semiconductors.

MECHANISM OF SECONDARY-ELECTRON EMISSION

Until recently it has not been possible to describe clearly the mechanism of secondary emission because of so much apparently conflicting experimental data reported in the literature. However, through the work of Bruining and DeBoer on clean metals as well as on compounds and composite surfaces, and that of Nelson, it is now possible to see what data are most reliable and to present at least a qualitative consistent picture of the mechanism of secondary emission. The views as presented here are derived mainly from Nelson[11] and Bruining and DeBoer.[22]

When the primary electron passes through the surface of a substance, its energy for the most part may be transferred to bound electrons at different distances from the surface. Very little energy can be transferred from the primary electron to the nearly free conduction electrons in metals. In order that the secondary electron be emitted after it has received energy from the primary electron, it must be able to travel through the substance to the surface and then have sufficient energy to overcome the work function of the surface. As has been pointed out, the work function of the surface exerts an important influence on secondary emission only for very-low-velocity primary electrons. Therefore, it may be concluded that secondary electrons inside a substance usually have ample energy to overcome the surface work function. Thus, it appears that the following factors are most important in determining what values of δ will be obtained.

1) The probability of transfer of energy in suitable small amounts from the primary electron to a number of bound electrons within the substance.

2) The transmission coefficient of the substance for low-velocity electrons.

Very little is known about the properties of solids which influence the first factor. However, it is reasonable to assume that the first factor does not vary widely for different solids and then to consider what properties influence the second factor. We know that metals have a large number of unfilled energy bands capable of absorbing low-velocity electrons and thus they have a high absorption coefficient for the yet unemitted secondary electrons which may account for the generally low secondary emission of metals. On this basis, metals which have the higher values of δ are those that have low atomic volume and a large number of bound electrons in the outer level, such as the heavy metals which have a high absorption coefficient for fast electrons and thus provide an opportunity for secondary electrons to be produced near the metal surface where they may readily escape. Metals with large atomic volume and few bound electrons in the outer levels allow deep penetration of the primary electrons and, therefore, the secondary electrons have long distances to travel to the surface with a good chance of absorption before they get there and thus few are emitted.

---

[11] Loc. cit.

[21] H. Bruining and J. H. DeBoer, "Secondary Electron Emission," Part IV—Compounds with a High Capacity for Secondary Electron Emission, Physica VI, No.8, pp. 823-833; August, 1939.

[22] H. Bruining and J. H. DeBoer, "Secondary Electron Emission," Part V—The Mechanism of Secondary Electron Emission, Physica VI, No.8, pp. 834-839; August, 1939.

The fact that the maximum value of δ occurs at lower voltages and the rate of fall of δ beyond the maximum is greater for the lighter elements of high atomic volume than for the heavy elements of low atomic volume, also fits in well with the absorption picture as described above. For, if we may say roughly that the optimum value of secondary emission occurs for given depth of penetration of the primary beam, then the primary electrons will reach this depth, at lower voltage in metals of high atomic volume than in those of low atomic volume. Also, if the rate of penetration of primary electrons is high, the rate of decrease of δ above the maximum with increasing voltage will be high.

Following this same picture, we conclude that the secondary emission of compounds and composite surfaces which are fairly good insulators is, in general, high because their absorption coefficient for low-velocity electrons is low. The absorption coefficient is low because they have few empty electron bands, a fact which accounts also for their low electronic conductivity. It will be noted from Table III that there are apparently two classes of compounds; those with values of δ much higher than for metals, and those with values of δ of the same order as for metals. These two groups of compounds have different empty electron energy bands, as follows:

1) The compounds with high values of δ have the empty electron energy band where secondary electrons entering it may be emitted without receiving more energy.

2) The compounds with low values of δ have an empty electron energy band where secondary electrons cannot leave unless they originate from surface atoms, or unless they receive additional energy to bring them to a still higher empty band from which they may leave the metal without still further energy.

Therefore, we may say compounds of the first group have low absorption for low-velocity secondary electrons, while those of the second group have a high absorption for low-velocity secondary electrons similar to metals.

Bruining and DeBoer[22] bring out one other interesting point regarding the correlation of secondary-emission characteristics with other properties, namely, the compounds with low secondary emission are those in which the long wavelength limit of the internal photoelectric effect is at longer wavelengths than that of the external photoelectric effect. Or, in other words, in compounds where the crystal lattice shows light absorption at longer wavelengths than that corresponding to the long wavelength limit of the external photoelectric effect, δ will be low (semiconductors). When, however, the long wavelength limit of the external photoelectric effect corresponds to the first absorption band on the long wavelength side of the absorption spectrum, δ will be high (colorless insulators).

It has been shown by Nelson[11] with carefully controlled experiments on thin films of magnesium oxide deposited by evaporation on a nichrome plate, that a positive charge produced in the insulating layer as a result of secondary emission is responsible for enhancement of the secondary emission from such surfaces. Although the dielectric strength of such thin films is not sufficient to give rise to thin-film field emission as described by Malter,[12] it is sufficient to give a field which will materially increase the emission of secondary electrons from the layer.

---

[11] Loc. cit.

[12] Loc. cit.

[22] Loc. cit.

Lecture 5

LUMINESCENT MATERIALS

H. W. Kaufmann

The term "luminescence" was originally used to mean the emission by a substance of visible light, whether as the result of excitation by some form of energy supplied from a source external to the substance itself (such as light, x-rays, ions, cathode-rays, or mechanical work), or as the result of a purely chemical process, such as the slow oxidation of phosphorus. This definition was logically extended to include the emission of ultra-violet and infra-red radiation as well; and it is so used today. The luminescence phenomenon of chief use in electron tubes and, therefore, of importance to us is that of cathodoluminescence, i.e., the emission of light as the result of cathode-ray bombardment. The technical requirement of a high vacuum which shall endure during the life of the electron device necessitates our excluding from consideration any cathodoluminescent materials which are unstable under reduced pressure at the baking-out temperatures required for the economical production of the requisite vacuum. This limitation compels us to focus our attention only upon inorganic solids of relatively high melting point and chemical stability. Experience shows further that, although there are glasses which glow under electron bombardment, their efficiency as light sources is much lower than that of a number of well-known and readily available crystalline materials which are commercially useful in vacuum tubes.

Proceeding from the purely technical viewpoint, we come first to discuss the silicates because they are efficient light sources, simple to produce, and stable during processing and use. The patriarch of silicates used in electron devices is willemite, whose formula is either Zn_2SiO_4 (as the general inorganic chemist writes it) or $2ZnO \cdot SiO_2$ (written in the terminology of the ceramist, who prefers to consider it as a compound of two oxides). Willemite exhibits an outstanding characteristic of many lumiphores (luminescent materials): if prepared in the crystalline form, with as few impurities as present chemical technique permits us to do, it rewards our care by refusing to give any usably intense luminescence even under vigorous bombardment. Addition of small amounts of manganese-oxide to the crystal structure, in proportions from one-tenth to two per cent by weight, causes the willemite to produce the brilliant green light so commonly seen on tuning indicators and ordinary oscilloscope tubes. The phrase "to the crystal structure" is important. It means that we cannot take pure zinc silicate, simply grind it well with manganese oxide and produce luminescence; but if we take that mixture and heat it to such a temperature that the manganese can actually be absorbed into the crystal lattice (which begins to occur slowly around 600°C, and proceeds rapidly at 900° or higher), we shall obtain the desired effect. The manganese is then called an activator. It is characteristic of activators that they are required in comparatively small amounts to produce luminescence. Manganese is one of the most common activators; copper is about as commonly effective; some others used in various connections are silver, bismuth, chromium, and various rare earths. Copper in willemite produces a blue hue.

Willemite is a very convenient material to manufacture, as it can be prepared by mixing intimately together the various oxides of which it is composed (together with a barium compound, which is needed to promote secondary emission) and firing the mixture to a temperature of 1200°C for a period of the order of half an hour. It can then be ground in a pebble mill to very small size without losing much of its efficiency and can be applied by spraying from an ordinary spray gun, either directly from a volatile suspending liquid, or in a nitrocellulose binder which is burned away at some time during the subsequent processing. Willemite, like most oxygen compounds of its kind, can be heated in air at any temperature up to that at which the zinc oxide becomes volatile (about 1400°C) without injury.

Luminescent screens applied to glass have an optimum thickness for any given velocity of bombarding electrons. For the conditions under which our willemite screens are produced and used, the optimum thickness corresponds to about 40 or 45 per cent light transmission of the sprayed screen.

The efficiency of an average willemite screen is about 2.3 candlepower per watt when it is scanned in a 6 by 8 centimeter pattern (30 by 10000 cycles) by a focused electron beam of 50 microamperes at 6 kilovolts. Values of efficiency for other screen materials to be mentioned later on will be on the basis of these same conditions. The variation of light output with voltage and current for several screen materials is shown in Figs. 1a, 1b, and 1c. From these figures it may be seen that the efficiency increases with increasing voltage, and decreases with increasing current.

A curve of the spectral distribution of the light emitted by willemite (Fig. 2) makes clear why it appears green to the eye. Since this color is definitely unpleasant for television reception, it is better to produce a screen having a yellow light output, which more nearly resembles the yellow light sources which by convention we have come to regard as white. Such a silicate can be obtained in a number of ways. Fusing the green (alpha) willemite at a temperature slightly over 1500° and cooling it under

suitable conditions will produce a yellow-luminescent substance (beta willemite) of good efficiency, and having a spectral distribution characteristic as shown in Fig. 2. The process of fusion, however, is costly because of the high temperatures involved, and the final product is difficult to remove from the platinum vessel in which it has been melted. A simpler way of obtaining a yellow-luminescent material is to replace part of the zinc in alpha willemite by beryllium. This results in a displacement of the green-willemite emission band toward the longer wavelengths as shown in Figs. 3 and 4. This displacement toward the red can be increased by increasing any of the following: the proportion of beryllium relative to zinc; the proportion of

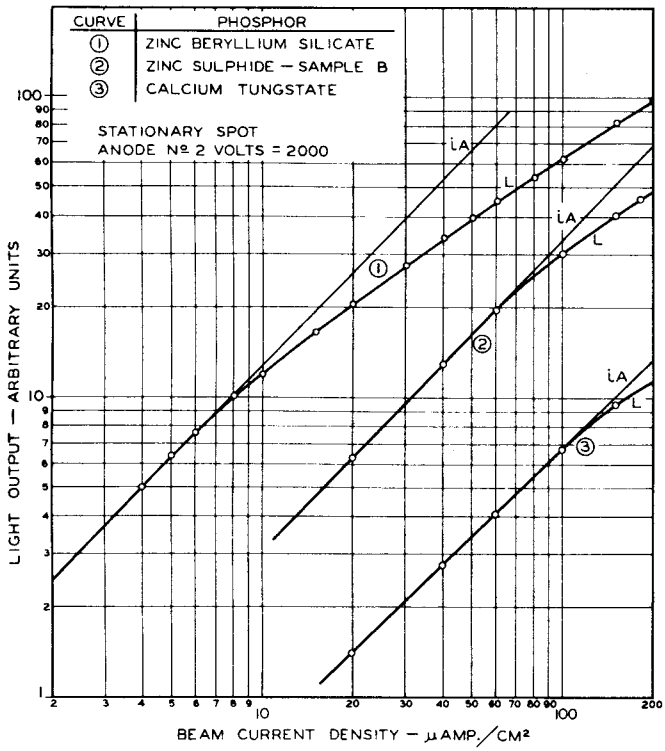


Fig. 1a

manganese; the proportion of silica relative to the zinc and beryllium oxides (taken together); or the firing temperature.

Zinc-beryllium silicate has the same general properties as willemite with the exception that its efficiency in candlepower per watt must be decreased the more the peak of emission is shifted toward the red end of the spectrum away from the point of maximum eye sensitivity. It is possible to obtain a given color by using a number of combinations of the variables mentioned above, and it is necessary to determine experimentally which particular combination is the optimum for the color desired. The material known as phosphor No.3 used in our television types 1800 and 1801 gives a pleasant yellow color. Its spectral distribution characteristic is shown in Fig. 5. The efficiency of phosphor No.3 is about 1.8 candlepower per watt.

It is also possible to obtain a good yellow light by bombarding calcium silicate, manganese activated. Here, however, we come into a difficulty due to a property we have not yet discussed, i.e., phosphorescence or persistence. Willemite and its beryllium-substitution deriva-

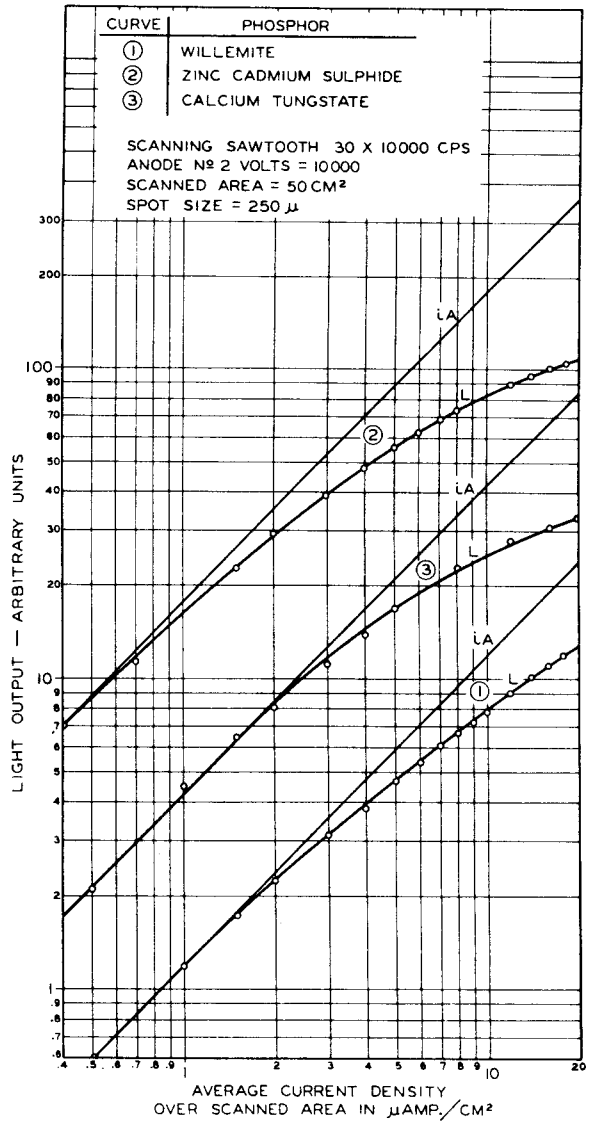


Fig. 1b

tives have a decay curve in which the energy drops to one per cent of its original value in roughly one twentieth of a second. Calcium silicate takes approximately one-tenth of a second to decay, and so television tubes cannot be screened with it because moving images would be followed by ghostly shadows.

Yellow highlights and black shadows in a picture are better than green and black, but they

are definitely less desirable than black and white. The observer may not object to white highlights and colored shadows, as in the rotogravure supplements, but color in what should be white highlights is "agin' nature." To produce a white-emitting screen, we can add thoria and zirconia to the zinc-beryllium silicate to increase the emission in the blue end of the spectrum (Fig. 6). It is possible to obtain a fair white

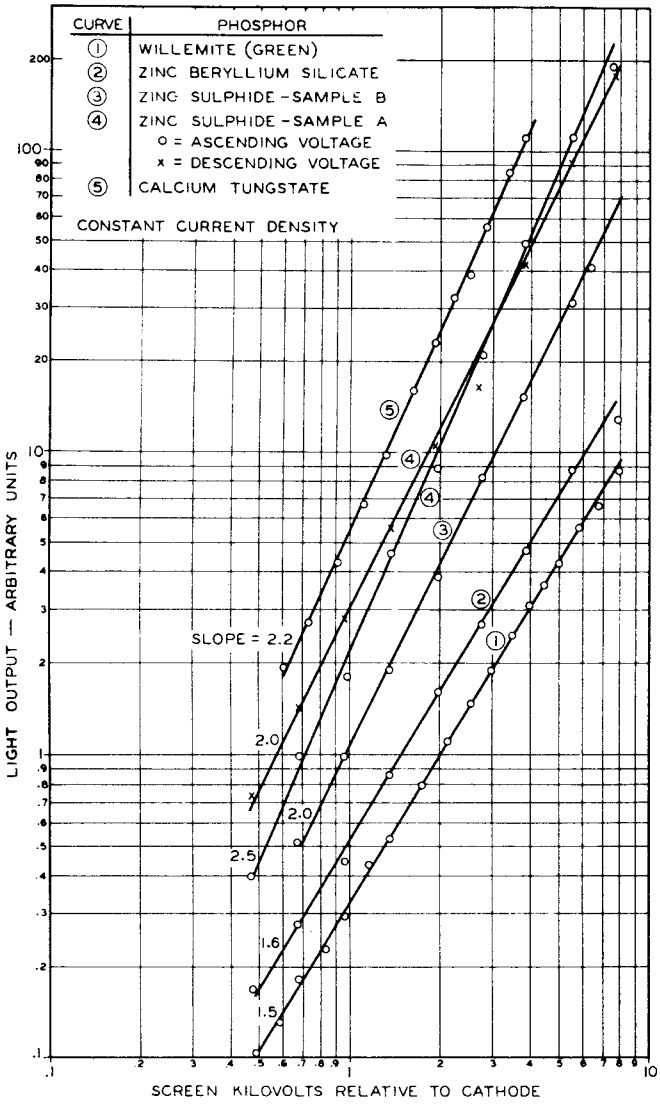


Fig. 1c

by such means, but the product has an efficiency of only 0.5 candlepower per watt. This efficiency is too low for use in tubes made according to present design criteria.

For the production of a satisfactory white cathode-ray-tube screen, we must turn to another class of compounds, the sulphides. If zinc sulphide is prepared by methods involving elaborate purifications to remove all traces of heavy metals, fluxed with suitable salts of one of the

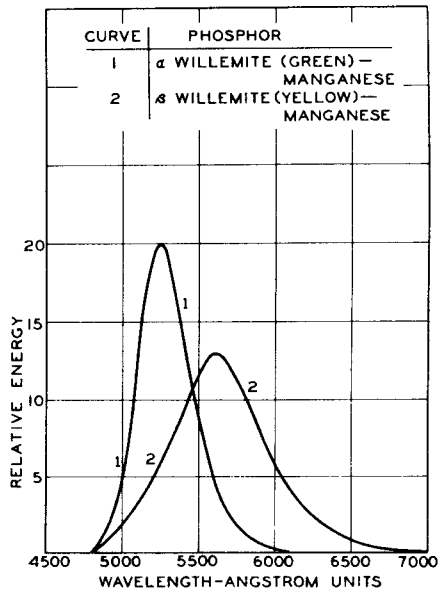
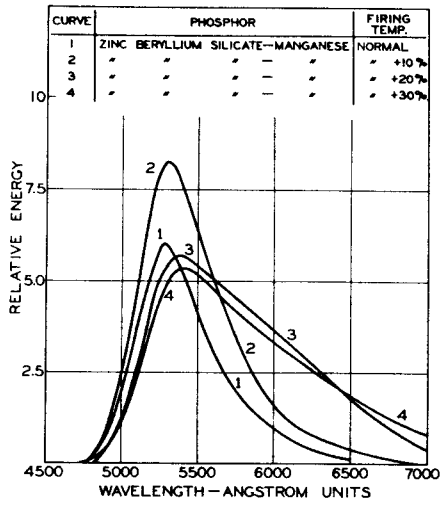


Fig. 2



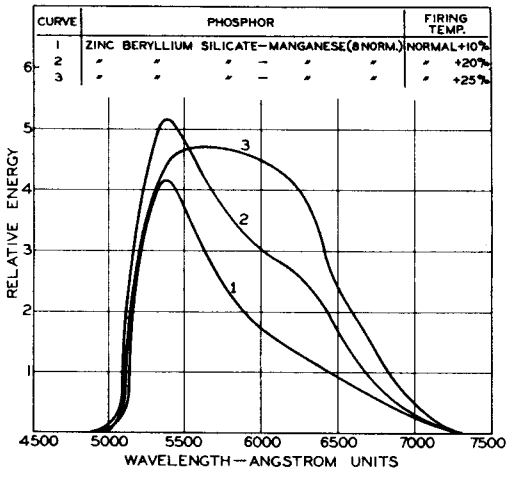


Fig. 3

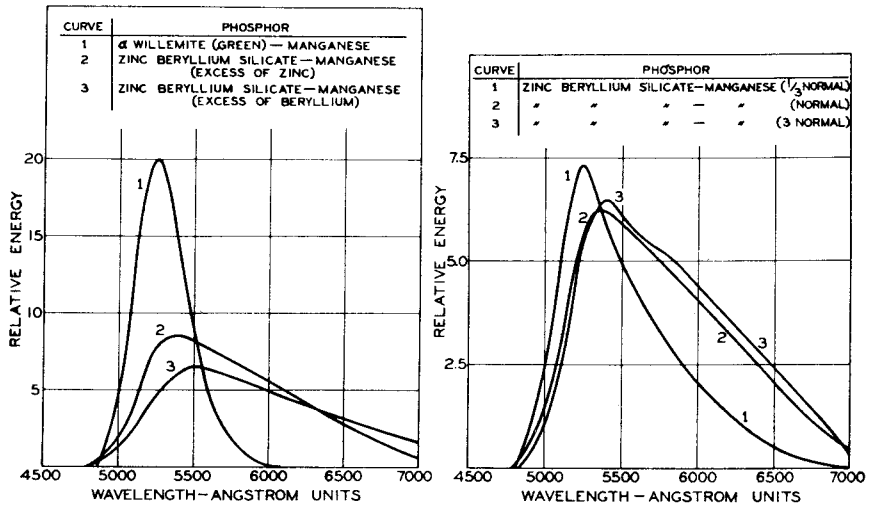


Fig. 4

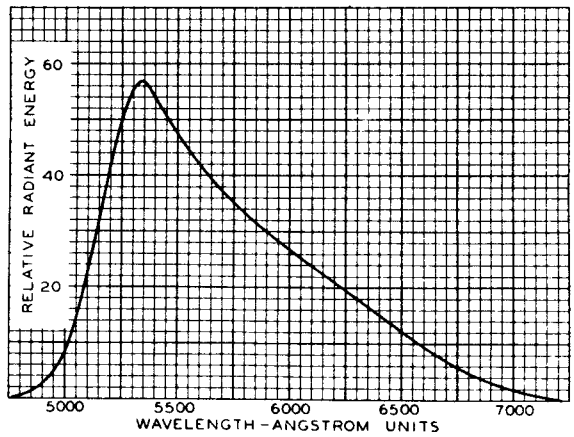


Fig. 5 - Spectral energy characteristic of phosphor No.3.

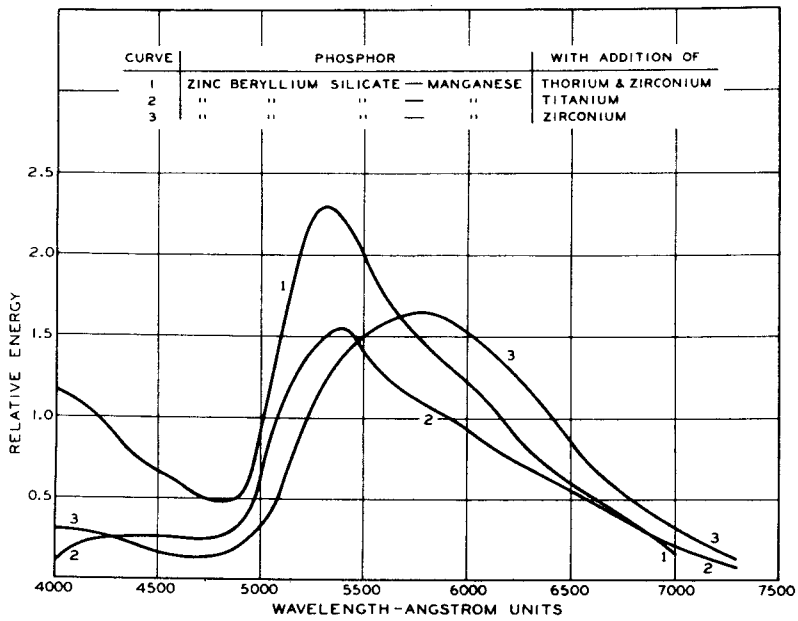


Fig. 6

metals of the alkaline or alkakine-earth groups (e.g., 2 per cent of NaCl), and crystallized at an elevated temperature, it will show a blue luminescence which consists of a rather broad band. The addition of small amounts of silver as an activator will narrow this band and produce a definitely deeper blue.[1] Replacement of the zinc by cadmium[1] produces a shift of the emission spectrum toward the red, as shown by curves Nos.3 and 4 in Fig. 7. By the use of zinc and zinc-cadmium sulphides of suitable blue- and yellow-light emission, it is possible to produce a mixture which emits white light when bombarded. The efficiency of the blue sulphide is somewhat over 1 candlepower per watt; that of the yellow zinc-cadmium sulphide has been measured as high as 4.5 in particular instances. Unfortunately,

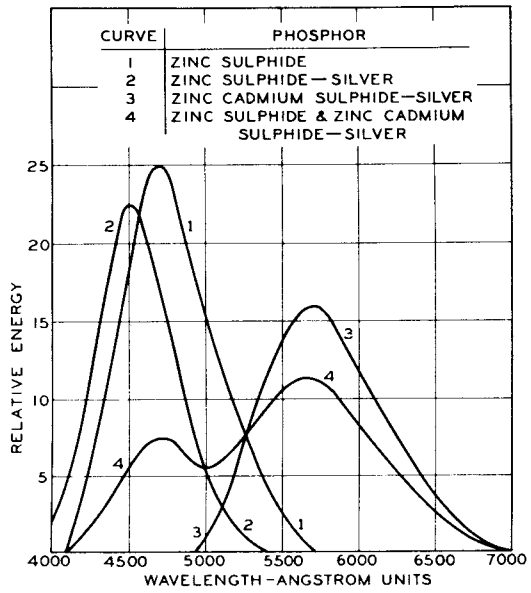


Fig. 7

the yellow sulphide is very sensitive to impurities and to heat treatment. The color of the emitted light may change after mistreatment, and the efficiency is almost certain to drop. It is customary abroad to use very long exhaust periods with low-temperature bakeouts, the process running up to a number of hours. This process by our standards, is considered as impracticable, and so the use of the yellow sulphides is definitely not regarded with favor, although the high efficiency obtainable under optimum conditions is undeniably attractive. The blue-emitting sulphide is much less sensitive to processing, and it is possible to mix it with yellow-emitting zinc-beryllium silicate to produce a satisfactory white screen. These screens can be processed by more-or-less standard methods, and produce tubes having ordinarily good life. They appear to have distinct commercial possibilities.

[1] K. Kamm, "Annalen der Physik," Vol. 30, pp. 333-353; October, 1937.

It is characteristic of the sulphides that they are much more sensitive to grinding than are the silicates, probably because their crystals are normally much larger than those of the silicates. Consequently, grinding of the sulphides beyond a certain point involves actual destruction of the crystals, while grinding of the silicates involves simply a separation of the crystals without destruction. For this reason it is necessary to crystallize the sulphides at comparatively low temperatures to insure that their crystal sizes will be small when they are to be mixed with silicates; otherwise, the difference in particle size will cause the two materials to segregate during spraying and produce a dappled screen.

This tendency of the sulphides to form coarse crystals, and the difficulty of grinding them satisfactorily led to the use of binders to hold them to the glass bulb walls. The alkali silicates, borates, and phosphoric acid were used for this purpose. They tend to be very hygroscopic and, if not completely freed from water during the exhaust process, they impair tube life. Even if the amount of vapor released should not be great enough to impair seriously the cathode surface, the presence of gas in the tube permits the screen to be bombarded with negative ions which produce a dark spot on the screen when magnetic deflection is used. The sulphides appear to be somewhat more sensitive than the silicates to damage from this source.

Characteristic curves of the phosphorescent decay of zinc sulphide and willemite are shown in Figs. 8a and 8b, respectively. The long persistence shown for the sulphide phosphor does not appear in the sulphides discussed above; the latter have a rapid decay to a low level, and any phosphorescence of long duration is of such low intensity as to be negligible. The silicates and silicate derivatives have a decay characteristic which is peculiarly well suited to television work, in that the intensity decreases fairly slowly in a way which tends to minimize flicker and ocular fatigue. The rapid decay from a high initial level which is shown by the sulphides tends to be tiresome during long periods of observation. The mixed silicate-sulphide screen usually has enough "carry-over" from the silicate component to prevent this effect.

A long-continuing phosphorescence in zinc sulphide may be produced by the presence of copper activator (in the proportion of about 10^{-6}) and the use of fluxes. When zinc sulphide is thus processed, a highly phosphorescent material is obtained. It is useful as a screen for the temporary recording of slow phenomena which require several seconds for their completion. The trace, which remains visible for a minute or more, presents a complete picture of the whole set of events. To excite the screen to its maximum phosphorescence requires the application of a certain amount of energy; if the beam is passed very rapidly over the screen, very little phos-

phorescence will be excited.[2] This type of screen is thus well suited for use in audio-frequency curve tracers, electrocardiographs, and similar instruments, but is not to be applied as a means of recording very high-speed phenomena, such as waves on transmission lines or switching

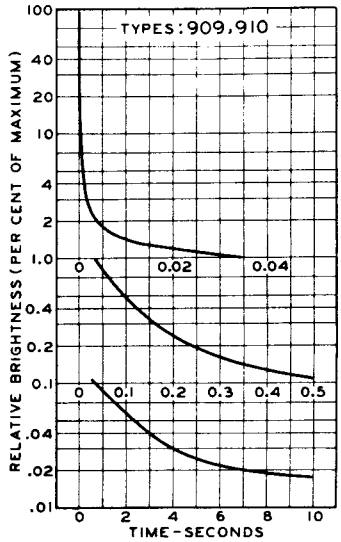


Fig. 8a - Persistence characteristic of zinc sulphide phosphor No.2.

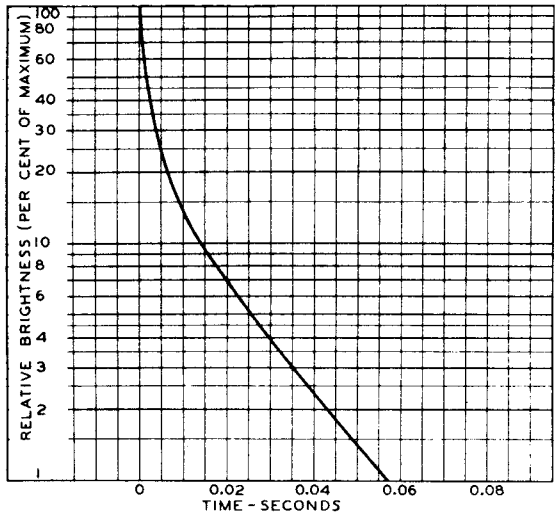


Fig. 8b - Persistence characteristic of willemite phosphor No.1.

transients. The long-persistence sulphide is easily excited to phosphorescence by visible light, such as is produced by an ordinary incandescent lamp; but infra-red radiation, which can be obtained from a low-temperature carbon lamp enclosed in an infra-red transmitting bulb which does not pass visible light, will quench the phosphorescence. Such a lamp is a convenient accessory to wipe off traces of phenomena previously observed, or to destroy background phosphorescence which may exist if the cathode-ray tube has been exposed to light just before it is to be used.

Calcium tungstate is a material which is useful because of its exceedingly short persistence (less than 30 microseconds). It is similar to the silicates in its ability to withstand heat treatment during processing and in general stability. It is probably unactivated and luminesces best when it is purest (although Swindells[3] has found that it must be lead-activated for x-ray work). It may be prepared conveniently by precipitation from a solution of a soluble tungstate, such as sodium tungstate. The precipitate, because of the extreme insolubility of the calcium tungstate, is not crystalline and is not appreciably luminescent. It may be crystallized by firing at a high temperature (1000°C or higher) and then shows an emission which appears blue to the eye, and actually extends out into the ultra-violet (Fig. 9). Its energy efficiency is good, but because it emits in a region where the eye

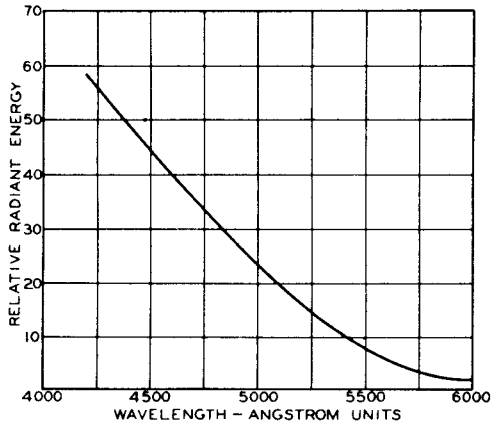


Fig. 9 - Spectral energy characteristic of phosphor No.5.

is very insensitive, its luminous efficiency is only about 0.1 candlepower per watt under ordinary operating conditions. Because commercial photographic film is very sensitive to blue and near ultra-violet light, traces on a calcium-tungstate screen may be photographed very well. Since the persistence is so short, the beam may be deflected in one direction only, and the time axis may be supplied by moving the recording film in a direction at right angles to the motion of the trace.

The three types of materials, i.e., silicates, sulphides, and tungstates, are the chief ones used for cathode-ray tubes at the present time. They can all be excited to some extent by

[2] M. von Ardenne, "Zeitschrift für Physik," Vol. 105, pp. 193-201; 1937.

[3] F. E. Swindells, "Luminescence of Alkaline Earth Tungstates Containing Lead," Jour. O.S.A., Vol. 23, p. 129; April, 1933.

ultra-violet emission of suitable wavelength, which may in particular instances be very short. In general, the optimum amount of activator for cathode-ray excitation is not identical with the optimum for ultra-violet excitation. The best method of preparation may differ for each. At the present time, there is a great deal of interest in the possibilities of fluorescent linings for gas-discharge lamps, in which very large amounts of ultra-violet light are produced. The literature on luminescence excited by ultra-violet light is much more voluminous and complete than that on the luminescence excited by cathode rays, partly because the technique of production of ultra-violet light is somewhat older. No attempt will be made here to cover this field; a very comprehensive survey may be found in the appropriate volumes of the "Handbuch der Experimentalphysik" and of the "Handbuch der Physik."

Candoluminescence is best exemplified by the ordinary Welsbach gas mantle in which a thorium-oxide base containing a small amount of cerium oxide is caused to emit light by being heated in a gas flame to a temperature which, while far above room temperature, is believed to be below that at which a black body would emit similar light. This phenomenon is not well understood. It has been found that if a sufficiently porous thorium-oxide body is prepared, it may be brought to incandescence in vacuo by the heating effect of an electron beam. To produce this effect, no cerium-oxide addition is required. This phenomenon appears to be a case of simple incandescence, and not of so-called candoluminescence The minimum energy input to produce light emission from bodies sufficiently dense not to sinter under bombardment is greater than cathodoluminescent materials can receive without undergoing decomposition. One difficulty in the preparation of such a screen is the comparative fragility of a sufficiently porous mass. If the mass is not porous, heat is lost by conduction to the adjacent portions of the screen to such an extent that the beam intensities at present available do not suffice to produce the requisite temperature.

REFERENCES

T. B. Brown, "Brightness of Cathodo-Luminescence at Low Current Densities and Low Voltages," Jour. O.S.A., pp. 186-192; May, 1937.

Engle and Hopkins, "Studies in Luminescence," Jour. O.S.A., pp. 599-615; December, 1925.

J. L. Gallup, "Vacuum Cell Luminescence Microscope," Jour. O.S.A., pp. 213-215; May, 1936.

Lenard, Schmidt, and Tomaschek, "Handbuch der Experimentalphysik," Vol. 23, Part I, Akademische Verlagsgesellschaft, Julius Klinghardt, Leipzig.

H. W. Leverenz, "Problems Concerning the Production of Cathode-Ray Tube Screens," Jour. O.S.A., pp. 25-35; January, 1937.

L. Levy and D. W. West, "Fluorescent Screens for Cathode-Ray Tubes for Television and Other Purposes," Jour. I.E.E., pp. 11-19; July, 1936.

W. B. Nottingham, "Potential and Luminescence of Insulated Willemite Cathode-Ray Screens," Physical Review, p. 591; April, 1937.

T. B. Perkins and H. W. Kaufmann, "Luminescent Materials for Cathode-Ray Tubes," Proc. I.R.E., pp. 1324-1333; November, 1935.

Pringsheim, "Handbuch der Physik," Vol. 23.

J. T. Randall, "Luminescence and Its Applications," Electrician, p. 118; February 5, 1937.

Lecture 6

CONTACT POTENTIAL, PUMPS, AND GETTERS

E. A. Lederer

Part I - Contact Potential

The justification for the existence of contact potential has already been discussed in Lecture 1. In that lecture we expressed contact potential in terms of the characteristic barrier action on the surface of metals which tends to prevent the egress of all but a few very fast electrons from a metal.

In order to convince ourselves that contact electromotive force or contact potential exists, we need not use any mathematics. Let us assume that two dissimilar metals A and B (Fig. 1) are joined at one place, are surrounded by a common vacuum, and are maintained at such temperature that both metals become electron emissive. In order to satisfy this latter condition, it is only necessary that the temperature of the system be slightly higher than absolute zero. Room temperature, for example, may suffice to dislodge a few electrons per unit time from each of the two metals.

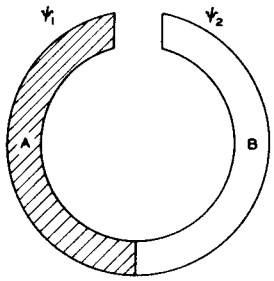


Fig. 1

Let us denote the work function, or the latent heat of electron evaporation, of metal A by ψ_1 and that of metal B by ψ_2; and let us assume that $\psi_1 > \psi_2$. Then, at a temperature conducive to electron emission, the electron current emerging from B and flowing through the gap to A would be larger than the electron current emerging from A and arriving at B. The net result would be an uncompensated electron current flowing continuously through the system. In other words, we would obtain electric power for nothing. This condition, as we know, is incompatible with the second law of thermodynamics. The electron currents received and emitted by A and B must be exactly equal. The reason for this equality is found in the difference of potential between the two metals at the gap. It is well to remember that the potential difference is sustained by the egress of electrons from both metals, and that the metal loosing more electrons in unit time by reason of its lower work function, will become positive with respect to the other metal.

Theory and experimental evidence lead to the following relation for contact potential (V):

$$V = \psi_1 - \psi_2 + P$$

The quantity P in the formula expresses the so-called Peltier effect and is a heat loss produced by the passage of current through the junction of the two dissimilar metals (Fig. 1). However, the magnitude of this effect is of an entirely different order than that of ψ_1 or ψ_2; in fact, it is so small that we are justified in neglecting it. We have become aware that minute surface contaminations exert a profound influence upon the work function and we, therefore, conclude that the contact potential is similarly affected. Our conclusion is indeed justified. From past experience we know that an oxide-coated cathode, for example, loses barium metal during activation and during life. The barium metal given off by the cathode condenses on the other tube electrodes which are cooler than the cathode. When the contact-potential difference between cathode and control grid is measured, it is not astonishing to find large variations which may amount to changes of 1 to 2 volts. These changes affect the actual grid bias of the tube and thereby interfere with its operation.

If we wish to measure contact-potential differences between cathode and control grid, we make use of the simple circuit given in Fig. 2. By means of the potentiometer P, a positive or negative potential E_r, with respect to the indirectly heated cathode, can be impressed on the control grid, and the grid current can be measured

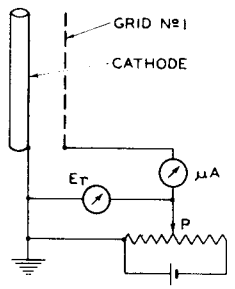


Fig. 2

with the microammeter. The measurement becomes more difficult with a filamentary cathode because of the voltage drop along the filament. In order to carry out the measurement with a filamentary cathode, a rotary commutator is used to advantage. It interrupts the power supplied to the filament during the moment the grid circuit, as shown in Fig. 2, is connected to one end of the filament.

With no power supplied to the filament the potential between it and the grid is not affected by a voltage drop. By means of the rotary commutator it is possible to maintain the filament at the desired temperature and to carry out instantaneous measurements as if the former were a unipotential cathode. It is evident that the frequency of commutation must increase with decreasing heat capacity (mass) of the filament.

If the electrons were to emerge from the emitter with no initial velocity, the current in the grid (collector) circuit would be zero at zero grid potential, or with $E_r = 0$ applied between cathode and grid. We know, however, that the emitted electrons emerge with initial velocities ranging between zero and a certain maximum velocity, the magnitude of which depends upon the temperature of the emitter. The result is that electrons can reach the grid or collector even when the grid is maintained at a negative or retarding potential E_r with respect to the cathode. We also know that the emitted electrons conform with Maxwell's velocity distribution (most probable distribution). The fraction n/n_s of the electrons capable of moving against a retarding potential E_r is given by Boltzmann's equation:

$$\frac{n}{n_s} = \varepsilon^{-\frac{E_r e}{kT}} \qquad (1)$$

In this equation n_s is the total number of emitted electrons, E_r the retarding potential, e the electronic charge, k the Boltzmann constant, T the temperature of the emitter in degrees K, and ε the base of the natural logarithms. Since a number of electrons per unit time flowing past a given point constitutes an electric current, equation (1) can be expressed by

$$\frac{I}{I_s} = \varepsilon^{-\frac{E_r e}{kT}} \qquad (2)$$

where I is the current to the collector with retarding potential E_r applied, and I_s is the saturation current at the temperature T in °K.

Taking the logarithm of equation (2) and plotting $\log_{10} I$ against retarding potential E_r, we obtain a line ABCD as shown in Fig. 3. From A to B the line is straight and has a slope with respect to the E_r axis expressed by

$$\frac{e}{kT} \log_{10} \varepsilon$$

The bend BC in the line occurs when I becomes nearly equal to I_s. Beyond the point C, all of the electrons emitted by the cathode are intercepted by the collector. If we extend both straight portions of the line beyond the curved section BC, they meet at S. Point S is at zero retarding potential if no potential other than E_r is acting. If a contact-potential difference exists between cathode and collector, the sloping section AB is displaced parallel to its former position, for example into position A'B'. The horizontal displacement along the E_r axis is a measure of the existing contact-potential difference.

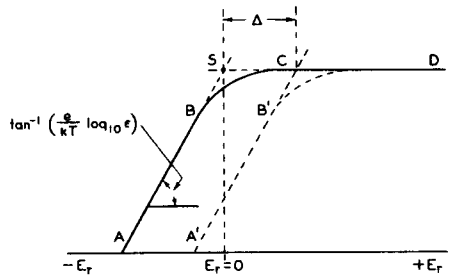


Fig. 3

The practical application of contact-potential measurements are important to the design engineer and the factory engineer because by such measurements changes in the surface conditions of grids can be determined. Furthermore, since the initial velocities of the emitted electrons are a function of the brightness temperature of the coating, and since the core material affects the brightness temperature, it follows that the core material affects the contact potential.

The curves of Fig. 4 show a practical application of measurements of retarding potentials. Retarding potentials are plotted against cathode brightness temperature. The tube under observation was a 6K7 using batalum getter and bright nickel plates. One object of the investigation was to determine what happened to the retarding potential during aging, and another was to determine whether or not barium from the getter activated the cathode.

Curve 1 shows measurements taken on the 6K7 before the getter was flashed. As a result, the presence of gas and the unstable condition of the cathode were suspected of affecting the readings. Curve 2 was taken after getter flashing. The gas current taken by the conventional method was less than 1×10^{-7} amperes. Curve 3 was taken after subjecting the cathode to a 220 per cent overvoltage heating (hot shot) of 1 minute; curve 4, after the same overvoltage for 2 minutes; and curve 5, after the same overvoltage for 3 minutes. The shift in these curves is considered to be due to the grid becoming progressively more electropositive as barium metal from the cathode was evaporated onto it. The very slight effect of additional hot shots after the initial one-minute cathode treatment indicates that enough barium was obtained for almost complete grid coverage in that time. The fact that only barium metal evaporates from an oxide-coated cathode has been mentioned in the section on oxide-coated cathodes.

In order that gas contamination of the other

electrodes might be studied, gas was released in the 6K7 by the following procedure. The cathode was maintained at 1100°K, grid No.1 was allowed to float, and a positive potential sufficient to give a cathode current of 40 milliamperes for 3 minutes was applied to the other electrodes. After this treatment, curve 6 was taken. It indicates that barium metal was removed from grid No.1 by chemical combination with the gas previously liberated.

From this description and the curves of Fig. 4, we can make the following deductions: (1) no barium metal from the getter reaches the cathode or the No.1 grid; (2) a hot shot of one minute furnishes enough barium metal from the cathode to coat grid No.1 at least with a monatomic layer; (3) successive additional hot shots merely produce more barium metal which may even out bare spots in the already existing barium layer on grid No.1; and (4) considerable gas is produced when heating or bombarding the electrodes, other than grid No.1. Although the nature of the gas is not known, we do know that it combines with barium metal and very probably is CO and H_2.

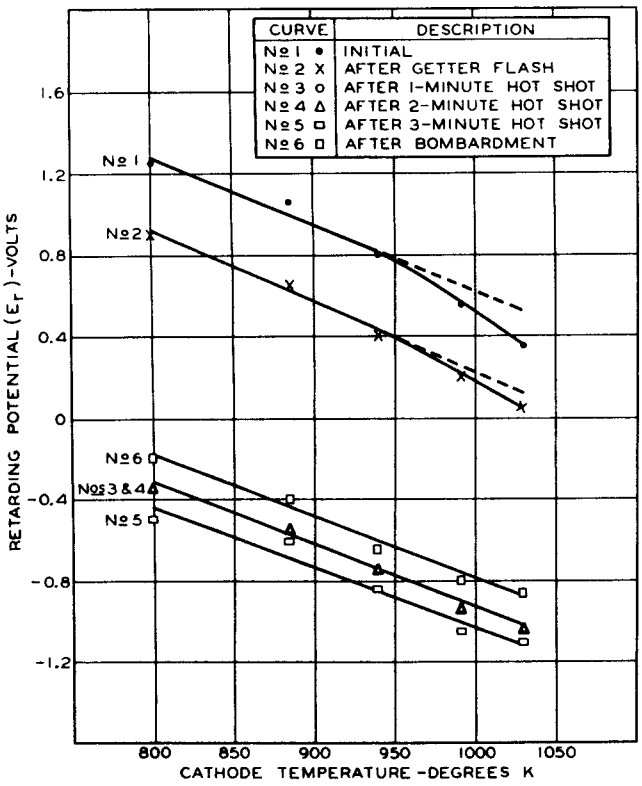


Fig. 4

The investigation of retarding potentials was continued by making measurements on a 6K7-G taken from stock. This tube had been aged previous to the measurements. The results of these measurements are shown in Fig. 5. It will be noted that all curves are practically straight and the slope appears almost constant. At first, the constancy of the slope and the straightness of the lines were surprising. However, the explanation of these facts rests again on equation (2), i.e., the relation between initial velocities and cathode temperature. After changing to common logarithms and substituting numerical values for Boltzmann's constant, the electronic charge, and changing from e.s.u. to volts, we can rewrite equation (2) as

$$\log \frac{i}{i_s} = -5050 \frac{E_r}{T} \qquad (2a)$$

In the equation, i is the collector or grid current (see Fig. 2) at the retarding potential E_r, and i_s is the saturation current which the cathode is capable of delivering at temperature T°K.

From measurements taken on a 6K7 tube, using an oxide-coated cathode with an emitting area of 0.718 sq cm, and determining E_r for i fixed arbitrarily at 5.0×10^{-7} amperes, we obtain the data in Table I.

Table I

T* °K	i_s amperes	i amperes	E_r volts	$\frac{i}{i_s}$	$\log_{10} \frac{i}{i_s}$	$\Delta E_r/100°C$
1000	48.8	5.0×10^{-7}	1.780	1.024×10^{-8}	2.010 - 10	
						-0.367
900	4.31	"	1.413	1.16×10^{-7}	3.065 - 10	
						-0.347
800	0.269	"	1.066	1.86×10^{-6}	4.270 - 10	
						-0.352
700	7.18×10^{-3}	"	0.714	6.96×10^{-5}	5.843 - 10	
						-0.368
600	39.5×10^{-6}	"	0.346	1.265×10^{-2}	8.102 - 10	

Average = -0.359

* Cathode temperature was measured in the middle of the cathode, i.e., at its hottest part. (Brightness temperature is used in both the calculated and measured values of the slope of the lines.)

From Table I we get the slope of the lines of Fig. 5 expressed in $\Delta E_T/100°C$ and from the constancy of this value, barring experimental errors, we have proof that the lines are straight

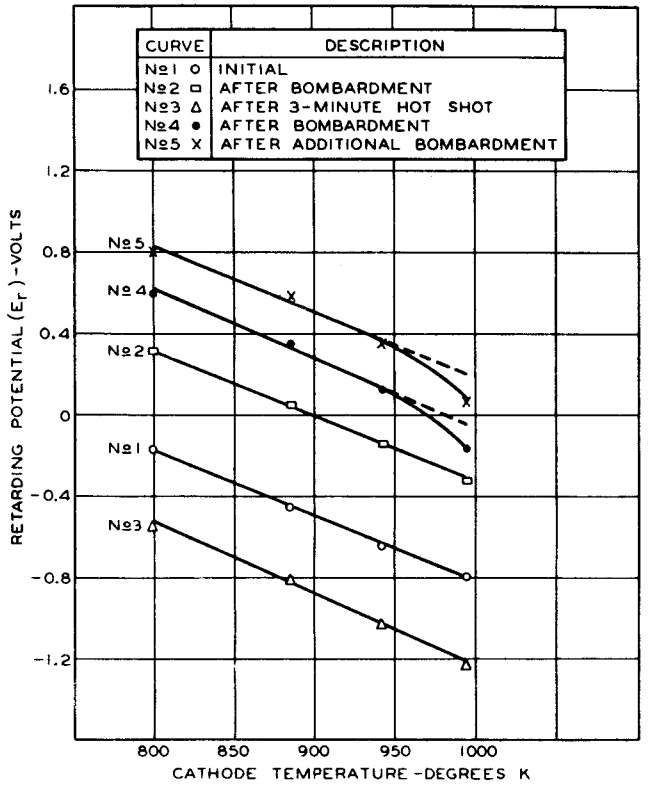


Fig. 5

over the charted range. It should be remembered, however, that $\Delta E_T/100°C$ is limited to the particular electrode dimensions and that brightness temperature was used for convenience while for a rigorous determination of $\Delta E_T/100°C$, true temperature should have been used.

REFERENCES

S. Dushman, "Thermionic Emission," Rev. of Mod. Phys., Vol. 2, p. 454; October, 1930.

R. M. Bowie, "This Matter of Contact Potential," Proc. I.R.E., Vol. 24, p. 1501; November, 1936.

Epstein, Ber. d. D. Phys. Ges., Vol. 21, p. 85; 1919.

T. C. Fry, "The Thermionic Current between Parallel Plane Electrodes; Velocities of Emission Distributed According to Maxwell's Law," Phys. Rev., Vol. 17, p. 441; 1921.

I. Langmuir, "The Effect of Space Charge and Initial Velocities on the Potential Distribution and Thermionic Current between Parallel Plane Electrodes," Phys. Rev., Vol. 21, p. 419; 1923.

O. W. Richardson, "Emission of Electricity from Hot Bodies." London, 1921.

R. H. Fowler, "Statistical Mechanics," Second Edition, The Cambridge University Press, The Macmillan Company, New York, 1936. p. 418, ff.

Part II - Vacuum Pumps, Vacuum Measurements, and Vacuum Technique

INTRODUCTION

Reviewing briefly the historical development of vacuum pumps will lead logically to a classification of pumps with respect to their principle of operation. The first vacuum pump, invented by Otto von Guericke (1602-1686) was a mechanical pump of the piston type. While this invention was scientifically of far-reaching importance, it remained practically unused for more than 100 years. The first practical use of the piston pump was made by James Watt (1736-1819) who improved it considerably. Watt used the pump to exhaust steam condensers by means of which he improved the efficiency of his steam engine.

Toward the middle of the past century after the scientific world received stimulation by the discoveries of Faraday, the need for better vacua became very important. Realizing that mechanical piston pumps could not be improved much further, Toepler and Geissler, independently of each other, substituted for the fabricated piston one made of mercury. In this simple manner, an absolutely vacuum-tight piston was made available; and furthermore, it did not require lubrication. Pumps of this type not only made possible the discovery of cathode rays, canal rays, and x-rays, but also made possible the beginning of the incandescent lamp industry. As it grew, it constantly demanded better and faster pumps. A notable improvement of the mercury piston pump was made by Gaede in 1905, and can be designated as a rotary mercury piston pump. It has found extensive use in the lamp industry. The rotary oil pump is also a piston pump but of improved design and is familiar to us as the "Cenco" pump or the "Kinney" pump.

The most fundamental advances in pump design are of rather recent origin. In these, the piston principle is abandoned altogether and advantage is taken of the properties of gases as revealed by the kinetic theory. The new principle of operation is typified by two kinds of high vacuum pumps, which are: (1) the rotary molecular pump (Gaede); and (2) the mercury diffusion pump (Langmuir). Although different in their mechanism, both types must be backed up by a fore vacuum which even today is produced by oil-lubricated piston pumps.

THE ROTARY MOLECULAR PUMP

We consider a gas as an assembly of a large

number of very small but perfectly elastic particles in random motion. They collide with each other and impact upon and rebound from the walls of the gas container. If V denotes the volume of the container, n the number of the particles, m the mass of one particle, c the mean velocity of the particles, and p the gas pressure per cm^2, then

$$p = \frac{nmc^2}{3V}$$

If one of the boundaries of the vessel is given a translatory motion in its own plane, all the gas molecules rebounding from that boundary will receive a velocity component parallel to the velocity of the boundary. Consider a channel of length L and height h in a solid wall as shown in Fig. 6, and denote by c the velocity of the

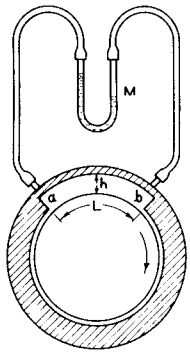


Fig. 6

moving boundary, and by v the viscosity of the gas. The difference in pressure Δp, between a and b as measured by the manometer M is given by:

$$\Delta p = \frac{6vcL}{h^2}$$

We note at a glance that the longer we make L and the larger we make c, the greater the difference in pressure. While Gaede was the inventor of the molecular pump, Holweck improved this pump considerably so that it could be used commercially for exhausting demountable high-power transmitting tubes. The Holweck pump is shown schematically in Fig. 7.

The diameter of the revolving brass cylinder is 15 cm, its length is 22 cm, and the clearing between cylinder and housing is approximately 0.001". The cylinder revolves at a speed of 4000 to 4500 rpm. The pump can lower the pressure in a 5-liter vessel from 0.1 mm to 0.001 mm in 10 seconds. The power input to the motor when the fore vacuum has been established is only 10 watts. One great advantage of the pump is that no liquid-air trap is required. One of these pumps was connected to the last two positions (previous to tipping off on an old 24-head exhaust machine and operated for over one-half year without any maintenance. A disadvantage of the Holweck pump is that it is expensive. Perhaps someone will design an equally good but less expensive pump.

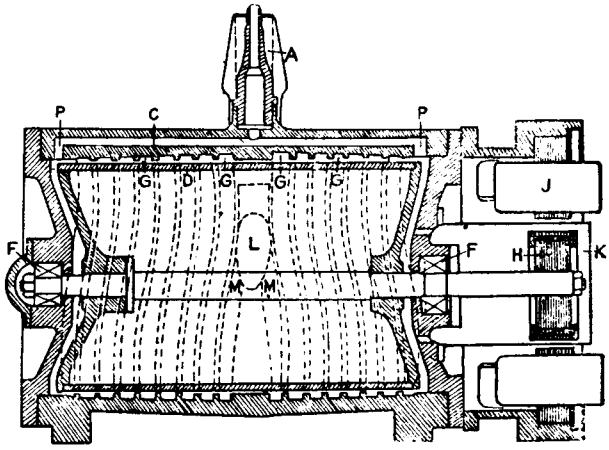


Fig. 7 - Holweck's molecular pump.
(From Comptes Rendus, Vol. 177, p. 44; 1923)

THE MERCURY DIFFUSION PUMP

Consider a porous plug P, a trap T, and a reservoir V, as shown in Fig. 8. When steam is blown through the tube S, it will diffuse through P into T and V, while any gas contained in T and V will diffuse simultaneously through P into S and be carried away. Therefore, the concentration of gas in T and V will decrease with time while the

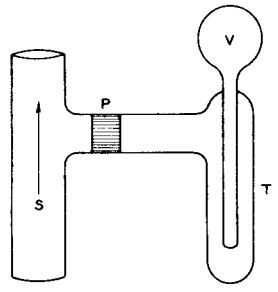


Fig. 8

concentration of steam will increase. If, then, we cool T so as to freeze out the steam, V will be eventually exhausted. It is apparent that the action of such a diffusion pump is slow. It occurred to Langmuir that the important feature was not the porous plug, but the condensation of the steam. Langmuir's condensation pump is shown schematically in Fig. 9.

Mercury vapor moves in the direction of the arrow through A into a wider tube B surrounded by a cooling jacket J. At atmospheric pressure, mercury molecules emerging from A would collide with gas molecules in B and mercury vapor would therefore diffuse into the vessel V. No such diffusion will take place if the mean free path of the

molecules of mercury vapor is great as compared to the dimension of the annular space between A and B, for then the molecules leaving A will travel in straight lines until they strike the wall B and condense there. The necessary conditions can easily be established by connecting B to a fore vacuum. Thus, a downward stream of fast-moving mercury atoms is established throughout A and B, and gas molecules diffusing from V into this stream suffer collisions with the mercury atoms and receive a velocity component which carries them downward and prevents their return. They are then removed by the fore pump.

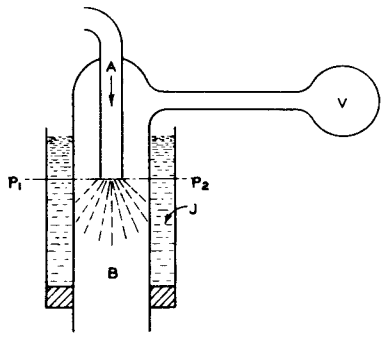


Fig. 9 - Langmuir's condensation pump (schematic).

Because this type of pump has no moving parts, and since it can be built of glass or metal, it is now widely used in vacuum technique. In recent years, suitable organic compounds, such as Apiezon oil, dibutyl-phthallate, and ethyl-hexyl-phthallate, having a low vapor pressure at room temperature, were found which could be used as a substitute for the mercury vapor. While mercury diffusion, or condensation pumps as they are frequently called, require a liquid-air trap between the pump and the vessel which is to be exhausted, oil-condensation pumps operate satisfactorily using a water-cooled trap and for highest vacuum can be made self-fractionating.[1] Modifications of the condensation pump are numerous and development work is still in progress.

It has been found that the speed of these modern pumps can be made as great as desired up to several hundred liters per second, the only precaution being that for the highest speeds the fore-pump pressure must be correspondingly decreased. The upper limit to the practical speed[2] is set by the dimensions of the connecting tubing.

COMPARISON OF PUMPS

The information in Table II is presented to

Table II

Type of Pump	Input Watts	Fore Vacuum Required mm of Hg	Vacuum Attained mm of Hg	Exhaust Speed Liters/sec.
1. Mercury with liquid-air trap:				
G. E. Metal	365	4.	4×10^{-7}	3
Glass (RCA make)- 1 jet - long divergent nozzle[2]	100	0.04	1×10^{-7}	50*
2. Ethyl-Hexyl-Phthallate with water-cooled trap:				
Glass - 3 jet- fractionating	175	0.04*	5×10^{-8}	15
Metal - 3 jet- fractionating	300	0.10	2×10^{-7}	240
3. Holweck Rotary Molecular	100 (motor)	15.	7×10^{-7}	4.5

* Estimated

[1] K. C. D. Hickman, "Trends in Design of Fractionating Pumps," Jour. of Applied Physics, Vol. 2, No.5, pp. 303-313; May, 1940.

[2] T. E. Phipps, H. M. Tenney, O. C. Simpson, and M. J. Copley, "A Study of the Speed of Divergent Nozzle Pumps," Review of Scientific Instruments, Vol. 6, pp. 265-267; 1935.

indicate the general operating characteristics of some available diffusion pumps, and of the older Holweck pump.

SPEED OF EXHAUST

The rate at which a given vessel can be exhausted depends on the diameter d of the connecting tubing, its length L, the difference in pressure between that in the vessel p_2 and that at the pump end of the tubing p_1, and on the molecular weight m of the gas. The quantity of gas q in cubic centimeters at 1 bar pressure which can be exhausted per second, is given by

$$q = 3809\sqrt{\frac{T}{m}}\,\frac{d^3}{L}(p_2 - p_1)$$

where d and L are expressed in centimeters, p_1 and p_2 in bars, and T in °K.

VACUUM MEASUREMENTS

We have not been able to produce an absolute vacuum. By means of vacuum pumps, we can only attenuate the gas in a container and approach a certain limit. It is not convenient to measure low gas pressures in fractions of an atmosphere. Instead, we measure them in terms of another unit called the "bar" which is equal to a pressure of one dyne per sq cm. The bar is sometimes called the "barye" or the "microbar." While the bar is the scientific unit of pressure, we often speak of gas pressure in terms of the height of the liquid column it will support. The liquid usually used is mercury and so we indicate pressures in centimeters, millimeters, or microns (thousandths of a millimeter) of mercury. A conversion table follows:

1 micron	=	1.342 bars
1 mm Hg	=	1342 bars
750 mm Hg	=	1006500 bars

The micron is extensively used because certain manometers are conveniently calibrated in terms of millimeters of mercury pressure.

In order to gain some conception of a pressure of one micron, let us consider the following case. A gram atom of any gas at 0°C and 750 mm pressure takes up a volume of 22.4 liters, or it can be accommodated in a sphere approximately one foot in diameter. There are 6.06×10^{23} molecules in a gram atom or 2.7×10^{19} molecules per cu cm. If this sphere were expanded to a diameter of 100 feet, the pressure would be approximately one micron and there would be 2.7×10^{-13} molecules per cu cm. A pressure of 10^{-3} microns (10^{-6} mm) can be easily reached with modern vacuum pumps and at this pressure a molecule of air would on the average travel a distance of approximately 250 feet before meeting another molecule.

MANOMETERS

Many manometers for measuring low pressures have been designed. There is one, the "McLeod gauge," with which the gas pressure can be derived from simple measurements of volumes and dimensions. Others make use of some properties of gases which vary with the number of molecules per unit volume but these manometers must be calibrated by comparison with another manometer, usually a McLeod gauge.

1. The McLeod Gauge

The McLeod gauge is by far the simplest manometer and depends on the applicability of Boyle's law. It consists of a glass vessel of volume V_0 to which is attached a glass tube C of small inside diameter, as shown in Fig. 10. The

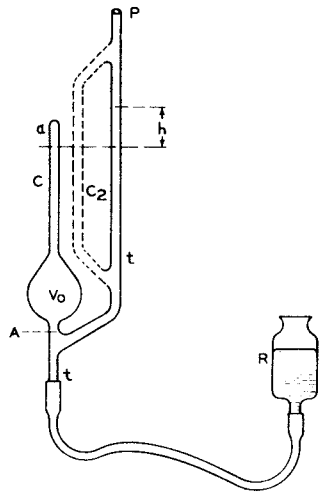


Fig. 10 – The McLeod Gauge.

glass vessel is connected as shown to a mercury reservoir R, and also at P to the system in which the pressure is to be determined. If we let V equal the volume of gas in C plus that in the vessel above the dotted line A, and X equal the unknown pressure in the device, we have:

$$X\,V = \text{constant}$$

Raising the reservoir R causes the mercury to ascend in tube t, and when it has reached line A, the volume of gas (V) is trapped in the vessel and tube by the mercury. On raising R further, the gas is finally compressed into the small volume a at the top of C under a pressure h, the difference between the mercury level in C and t. Knowing the diameter of C, we can easily determine V, and h can be read with a scale. Therefore,

$$a\,h = \text{constant} = X\,V$$

and the gas pressure in the system is

$$X = \frac{a\,h}{V}$$

In order that low pressures may be read with the McLeod gauge, C is a capillary tube. Then, to eliminate errors in the reading of h, t has to be made with the same inside diameter as C. But with this arrangement, the speed of exhausting the gauge is cut down. Therefore, another capillary tube C_2 is attached to t, as shown by the dotted lines. Pressure h is then indicated by the difference between the mercury level in C and C_2.

The McLeod gauge is very satisfactory for non-condensible gases. The introduction of water or oil vapors, however, will cause great errors. Therefore, the gauge should be so arranged as to permit a thorough bakeout from time to time. Furthermore, a liquid-air trap should be inserted between gauge and vacuum system. The gauge is calibrated by "weighing it out" with mercury. From the weight of the mercury and its density at the temperature during calibration, the volume of V_0 and that of C is computed. Due to irregularities in the diameter of capillary tubing, it is worth while to determine its average diameter by measuring the length of a mercury column at different places along the entire length of C. C_2 and C should be made from the same piece of tubing. Before calibration, the gauge is thoroughly washed with chromic acid, then with distilled water, and dried in an oven. It should be remembered that a dirty McLeod gauge is worthless.

2. The Pirani Gauge

Suppose a wire is heated in a gas-filled envelope by passing an electric current through it. After temperature equilibrium is reached, the heat loss of the wire is due to radiation and to convection through the gas. At low wire temperatures, most of the heat loss is due to convection. Suppose, now, that the envelope is connected to a vacuum pump. As the gas surrounding the wire is pumped out, the heat loss decreases and the temperature of the wire increases. The wire temperature can be determined by measuring its electrical resistance. The foregoing is a simple explanation of the "Pirani gauge," the electrical diagram of which is shown in Fig. 11.

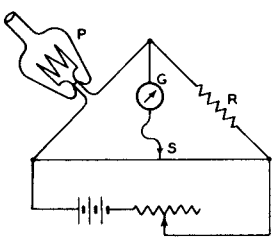


Fig. 11 - Electrical diagram of the Pirani gauge.

The Pirani tube P which is attached to the vacuum system contains a long, fine platinum wire or ribbon and forms one arm in a Wheatstone bridge. R is a resistance equal to that of the wire in the tube with high vacuum under the conditions of operation. S is a slide wire, and G is a galvanometer. The current flowing through P and R heats the platinum wire to about 300°C. In order to make the arrangement less sensitive to changes of room temperature, R can be replaced by an exhausted tube identical with P. In this case, it is also advisable to arrange both tubes close together and immerse them in a bath of constant temperature.

The gauge can be made very sensitive to small pressure changes but its useful range is not very large. For example, a gauge which permits measurements by large galvanometer deflections between 1 and 150 microns, has very rapidly decreasing sensitivity as the pressure is increased above 150 microns. An advantage of the Pirani gauge is that its continuous indications show rapid changes of pressure. This feature is not found in the McLeod gauge which only permits pressure checks from time to time. A disadvantage of the Pirani gauge is that it must be calibrated against a McLeod gauge for every gas. Thus, accurate measurements with a Pirani gauge require a knowledge of the gas present in the system.

3. The Ionization Gauge

An extremely sensitive gauge for measuring the lowest pressures is the ionization gauge. It has the limitation given above for the Pirani gauge but its simplicity and sensitivity make it a very useful instrument. It consists of a triode (preferably one with a cylindral plate and a pure tungsten filament) sealed onto the vacuum system. There are two methods of using this gauge.

In method A, the grid is maintained at a negative bias, while the plate is at a positive potential. Positive ions produced by collision between electrons and gas molecules are collected by the grid. The ratio of grid current I_c to plate current I_b is proportional to the gas pressure. Let us consider a specific case. Assume that the triode has a cylindrical plate with a diameter of 1 cm, and that the ratio of I_c/I_b = 1/1000. This ratio indicates that among 1000 electrons collected by the plate only one electron impinges upon a gas molecule; and that the mean free path of the electrons is 1000 times as large as the actually covered distance, which according to our assumption is 1/2 cm. The mean free path at atmospheric pressure is about 10^{-5} cm and increases proportionally with decreasing pressure. Thus, a mean free path of 1000 x 1/2 cm corresponds roughly to a pressure of 2×10^{-8} atmospheres or 1.5×10^{-5} mm.

An empirical formula for use in determining pressure with the ionization gauge has been evaluated by Moeller, and is as follows:

$$p = K \frac{I_c}{I_b} d$$

in which p is the pressure, K is a constant, and d is the grid to plate distance.

Method B is still more sensitive than method A. In method B, a positive potential is applied to the grid, while a negative bias is applied to the plate. This arrangement is similar to that used for production of Barkhausen-Kurz oscillations. The better sensitivity may be due to the fact that electrons oscillate around the grid and thereby increase the probability of ionization. With method B it is impossible to estimate the magnitude of the gas pressure from the ratio I_c/I_b.

Calibration of ionization gauges for every particular gas with a given tube electrode arrangement should be made with a McLeod gauge and the values so determined can then be extrapolated to lower pressures. See curves in Fig. 12.

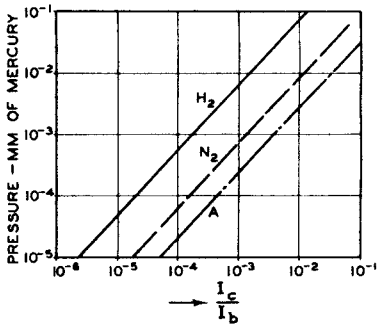


Fig. 12

There are still other vacuum gauges like the quartz-fibre gauge, the Kundson gauge, and the thermocouple gauge, but their use is limited to laboratory measurements because they require special precautions or they do not offer any particular advantages over the gauges previously described.

In all vacuum work there is one rule which can not be disregarded or failure will result. The rule is, "Cleanliness is next to godliness." We have already discussed how to clean a McLeod gauge. The same rules apply to the cleaning of all vacuum apparatus. Rubber connections and wax seals should only be used where absolutely necessary. Stopcocks should be used sparingly and great care must be exercised in the selection of the stopcock lubricant. Rubber tubing as used for McLeod gauges should be boiled for hours in caustic soda, scrubbed inside with a brush to remove sulfur dust and talcum, rinsed with water, and dried with air. Rubber tubing to be used on a vacuum system should not come in contact with carbon tetrachloride, alcohol, ether, or oil. An exception to this rule is found in the permissible use of castor oil when it is applied sparingly to exhaust ports. Our practice of manufacturing tubes on machines using rubber connections is feasible only because we do the final exhaust, after sealing off, with the getter.

REFERENCES

S. Dushman, "Production and Measurement of High Vacua," published by the General Electric Review, 1922.

S. Dushman, "Recent Advances in the Production and Measurement of High Vacua," Jour. Franklin Inst., Vol. 211, No.6, pp. 689-750; June, 1931.

L. Dunoyer and J. H. Smith, "Vacuum Practice," D. Van Nostrand Company, New York, 1926.

F. H. Newman, "Production and Measurement of Low Pressures," D. Van Nostrand Company, New York, 1925.

Part III - Getters

INTRODUCTION AND HISTORY

The term "getter" was coined years ago, probably by engineers of the Edison Lamp Works formerly located here at Harrison. A getter is a chemical pump arranged in a vacuum or gas-discharge device for the purpose of removing undesirable gas residues. In radio receiving tubes we use getters in order to reduce gas residues initially to a tolerable level. This function of the getter we have already mentioned in connection with speed of exhaust. Further, we wish to maintain a low gas pressure during the useful life of the tube and so the getter must resume the function of removing gases as soon as they are liberated from the electrodes or from the envelope. Because of this second function, certain getters are also referred to as "keepers." Sometimes the word "getter" is given a wider meaning and is then used to designate those substances introduced into tubes by the common getter technique, but intended for other purposes than gettering. An example is the use of caesium metal in phototubes.

A getter is chemically a highly active material and combines rapidly with gases. Because of this very essential quality, it is necessary to protect the getter in some way when fastening it to the mount before the sealing and the evacuating operation. The protected getter, is, of course, inactive and is properly referred to as the getter source.

In the scant and scattered literature on getters, we find that Malignani is often credited with having invented phosphorus getter in 1896. Metallic getter, especially calcium getter, is said to have been invented by C. Soddy in 1906. I was greatly surprised a few years ago to find during a patent search that an electrically flashed magnesium getter was patented by Fitzgerald in 1883 (U.S. Patent 286,916). With the commercialization of the radio receiving tube, the patent literature on getters has increased rapidly. We shall discuss only such getters as are or have

been used for factory production.

Chemically active materials which in one form or another are used as getters are given in the following list.

Phosphorus
Magnesium

Calcium, Strontium, Barium } Alkaline earth metals

Lithium, Sodium, Potassium, Rubidium, Caesium } Alkali metals

Cerium, Lanthanum*

The properties of each of these materials will be discussed later on.

ACTION OF CLEAN-UP

If we vaporize or disperse a getter material in an imperfect vacuum, the probability of combination between the vapor and the residual gas is very great. Hence, the clean-up action is very fast. In describing this action we refer to (a) gettering by dispersion during which the vaporized getter combines rapidly with the residual gas and the surplus of the vapor settles on the cooler sections of the bulb or electrodes; (b) contact gettering during which the getter deposits combine slowly with residual gas which comes in contact with them due to temperature agitation; and (c) clean-up by ionization during which a rapid chemical reaction takes place between a getter deposit and an ionized gas. These three phases of gas clean-up are generally observed with any getter.

SHIELDING

Getter deposits are often undesired on tube electrodes and particularly on the insulating members of the tube structure. In order to keep the getter from depositing where it is not wanted, shields are used. Shielding is only effective if the mean free path of the getter vapor is large as compared with the distance between origin of vapor and location of deposit. When this is the case, the getter vapor travels in straight lines. Fortunately, this condition ordinarily exists in all tubes we are dealing with. There is also one further point which should be remembered in connection with dispersed gettering. Even if the pressure in the device is too high at first to satisfy the straight-line requirement, the pressure will fall rapidly after traces of the getter are vaporized.

* and some other rare earth metals (many of which constitute misch metal)

PHOSPHORUS

Phosphorus occurs in several modifications, two of which, the yellow and the red, are of interest to us. Yellow phosphorus melts at 44°C and ignites in air when heated to 50°C. If yellow phosphorus is finely divided (such as it is on filter paper dipped in a solution of yellow phosphorus in carbonbisulfide), it ignites even at room temperature. Yellow phosphorus is very poisonous.

Heating yellow phosphorus in a closed container (under pressure) for several days at 260°C gives the red phosphorus. Red phosphorus ignites in air above 400°C and is chemically much less active than the yellow modification. When red phosphorus is heated in vacuum, it reverts again to the more volatile yellow modification. These properties of phosphorus make it very convenient for use as getter, chiefly to remove traces of oxygen.

The lamp industry uses a suspension of red phosphorus into which the tungsten filaments are dipped. A trace of red phosphorus remains on the wire. It is chemically inactive in air during mounting of the filament, and even during sealing-in. After the lamp is exhausted and the filament is heated for the first time, it is reconverted to the yellow modification and is then capable of removing oxygen contained in the residual gas.

The same technique was used with the early types of radio receiving and transmitting tubes employing pure tungsten filaments. Since an excess of phosphorus did not interfere with the operation of the tube, a liberal supply of it was painted on the plate from which it was liberated by high-frequency heating of the plate or by internal bombardment of the plate. As mentioned in connection with the properties of the thoriated filament, red phosphorus was also used in connection with magnesium getter in the types 199 and 201-A. Both of these types used a fine thoriated-tungsten filament. The beneficial action of phosphorus consisted of cleaning up the layer of nickel oxide on the plates. As a result, the "poisoning effect" of the nickel oxide on the filament was eliminated.

Phosphorus getter cannot be used in tubes with oxide-coated filaments for it would combine immediately with the free barium in the coating. Like oxygen, phosphorus poisons the emission of oxide-coated cathodes. Back in 1931, a 24-head exhaust machine on which 201-A's with phosphorus and magnesium getter were made caused considerable shrinkage when it was used for exhausting type 227. It was found necessary to change the port rubbers and sweeps, and to clean the manifold thoroughly to remove all traces of phosphorus vapor.

MAGNESIUM

Magnesium metal is chemically very active. It reduces water vapor slowly at as low a temperature as 100°C and it has a considerable vapor pressure

at 600°C. When magnesium powder is blown into an alcohol or Bunsen flame, it combines instantaneously with the oxygen in the air and also partially with the nitrogen. We have all seen the action of flash-light powder. Solid magnesium metal, however, can be stored in air for years without showing signs of oxidation or corrosion. A fresh cut of magnesium metal is covered immediately by a molecular film of magnesium oxide which like an impenetrable varnish prevents the metal from further attack by the air. Breaking the continuity of this oxide film by amalgamation with mercury, for example, one can observe rapid oxidation at room temperature.

The use of magnesium as getter is based on its properties just described. As we know, no special precautions are needed in protecting magnesium getter during handling and sealing-in. It is dispersed in the tube by heating it to a temperature of from 500° to 800°C, and forms a mirror-like deposit on the glass bulb. The clean-up by magnesium metal during dispersion is very rapid. Its gettering by contact is slow and is ultimately limited by the formation of a molecular layer of oxide on the surface of the getter deposit. However, such a surface-contaminated deposit may still clean up large amounts of gas due to gettering by ionization. It is well known that tubes with magnesium getter remain good when in use because of the continuous gettering by ionization but tend to become gassy when stored. Therefore, magnesium is not a good keeper. In Fig. 13 is

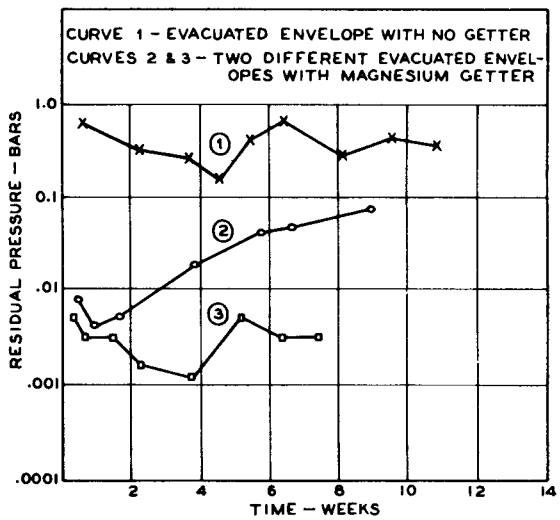


Fig. 13
(From S. Dushman, J. Franklin Inst., Vol. 211, pp. 689-750; 1931)

shown the relation between residual pressure in evacuated envelopes and time. Each curve represents measurements on one envelope. The uppermost curve was taken with a blank envelope (no getter). The two other curves show the keeping properties of magnesium getter in two different envelopes. Pressure measurements were made by means of ionization gauges sealed to the envelopes.

Magnesium metal is produced by electrolysis and while it is brittle at room temperature it becomes very ductile at about 200°C. Magnesium wire is made by extrusion at elevated temperature (450°C). To protect the metal from oxidation, the process is carried out in an atmosphere of illuminating gas. Therefore, magnesium contains dissolved amounts of light hydrocarbons which are liberated during flashing. In flashing a small piece of magnesium completely, we find generally a black residue which consists of carbon and magnesium oxide.

CALCIUM, STRONTIUM, AND BARIUM

The processes of preparing the alkaline-earth metals for gettering purposes are so much alike that they can be treated together. It has long been known that among these metals barium is chemically the most active; hence, all efforts have been directed toward obtaining and using barium as a getter while calcium and strontium have served only for casual experiments. Therefore, it will be sufficient to discuss barium getter. The methods of preparation used for barium are applicable to strontium and calcium. When barium metal is exposed to the atmosphere, it reacts immediately with oxygen, carbon dioxide, and water vapor. Consequently, it cannot be used in its metallic form but must be introduced into the tube indirectly, i.e., it is either protected or in the form of suitable compounds. Methods which have been and are still employed make use of (1) copper-, nickel-, iron- or aluminum-clad barium, (2) barium azide (BaN_6), (3) barium-magnesium and barium-aluminum alloy, and (4) batalum and batalum ribbon getter.

1. Copper-, Nickel-, Iron-, or Aluminum-Clad Barium Getter

One way of protecting barium from the atmosphere is to insert a stick of barium metal into a copper tube having an inside diameter of approximately one inch. The ends of the tube are closed by plugs and this assembly is then worked into a wire by rolling and drawing. The final product, for example, may be a wire having an outside diameter of 0.075 inch with a barium-metal core measuring 0.060 to 0.065 inch in diameter. Sections about 1/4 inch long are pinched off from this wire by means of a dull cutting tool. A section, often called a pellet, is then fastened upon a nickel or iron getter holder and/or shield from which the barium metal can be flashed by heating it to about 1000°C with high-frequency induction currents. This method of gettering was worked out about ten years ago and has been widely used in the manufacture of the larger glass receiving tubes.

It should be noted that with this method the

pinched ends of the pellets do not have an air-tight seal. Therefore, corrosion of the barium metal can take place, and once started, it progresses inward rapidly. Consequently, the getter action is satisfactory only with fresh pellets or those on which corroded ends have been cut off. As an aid in evacuation, degassing of the getter previous to flashing is essential. Changing the protective sheath from copper to nickel, aluminum, or iron has little influence upon the general qualities of the pellet.

The flashing occurs when enough pressure is built up inside the pellet to eject a mixture containing barium vapor as well as molten barium, and sometimes molten copper or aluminum, depending on whichever material is used for the sheath. Particles of the molten material often stick to the glass and cause microscopic surface cracks which widen in three or four weeks to form air leaks.

The barium deposit produced by this method generally has a dark color indicating that the presence of much gas during flashing reacted with the barium to form an impure deposit. Pure barium-metal deposits are almost indistinguishable in appearance from a magnesium deposit.

Recently, iron-clad barium wire has been proposed for metal tubes. In this form the wire has an outside diameter of 0.015 to 0.020 inch. The thickness of the iron sheath is from 0.002 to 0.0025 inch. A section of such wire is heated by passing current through it. The use of iron as a protective casing permits spot-welding the section between two lead wires. In order to facilitate the escape of barium-metal vapor and at the same time to produce a directional flash, the casing is ground flat, as shown in Fig. 14. The barium vapor breaks through the iron casing where

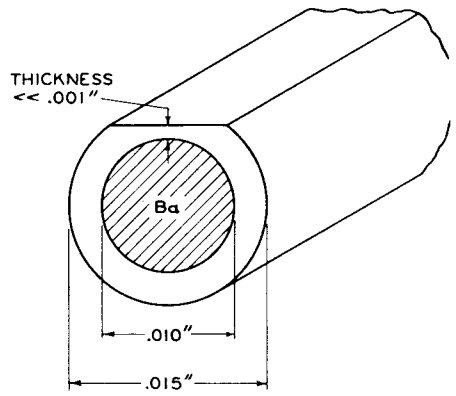


Fig. 14

the wall thickness is a minimum. Besides, the flat surface of the getter wire is intended to act as a shield confining most of the getter deposit to the section of the bulb near the flat surface and limited roughly by a plane coincident with it.

2. Barium-Azide Getter (BaN_6)

In Lecture 2 on oxide-coated cathodes, a brief discussion of the barium-vapor process, carried out with barium-azide as the barium source, was given. Much more barium metal was used with this process than was needed for filament activation, and so the surplus served as getter. Later, barium azide was used in Europe as a barium source for tubes using ordinary oxide-coated filaments. The getter source was made as follows: A nickel-mesh disk 3/4 inch in diameter with a centered hole 1/4 inch in diameter, was dipped into a concentrated solution of BaN_6 in water, and then allowed to dry. When this prepared disk (often designated by the name "life-saver") was heated in the tube during bakeout, nitrogen was liberated at about 150°C. The remaining barium metal was flashed from the life-saver by applying high-frequency heating. The life-saver shape was chosen because it heats up uniformly in a high-frequency field. A round disk made from material of poor heat conductivity (such as mesh) without a center hole heats at the edge first while the center is relatively cold. No efforts were made to produce a directional effect except that provided by the plate which shadowed the tube structure and the mica spacers.

Examining this getter critically, we find that the decomposition of BaN_6 is not complete but that the barium metal formed combines with part of the nitrogen. The result is that the surface of the life-saver has a mixture of barium nitride and barium metal. Consequently, when high-frequency heating is applied to the disk, the flash is accompanied by a second evolution of nitrogen gas. As a rule, the getter deposits appear dark, indicating vaporization in the presence of gas.

Fabricating the getter is objectionable owing to the danger of handling barium azide. The dry azide is very explosive and traces dropped on the floor may cause fire if they are stepped on. The saturated azide solution (17% maximum at room temperature), is not dangerous except that it is highly poisonous like all soluble barium compounds when taken internally. BaN_6 can be readily oxidized with potassium permanganate ($KMnO_4$), and so the work of dipping the nickel-mesh disks in the azide solution is carried out over a lead-lined sink which is flushed from time to time with $KMnO_4$ solution.

Reimann has studied the activity of barium-metal deposits in cleaning up gases. As a result of experimental evidence, he contends that a black barium-metal deposit cleans up gases by contact gettering about 4 to 5 times as fast as a bright metallic-barium mirror. The reason was sought in the difference of actual to apparent surface which no doubt is greater with the black deposit. Mr. Wamsley and I have observed exactly opposite effects. While this question should be definitely settled by further experiments, it is of no great practical consequence because contact gettering takes place only during shelf life, and as long as the gas pressure does not increase, the quality of the tube remains unaffected.

It is well known that contact gettering in

tubes using pure magnesium getter is so poor that the gas pressure increases on standing. This drawback has been overcome completely by the use of barium getter which permits the maintenance of a vacuum of the order of 10^{-7} mm for as long a time as free barium metal is available.

3. Barium-Magnesium and Barium-Aluminum Alloy Getter

In order to protect barium metal from attack by the atmosphere, attempts have been made to alloy it with other metals. When the alloy is heated in vacuum, barium metal is vaporized to serve as the clean-up agent. Barium-magnesium alloy has gained wide usage. If the barium content of this alloy does not exceed 25 per cent by weight, the alloy is stable in dry air and can be stored for long periods. However, owing to the difference of vapor pressure of the two constituent metals, large amounts of magnesium are vaporized during flashing before dispersion of the barium takes place. For example, if we use a piece of alloy weighing 5 milligrams, about 3 to 4 milligrams of magnesium must be evaporated before any but traces of barium are dispersed. This quantity of magnesium is very large for a medium-sized glass tube. Fortunately, with the magnesium assuming the role of cleaning up the bulk of the gas, only very small amounts of free barium are required to serve as a "keeper" during shelf life. A serious disadvantage in the use of barium-magnesium getter is that one cannot tell by visual inspection of the flash whether or not any of the barium metal has been dispersed.

No such limitation exists with barium-aluminum getter. This alloy contains from 45 to 50 per cent barium metal by weight. When this alloy getter is flashed, barium evaporates first; and if nickel or iron tabs (or getter holders) are used, no aluminum is dispersed. The amount of barium flash can, therefore, be judged by inspection. It is obvious that an alloy should contain as large a percentage of barium as is comparable with good stability in air. This maximum barium ratio is attained at about 50 per cent. Such an alloy decomposes slowly at 70 per cent relative humidity at 80°F.

Barium-aluminum alloy in reality is a chemical compound because during its preparation the heat of the melt increases exothermically from about 700° to 1700°C. Consequently, energy has to be expended to liberate barium metal during the flashing operation. Hence, barium-aluminum getter requires a high flashing temperature. Fortunately, this temperature is not so high as to prevent us from using nickel or iron tabs, although care must be exercised not to melt the tabs. It is sometimes advantageous to use a small amount of magnesium as precleaning agent with the barium-aluminum alloy. The magnesium may be mechanically admixed to the barium-aluminum alloy, or mixtures of barium-magnesium and barium-aluminum alloys may be used. Such mixtures were employed at the beginning of metal-tube manufacture. Although heating of the getter was carried out at the maximum permissible shell temperature, chemical analysis revealed that practically no barium (only magnesium metal) was flashed.

Barium-magnesium as well as barium-aluminum alloys, with barium ratio as given above, are brittle at room temperature. It is customary, therefore, to pulverize the alloys and compact the powder into small pellets on a pill-press. According to usage, the weight of getter pellets varies from 2 1/2 to 30 milligrams. When such pellets are flashed, variable amounts of gas are liberated. The amounts vary within such wide limits that a special study of these conditions has been started.

Another factor affecting the temperature of flashing is the thermal contact between tab and pellet. Latest developments take this factor into consideration. Hence, more uniform results are obtained with so-called imbedded getters, i.e., pellets pressed into the getter tab.

Both types of alloy getters are widely used in the manufacture of glass tubes and undoubtedly the results are better than if magnesium getters were used. However, there is still room for improvement. A barium getter for glass tubes requiring a low flashing temperature and permitting visual inspection of the amount of barium deposit may improve quality and decrease shrinkage considerably.

4. Batalum and Batalum-Ribbon Getter

Batalum getter has been devised for use in metal tubes. It is a reaction-type getter and derives alkaline-earth metal from its oxide which is reduced at elevated temperature with tantalum metal. Since we desired to produce barium with tantalum, the name "batalum" was coined to indicate this combination. The reaction is expressed by the formula:

$$5\ BaO + 2\ Ta \rightarrow Ta_2O_5 + 5\ Ba$$

Owing to its volatility, the barium metal is removed from the sphere of reaction and, therefore, it proceeds in the direction of the arrow. However, Ta_2O_5 (tantalum pentoxide) reacts with barium oxide forming barium tantalate, according to the formula

$$3\ BaO + Ta_2O_5 \rightarrow Ba_3Ta_2O_8$$

Therefore, the complete process can be represented by

$$8\ BaO + 2\ Ta \rightarrow 5\ Ba + Ba_3Ta_2O_8$$

Accurate measurements of the barium yield and the original amount of barium compound (weighed as carbonate) indicate the general correctness of this formula. The temperature required for the reaction is about 1300° - 1500°C. The process is most conveniently carried out by coating a tanta-

lum wire with barium and strontium carbonates, converting the carbonates to the oxides at about 1000° - 1100°C, and then raising the temperature to about 1300°C at which barium and strontium metal evaporates. Barium and strontium carbonates were used, because, as we know, barium carbonate alone melts and would cause sputtering. The reducing agent, the tantalum, in form of a wire has two main advantages. First, it affords control over the reaction by limiting the reaction surface to that of the wire. Thus, in controlling the wire temperature, we can start and stop the evaporation of barium at will. The second advantage is of chemical nature, and results from combining the wire (which serves as a heater) and reducing agent into one and the same unit.

In its first form the batalum getter consisted of a tantalum coil sprayed with the double carbonates (presintered in air to increase their density) and provided with a shield for directing the flash.[3] Flashing of the getter was accomplished by heating the tantalum wire electrically. While the results obtained with this getter were satisfactory, it has two disadvantages. One was its cost and the other was a processing requirement. The carbonate coating required decomposition during exhaust. Both disadvantages were eliminated by the so-called "ribbon getter." This getter consisted of a tantalum strip 1.5 x 0.1 x 0.0025 cm, with the middle section formed into a channel approximately 1 cm long. The barium compound, mixed with a nitrocellulose binder was filled into the channel on an automatic machine which formed and cut the channel and dried the coating. Since the channel which was coated only on the concave side was in itself a shield, and since the preparation of the getter was carried out entirely on a machine, the cost of the getter was only a small fraction of that of the coil getter.

The disadvantage of decomposing the carbonates during exhaust has been overcome by using barium-berylliate as coating material for the ribbon getter. This is a compound formed from the basic barium oxide with the amphoteric beryllium oxide and is approximately represented by the formula $BaBeO_2$.

Because of its advantages, the ribbon getter has found wide application in metal tubes. Both types of batalum getters (made in our own factory under controlled conditions to insure quality and purity) yield alkaline-earth metal of very high purity. The getters may be flashed partially or completely on the exhaust machine, and the flash is controlled completely by the temperature of the tantalum carrier. For most purposes it is advantageous to flash the getter after sealing off, i.e., before or during aging of the tube.

[3] E. A. Lederer and D. H. Wamsley, "'Batalum,' A Barium Getter for Metal Tubes," RCA Review, Vol. 2, pp. 117-123; July, 1937.

THE GROUP OF ALKALI METALS

The alkali metals, lithium, sodium, potassium, rubidium, and caesium are chemically very active but cannot be used for radio receiving-tube getters because of their high vapor pressure. Their use is confined to special devices, such as caesium-vapor detector tubes, phototubes, vapor lamps, and gas-discharge devices. In the latter, they serve as a means for reducing the "cathode fall." An alkali metal is introduced into the device either as a metal by vacuum distillation or is produced by reduction. While it is relatively easy to produce an alkali metal by reaction with a suitable reducing agent (vacuum decomposition of the tartarates, iron, aluminum, or calcium with the halides, oxides, carbonates, etc.), it is advantageous to use alkali compounds free from water of crystallization and such reactions as do not produce gas. For example, caesium metal is conveniently produced by heating an intimate mixture of caesium dichromate or permanganate with silicon or boron. The formulas are:

$$Cs_2Cr_2O_7 + 2\ Si = Cr_2O_3 + 2\ SiO_2 + 2\ Cs$$

$$2\ CsMnO_4 + 2\ Si = 2\ MnO_2 + 2\ SiO_2 + 2\ Cs$$

Both reactions are exothermic and once started at a temperature of about 900°C proceed until the reducing agent is used up. Therefore, the control is poor and reflashing of the getter is obviously useless. Notwithstanding these facts, reactions using purified silicon are extensively used in the manufacture of phototubes and Iconoscopes. Rubidium and potassium can be made in a similar manner. Lithium and sodium can be produced by reacting the respective berylliates with tantalum, titanium, or aluminum.

MISCH METAL

Misch metal is a mixture of the rare-earth metals and contains much cerium and lanthanum and in addition variable amounts of sodium but not over 5 per cent. This material has been proposed often as getter for various devices. There is no doubt that the heat of combustion of both cerium and lanthanum is very high and that misch metal may be an excellent material for dispersion gettering to remove traces of air. Misch metal for contact gettering is obviously very inferior to barium, because the former can be stored in dry air for many weeks while the latter reacts rapidly with the oxygen in the air. It is believed that misch metal is covered by a molecular layer of oxide which protects the metal from rapid oxidation. Misch metal is now used in gas-filled voltage regulators (type 874) as an oxygen getter and as a means to maintain the uniformity and magnitude of the cathode fall during life.

REFERENCES

S. Dushman, "Recent Advances in the Production and Measurement of High Vacua," Jour. Franklin Inst., Vol. 211, No.6, pp. 689–750; June, 1931.

A. L. Reiman, "Clean-up of Gases by Magnesium," Phil. Mag., Vol. 16, pp. 673–86; 1933.

Andrews and Bacon, "The Comparison of Certain Commercial Getters," J. Amer. Chem. Soc., Vol. 53, p. 1674; 1931.

A. D. Power, "Gettering of Hydrogen," Ref. 6—References a, Part III.

A. D. Power, "Gettering of Nitrogen," Ref. 6—References b, Part III.

Lecture 7

METALLURGICAL PRINCIPLES

S. Umbreit

INTRODUCTION

"Careful research into the nature of the metallic state has yet to discover, with any certainty, its essential quality. We do not yet know, for sure, what it is that makes the metallic elements different from, if not actually unique among created things. We do know, of course, that metals have certain more or less characteristic properties — they are hard, strong, and tough, and good conductors of heat and electricity, and these so-called metallic properties provide a definition of a sort. However, it is the structure rather than the behavior of the metals which gives the really fundamental information necessary to the understanding of metallicity.....

"As everyone knows, metals owe their industrial importance to the fact that they possess, often in a peculiar sense, useful properties, such as strength, plasticity, electrical conductivity, and ferromagnetism. These and other properties find their origin in the internal structure of the metal or alloy. What a metal does is because of what it is. Metallic behavior is simply an outward manifestation of what is, and what goes on, unseen, inside..... The basic fact underlying nearly all metallic behavior is that a metal is crystalline in its nature; any mass of metal that has been formed in the ordinary way is made up of many small crystals tightly bound together."*

These words of L. R. VanWert,[1] of Harvard, state the subject very well.

NUCLEI AND GRAIN GROWTH

If a piece of metal is polished, appropriately etched, and examined under a microscope, it is at once apparent that the mass is granular in character but not in the sense that each grain visible under the microscope has a definite shape and proportion as do mineralogical specimens. The irregular polyhedral shape of the metallic crystallite is due to the fact that it grew rapidly in a liquid of high concentration, whereas the mineral grew at an exceedingly slow rate from a dilute concentration. However, both crystals do have the common feature that the atoms are arranged with great regularity both as to position and spacing.

This regularity distinguishes the crystalline solid from the liquid and from glasses in which the atoms are arranged with random orientation and spacing. This appearance of the crystalline material in a liquid is called freezing which occurs in a pure metal at a single fixed temperature and coincides, in most cases, with the temperature at which the crystalline phase disappears on heating.

During the cooling of a melt of metal, crystallization centers or nuclei are formed spontaneously at the surface of the container. From these nuclei, the crystals grow in all directions but more rapidly in certain directions than in others. These crystals grow together so that the boundary planes form irregular polyhedra due to the mutual interference of one another. The freezing of a melt is not a simple affair but depends upon two factors which determine whether the melt will freeze on cooling or whether it will become a glass, and if it crystallizes, whether the grains will be coarse or fine. These two factors are: the number of nuclei and the rate of grain growth. Both factors must be large at the same temperature to produce a granular solid. The number of nuclei and the rate of grain growth are generally large in the case of metals but not so in the case of the alkaline silicates which constitute our glasses. In the latter case, the appearance of nuclei does not occur until the liquid has cooled to a point where the viscosity is high and the rate of grain growth has diminished to zero. Consequently, the mass remains glassy.

The difference between the behavior of a molten mass of a crystalline substance and that of a vitreous or glassy substance upon cooling, is that the former reaches a temperature where the crystalline, solid phase appears and further cooling results in a completely solid mass. Heating causes the disappearance of the crystallites at the same temperature. In other words, there is an abrupt change in viscosity or rigidity when the freezing or melting temperature is passed.

With a vitreous substance like glass or pitch, there is no abrupt increase in viscosity on cooling but rather a gradual increase extending down to or below room temperature. The mass gradually changes from a liquid to a solid and it is hard to say when it is a liquid and when a solid.

The behavior of a pure metal and glass is illustrated in Fig. 1, which shows the rate of grain growth (GG) and number of nuclei (NN) plotted against decreasing temperature. In metals the number of nuclei present is large at a temperature where the rate of grain growth is large; consequently, crystallization occurs. However, in glass the number of nuclei present in the temperature range of appreciable grain growth

---

[1] L. R. VanWert, "Inside a Metal," Mining and Metallurgy; October, 1937.

* Quotation with permission of Mining and Metallurgy.

is practically nil; so, the mass remains vitreous.

If the number of nuclei is great, there will be a large number of fine grains of about equal size because a large number of grains start to grow at the same time and with equal speed, but the growth is soon stopped by mutual interference. If, however, the number of nuclei is small and the rate of growth is large, then the

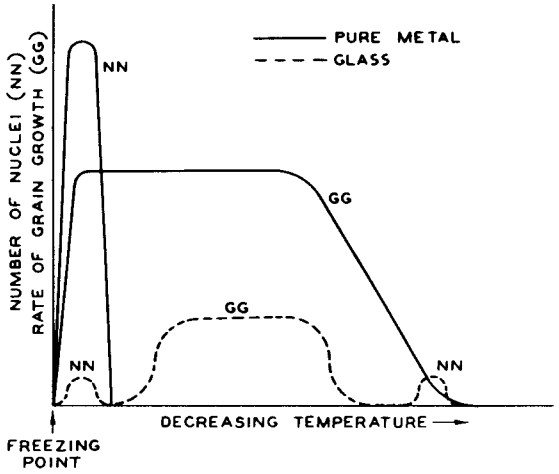


Fig. 1

grains will be large. Variations in rate of nuclei formation and growth account for the variation in grain size of different metals. Steel or tungsten has a small grain size whereas brass always has much larger grains which are plainly visible on the surface of any brass door knob frequently used and exposed to the weather.

This matter of grain growth under fixed conditions can be noted elsewhere in ordinary life. The large patterns of frost on a window pane in the winter are the result of slow growth starting from a limited number of nuclei. The phenomenon differs from the freezing of metals in that the solid is deposited from the vapor phase rather than from the liquid phase, but the mechanism is the same. A practical application of the knowledge of crystallization processes is found in candy making, which is essentially a freezing operation.

ATOMIC ARRANGEMENT

If the metallic grains which have developed during solidification are subjected to the monochromatic x-radiation of molybdenum, it will be found that the grains act like diffraction gratings and refract the x-rays an amount which depends upon the spacing of the atoms, because the spacing of the atoms in metals is comparable in size to the wavelength of x-rays. In a pure metal, all the atoms in a single grain are arranged in a definite manner and at definite intervals from one another in directions which are straight lines. In other grains, the arrangement and spacing are identical, but the directions of the axes are somewhat different. From the nature of the diffraction pattern produced, it is possible to determine the arrangement and spacing of the atoms. It has been found that most pure metals fall into three groups. The more ductile metals have atoms situated at the corners of a cube and at the center of the cube faces. This arrangement is called "face-centered-cubic" as shown by Fig. 2; Al, Ca, Fe between

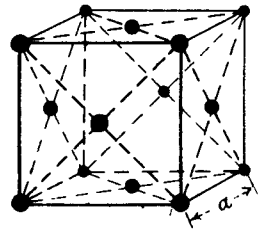


Fig. 2 - Face-Centered-Cubic Crystal Structure
(Courtesy of Metals and Alloys)

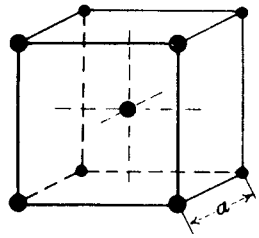


Fig. 3 - Body-Centered-Cubic Crystal Structure
(Courtesy of Metals and Alloys)

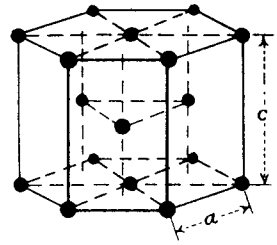


Fig. 4 - Hexagonal-Close-Packed Crystal Structure
(Courtesy of Metals and Alloys)

910° and 1400°C, βCo, Ni, Cu, Sr, Rh, Pd, Ag, Ba, βCe, Ir, Pt, Au, Pb, and Th are in this class. Harder, more brittle metals have atoms at the corners and center of the cube giving rise to the term "body-centered-cubic" shown in Fig. 3. These metals are Ti, Na, K, V, Cr, Fe below 910° and above 1400°C, Mo, Ta, and W. There is another grouping called "hexagonal-close-packed" that is obtained when balls are packed as closely as possible starting with a triangular unit. This arrangement is shown in Fig. 4. The metals falling in this class are Be, Mg, Ti, Cr, αCo,

Zn, Zr, Ru, Cd, αCe, Hf, Rn, Os, and Th. There are a few other types of more complicated symmetry that are less important.

It should be noted here that because of the regular spacing of atoms in metals, a large number of sets of parallel planes containing atoms can be found oriented variously to principal axes. The spacing between adjacent planes and the density of atoms in each plane depend upon the orientation. If the spacing between planes is relatively large and the bond between planes consequently weaker, the grain will distort during cold work by slipping along these planes, which are called "slip-planes." Ductile materials are distinguished by their ability to slip on these planes. The most ductile metals all have a face-centered-cubic arrangement which has four directions of easy slip, whereas the least ductile metals of the hexagonal-close-packed system have only one direction.

The atomic arrangements of practically all compounds and minerals, however complex, have been determined by x-ray analysis and classified by the Braggs in England during the last twenty years. Furthermore, much has been learned about the effect of cold work and annealing of metals which could not have been established in any other way. It is possible to use the x-ray method to determine the directional variations in mechanical properties of sheet material but other methods are available that are more sensitive.

With pure metals, the freezing and cooling to room temperature usually occur in a simple manner. Some metals like iron or nickel freeze at a definite temperature but on continued cooling reach a temperature where the atomic arrangement or some physical property changes abruptly. These points are called transformation or transition points. Iron freezes with the body-centered-cubic atomic arrangement at 1530°C but, with continued cooling to 1400°C, it has a transition point where the arrangement changes to that of the face-centered-cubic which is stable down to 910°C. At this temperature, it again transforms into the body-centered arrangement of atoms which is called alpha iron. Further cooling to 770°C brings about a change in the magnetic properties; iron is magnetic below 770°C and non-magnetic above.

The cooling of nickel to room temperature is unaccompanied by any atomic rearrangement but nickel becomes magnetic at 360°C and remains so below this temperature. The coefficient of electrical resistance of nickel is also different above and below this temperature which is called the Curie point.

Other elements show transition temperatures. For instance, the allotropic forms of carbon have a transition point where graphite and diamond are in equilibrium. Silicon has abrupt and reversible changes in electrical resistance at 214° and 435°C; titanium at 310°. Tin has transitions at 18° and 161°C.

The melting and transformation points of metals are usually determined by making cooling or heating curves which show discontinuities of rate at the temperature of transition or freezing due to absorption or liberation of heat. The duration of the arrest is proportional to the heat of fusion or transition and inversely proportional to the rate of cooling. The heat of transition is much smaller than the heat of fusion but is sufficiently pronounced to be used as a means of identifying the transition temperature.

MAGNETIC PROPERTIES

Metallic bodies may either concentrate or disperse magnetic lines of force. Metals which concentrate the lines of force are called paramagnetic; those that disperse them are diamagnetic. Ferromagnetic metals such as iron, nickel, and cobalt are distinguished from other paramagnetic metals by their marked capability of concentrating the lines of force.

The intensity of magnetization of ferromagnetic metals at constant field strength decreases only slightly with rising temperature but drops practically to zero at a definite temperature which is characteristic of the metal.

GASES IN METAL

The subject of gases in metals is an important one to us since our success in making tubes depends largely on our ability to remove gas from a bulb composed of or containing metal parts and to prevent the appearance of gas during the life of the tube. Unfortunately, little is known of the real nature of gas in metals; the operations of reducing the metal from ore depend upon gaseous reactions which permit the metal, which is usually molten, to dissolve gases readily. Generally, gases are more readily soluble in molten than in solid metal, but even vacuum melting will not remove all the gas from metal. It appears that gas can be retained by metals in several ways that are difficult to differentiate. Gas may be occluded or adsorbed on the surface, it may be dissolved in the metal as air is dissolved in water, it may be in combination with the metal as the oxide, or hydride, or it may result from a reaction between constituents.

Most metallurgists are interested primarily in the gases which are precipitated during the freezing of the metal and which produce blowholes, pipes, and porous metal in the cast ingot and cause difficulty in working the metal. But we are interested in the amount of gas that is liberated when the metal is heated in a good vacuum. The surface of the metal contains much adsorbed gas which can be easily removed by low-temperature annealing in hydrogen. Some metals contain oxides that are readily decomposed with heat under vacuum. The decomposition liberates oxygen. Such metals are silver, copper, iron, and nickel. Gases from oxides are easily removed but the gases that are dissolved or that result

from chemical reactions present a problem. In nickel or iron, if the oxide and carbon are present, they may react to produce carbon monoxide which diffuses slowly to the surface. The rate of diffusion of most gases, except hydrogen, is so slow that the ordinary firing and exhaust schedules used in tube manufacture can remove only a little from the surface. More diffuses slowly to the surface during life of the tubes and must be removed by the getter. A number of years ago, some experiments with nickel and iron indicated a higher rate of diffusion of gases in nickel than in iron. The higher diffusion rate in nickel permitted more gas to be removed from it during firing at 1000°C and exhaust than from iron. It was also found that with longer periods of firing, the amount of gas liberated during life was less, and that when the firing was continued for more than 8 or 10 minutes the rate of liberation of gas was reduced very little. Although some speak of hydrogen as replacing the gases in metal during our annealing operation and then being easily removed during exhaust, the hydrogen is only effective as a protective atmosphere that permits continued heating at elevated temperature without oxidation. Of course, hydrogen will reduce oxides and produce a bright surface from one which had previously been oxidized and make a clean part that can be welded easily. But the primary function of hydrogen is the prevention of oxidation during the period at high temperature necessary to permit the gases near the surface to escape from the metal.

COLD WORK

When a metal is subjected to permanent deformation or cold work, some of its properties are changed. This cold work may be accomplished in several ways, such as by rolling, drawing, hammering, swaging, bending, stamping, extruding, and stretching. By means of one or more of these operations, most of the metal used in construction is produced in a few simple shapes, such as wire, rod, tubing, sheet, strip, plate, rails, I, T, and angle sections. From these standard shapes are fabricated most of our metallic articles, the most important exception being castings for such parts as are too complicated to be made from plain shapes.

Consequently, most metal in use has at one time or another experienced the effect of cold work. When a piece of soft metal wire or rod, such as nickel, is subjected to tension as in stretching, the metal elongates slightly. As the load increases, the length increases proportionally until a point is reached where the increase in length is greater than expected. The maximum stress at which the elongation is proportional to the load, is called the proportional or elastic limit. If the load is removed, the length of the specimen returns to its original value, and the elongation of the specimen has been elastic. When, however, the load is increased to a value greater than the elastic limit, the specimen elongates at a much greater rate than before. Now if the load is relieved, the specimen is longer than it was originally by the amount of elongation occurring above the elastic limit. This increase in length represents the plastic or permanent deformation and removal of the load does not completely relieve the internal stresses.

The tensile test is a common physical test which is readily applied to metals. In the elastic range of the test, the ratio of the load to the resultant elongation is called the modulus of elasticity, or Young's modulus.

Actually, annealed metals can be stretched elastically less than 1 per cent after which they elongate rapidly or break suddenly if they have been hardened by cold work.

As the tensile test is continued to the point of fracture, the specimen elongates rapidly when the stress exceeds the elastic limit. This elongation is largely localized at a point which is reduced in cross-section and ultimately becomes the point of fracture. The maximum load required to break the specimen divided by the original area is called the tensile strength.

The reduction of area at the point of fracture expressed in per cent of the original cross-sectional area of the specimen is an important figure because it is an index of the ability of the metal to work-harden.

The total elongation is expressed as a percentage of the total length of the specimen but since the elongation is greater at the point of fracture because of the "necking down," it is customary to give the elongation for a 10-inch length and for the 2-inch length including the fracture. These are the values obtained from the tensile test.

Generally, soft or annealed face-centered-cubic metals have large elongation (20 to 50 per cent), large reduction in area, a definite yield point, and a low tensile strength. When these same metals are subjected to cold work, the elongation drops to 1 or 2 per cent, the area does not decrease, the elastic limit is identical with the tensile strength which is high, and the modulus of elasticity does not change. From these values, we learn something about the following properties of metals.

Ductility is the property by virtue of which a metal can be drawn out by means of tension without rupture. If a metal has high elongation and large reduction in area together with enough tensile strength to exceed the friction of the wire in the drawing die, the metal is considered to be ductile. If the tensile strength is low, the wire does not have sufficient strength to permit fine wire being drawn.

Malleability is the property by virtue of which a metal can be deformed by compression, as in hammering or rolling. Of course, tensile strength is not a requisite for malleability but compression strength is. Metals which are duc-

tile are also malleable but the relationship is only approximate.

Resilience is measured by the elastic energy which is stored up in the metal and is greatest in metals with low elongation and high tensile strength. Piano strings are made of hard-drawn steel wire for this reason.

Toughness and brittleness are, in a sense, opposite terms. Toughness is measured in terms of the work required to break the metal. If a stress-strain diagram is drawn for a specimen, the area under the curve is a measure of its toughness or resistance to fracture. Toughness is accompanied by high tensile strength and high elongation. Brittleness is the property of a metal by virtue of which it breaks without much deformation.

Ability of a metal to work-harden may be determined from the shape of its stress-strain curve and would be indicated by low reduction of area, high elongation, low yield point, and high tensile strength.

Hardness is a difficult property to evaluate because there are several kinds of hardness, such as resistance to abrasion, resistance to penetration, rebound hardness, and cutting hardness. These properties can not be determined from the tensile test but require individual tests for each kind of hardness.

Curve A of Fig. 5 is a tensile diagram for a hard-drawn metal which is strong, resilient,

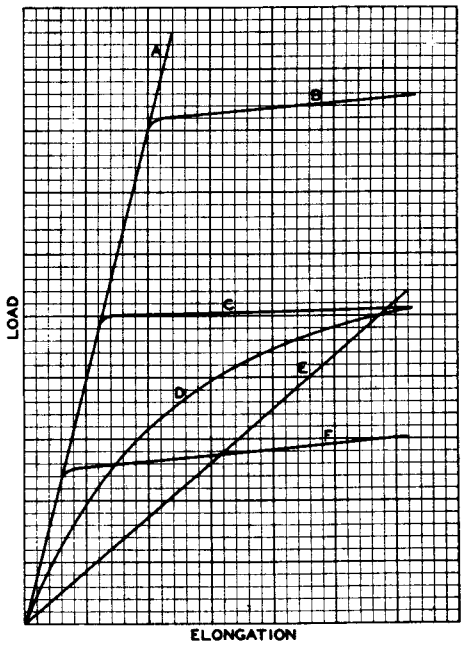


Fig. 5

and brittle; it has no ductility, malleability, or toughness. Similarly, curves B, C, and F show the behavior of three types of ductile, malleable metal. B, of course, is the strongest and toughest, and almost as resilient as A. F is a weak, ductile metal but although it is tougher than A, the latter is much stronger.

Curve D is for a metal which has good working qualities. Although it has the same tensile strength and elongation as has C, it is not so tough as C, and requires less work to rupture it; it work-hardens more rapidly, takes a permanent set even at low loads, and does not tend to localize the fracture by necking down. These characteristics are shown by the steepness of the curve over the plastic range.

Curve E is for a resilient metal that is more flexible than A. A spring of this metal would deform more for the same load than would one of the metal represented by curve A.

When a ductile metal is subjected to cold work, that is, plastic deformation below its annealing temperature, it is found that the elastic limit and the tensile strength increase, and the elongation decreases as the amount of cold work increases. The modulus of elasticity may change to the extent of 20 per cent; a similar change may be found in the electrical conductivity; and a slight decrease in density is also noted.

About 35 years ago, Beilby advanced the view that there was reason for supposing that metals exist in the amorphous or vitreous as well as in the crystalline state. In metal aggregates, the crystalline grains are surrounded by continuous films of amorphous material.

At low temperatures amorphous metal is harder and stronger than crystalline metal because, in the latter material, there are definite directions of weakness resulting from the regularity of the atomic arrangement. Near the melting point, the crystalline material is stronger but is surrounded by weaker, softer, viscous, amorphous metal having the properties of glass at elevated temperature.

When a crystalline material is subjected to cold work, the grain is distorted in a discontinuous stepwise manner; the grain breaks up into fragments by slipping along planes of weakness called slip-planes. This rubbing together of two surfaces displaces atoms from the regular crystalline arrangement, and produces amorphous metal in the plane of slip. When this slippage on a given plane is interrupted by interference with other adjacent grains, the amorphous metal becomes rigid and stronger than the crystalline material which it cements together. Further cold work produces slippage at other slip-planes so that the whole aggregate is broken down into fragments cemented together by hard amorphous metal.

It has also been found that, if a metal is broken at a temperature near the melting point, the fracture occurs in the grain boundaries but, when broken at a low temperature, the fracture occurs within the grains. The presence of amorphous metal in the grain boundaries would explain these observations, since the vitreous cement is stronger at low temperature and weaker at high temperature than is the crystalline material.

Tests of Jeffries and Sykes, where various rates of loading at elevated temperature were used, showed the properties of vitreous material. Very rapid fracture showed high tensile strength with a ductile transcrystalline fracture, while slow fracture showed a low tensile strength and a brittle intergranular fracture. In this test, it was assumed that the properties displayed by pitch exist in amorphous cement at elevated temperatures, namely, that it is relatively soft and plastic to slow loading but hard and brittle to rapid loading or shock.

The crystal structure of each grain is oriented differently from that of its neighbor, depending upon the orientation of the nucleus from which it grew. When the growth of two grains is interrupted by mutual interference, there exists between the two grains a zone of discontinuity, the nature of which is not exactly known. It may contain impurities or amorphous vitreous material; at least, it displays the properties of amorphous metal.

GRAIN GROWTH AND RECRYSTALLIZATION

When a metal freezes or crystallizes, grains grow in size and shape, depending upon the condition of crystallization. If the metal has not been cold-worked, the grain size of cast metal can not be increased by annealing at any temperature below the melting point. If the metal is worked at a temperature above the annealing point, that is, if it is hot-worked, new grains are formed, which are usually smaller. If this metal is heated to a higher temperature, fewer, larger grains result; this effect is called grain growth.

However, if a metal is plastically deformed below the annealing temperature, or cold-worked, it is found that the grains are distorted and broken. When the metal is heated, the distorted or elongated grains are replaced by new grains of polygonal form called "equiaxed grains" because the length of the grain in any direction is about constant. The change is called recrystallization and occurs at a temperature which depends on the metal and the amount of cold work. The lowest temperature of recrystallization visible microscopically for several metals after they have been subjected to severe cold work, is given in the following tabulation.

W	1200°C	Cu	200°C
Ta	1000	Al	150
Mo	900	Mg	150
Ni	600	Cu	room temp.
Pt	450	Zn	room temp.
Fe	450	Pb	below room temp.
Au	200	Sn	below room temp.
Ag	200		

The recrystallization temperature is lower, the lower the working temperature, the purer the metal, the smaller the initial grain size, and the longer the time of annealing. The progress of recrystallization can also be followed by noting the annealing temperature at which the electrical conductivity, tensile strength, and percentage elongation change greatly.

When pressed powders are sintered together, grain growth occurs at temperatures very close to the melting point. But, before grain growth can occur, consolidation by diffusion must take place. Grain growth or possibly recrystallization occurs in electrodeposited metals because, although they are crystalline, they are deposited in a strained condition which is conducive to an alteration in grain size or shape. When iron passes through the α-γ transformation, the change in atomic arrangement is accompanied by recrystallization without the presence of strain in the grain.

In cold-worked metal, the grains are elongated in the direction of the flow of metal and in a hard-drawn wire, such as tungsten, the grains appear as fibers or threads. At low temperature, the time required for recrystallization may be days; near the melting point, it may be seconds.

During recrystallization, strain-hardened metal changes to unstrained or annealed metal and small annealed grains then grow into larger grains. Actually, recrystallization is the change from the strained fragment of a grain into the annealed, strain-free grain of small size; grain growth follows recrystallization and changes a small grain into a large one by absorption of its neighbors. During annealing, these two operations occur almost simultaneously.

The time of exposure has an important effect on the temperature of recrystallization; the longer the annealing, the lower the temperature of recrystallization. The lower the temperature of deformation, the lower the temperature of recrystallization. Also, purity of the metal has much to do with the recrystallization temperature, for solid solutions have higher temperatures than pure metals. The greater the amount of strain hardening, the lower the recrystallization temperature.

With large grains only slightly deformed, it is impossible to have a uniform condition of cold work; and so such metal has a recrystallization range. The greater the amount of cold work, the narrower and lower is this temperature range.

Any foreign material in the metal affects its grain growth characteristics; foreign material may be present in solid solution, it may be a mechanical inclusion, such as slag, or as another constituent of an alloy. A good example of inclusion is the thoria in tungsten used for filaments in certain types of vacuum lamps and in radio tubes. Thoria was added to tungsten to prevent "offsetting" by obstructing grain growth normal to tungsten and producing many fine grains instead of a few large ones. In this way, the interlocking effect of the fine grains

prevented the side-slip which would occur in the plain surface between two large grains. Thoria is not introduced into filaments of radio tubes for the same purpose, but the obstructional effect is present.

The thoria is added to tungsten oxide or tungsten powder in the form of thorium nitrate in solution. When the mixture is evaporated to dryness, thoria is left finely divided and evenly distributed throughout the powder. When the metal is worked into wire, the thoria particles are drawn out with the wire into bead-like chains, because thoria is not so plastic as the metal. These particles offer less obstruction to grain growth in the direction of drawing the wire than they do radially; consequently, the tungsten grains are longer in the direction of working.

Frequently, it is found that the results of annealing contradict the laws of normal grain growth. But usually this will be found to be due to lack of homogeneity in composition, structure, strain, or temperature.

Strain gradients are introduced by drawing or rolling which produce greater stress at the surface of the metal than in the interior. During heating and cooling, there are always temperature gradients especially if the mass of metal is large and its thermal conductivity is poor. Under suitable conditions, these gradients result in very large grains due to exaggerated or abnormal grain growth. Single-crystal tungsten wire is produced by large temperature gradients that permit a crystal to grow along a wire at a rate higher than the speed with which the wire moves into the hot zone.

H. M. Howe has called this abnormality in grain growth "germination," and the annealing range conducive to it, the "germinative range." Many cases of germination have been reported and in most cases have been produced by subjecting the metal to slight strain followed by annealing at low temperature. Plain steel containing 0.05 to 0.12 per cent carbon shows this phenomenon quite clearly. Ruder has used this method to produce coarse grains in silicon steel.

Another case of germination occurs in drawn tungsten wire which has only small amounts of impurities or additions, such as are found in our non-sag wire. This wire, which has been drawn in the usual way and is thoroughly cold-worked, can be heated so that it will have either a coarse-grained or a fine-grained structure. If the wire is heated suddenly to a high temperature, the grains will be fine but if it is heated slowly through a narrow range of temperature, the grains become very coarse.

This grain growth is now known to occur by the gradual, though rapid, absorption of one grain by another, that is, the boundary separating the two grains migrates until one grain disappears entirely. This requires the transfer of atoms from the lattice of one grain to that of the growing grain. However, similarity of orientation in two adjacent grains is not considered a prerequisite for grain growth but rather the existence of thermodynamical instability of sufficient intensity to overcome the obstructional factors at a given temperature.

Jeffries and Archer believe that absolute and relative grain sizes are the chief factors in determining the force or tendency to grain growth. In aqueous solutions of salts, the tendency is known to be due to the greater solubility of the smaller crystals. Generally, in a mixture of grains of various sizes, the larger tend to grow at the expense of the smaller. Grain growth would not occur if the grains were all of the same size, because each grain is as potent as its neighbor and no change results. Grain growth depends upon difference in grain size where the large grains have greater power of reorienting the atoms in the lattices of neighboring grains. The thing that limits the grain growth and prevents metals from being single-crystalline is the fact that the increase of reorienting power with size is most pronounced in fine grains and diminishes rapidly with grain size.

The growth of grains in a metallic aggregate is a competitive process, each grain competing with its neighbor for the atoms in the grain boundary. The more powerful grain grows at the expense of the weaker and thereby becomes a trifle stronger.

The three important factors in grain growth are grain size, temperature, and obstructions. When a grain is cold-worked, a large number of fragments are formed, some large and some small; each, however, is a crystal fragment capable of growth when temperature permits. If the temperature is raised rapidly to a point above the grain-growth range, all fragments have about the same power to grow, and each is opposed by its neighbor. Consequently, only a little grain growth occurs. If, however, the temperature is raised to a point just below the grain-growth range, only a few fragments have sufficient energy to force their orientation upon their neighbors, and they grow with little or no opposition because they are surrounded by small, easily absorbed fragments. If sufficient time is given, the aggregate will consist of only a few large grains. This condition is designated as "exaggerated grain growth." Most annealing operations are performed at temperatures high enough to produce a moderate grain size, since coarsely grained metal is not particularly desirable in industrial applications.

ALLOYS

An alloy is a mass of metal containing more than one element, which has solidified from a melt that was a single homogeneous liquid at some temperature above the melting range. Molten lead and aluminum are practically insoluble in each other and when such a mixture of these metals is frozen, a layer of lead and a layer of

aluminum are formed. This is no more an alloy than a mixture of oil and water is a solution.

The simplest form of alloy is the solid solution such as is formed when copper and nickel alloy. If an alloy of these two metals in any proportion is inspected microscopically, it has the appearance of a pure metal because there is only one phase, the solid solution of one in the other. The equilibrium diagram is shown in Fig.6. In this case, the melting point is roughly proportional to that of the two metals; however, the alloy does not have a single melting temperature but is pasty over a small range of temperature.

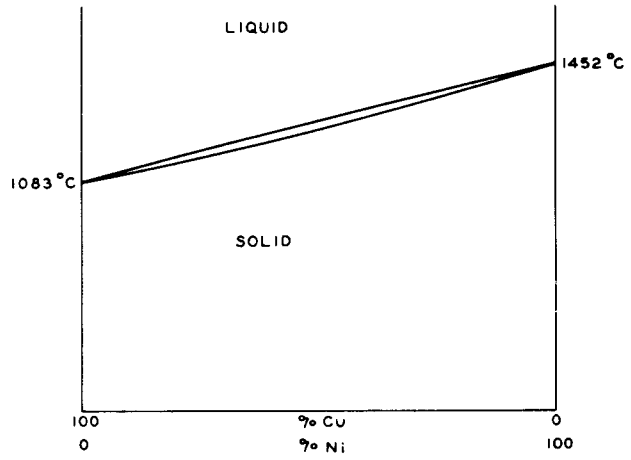


Fig. 6

When a solid solution is formed, the resulting alloy is usually harder and stronger than either metal; its electrical and thermal conductivity, and its vapor pressure are less; and its recrystallization temperature is usually greater.

Another simple type of alloy is one in which each metal is completely insoluble in the other in the solid state. Its equilibrium diagram is shown in Fig. 7. Line AE represents the solubility of silver, and BE the solubility of lead. If the alloy is rich in silver and cools from the liquid state, pure silver will be precipitated when the temperature reaches that shown by the line AE; further cooling results in continued precipitation of silver and the remaining liquid has the composition shown by the line AE. In a lead-rich alloy, pure lead is precipitated when line BE is passed. As the composition of the liquid changes with cooling, a point is reached where the two lines cross; below this temperature the liquid is saturated with respect to both metals and they separate side by side to form a peculiar structure which is called a "eutectic." It has the lowest freezing point of all silver-lead alloys and no alloy is completely frozen until it reaches the eutectic temperature which is about 310°C.

An ordinary example of this type of solubility is that of ice and common salt. The lowest temperature at which these can exist in solution is about -20°C. Consequently, salt cannot be used to thaw ice if the ice is colder than -20°C (-5°F). It should be noted that the chilling effect of ice and salt mixtures is due to the fact that salt dissolves in water or reacts with ice with the absorption of heat and the resultant solution is liquid to -20°C.

There is another type of alloy that is more common than either of the two types given above

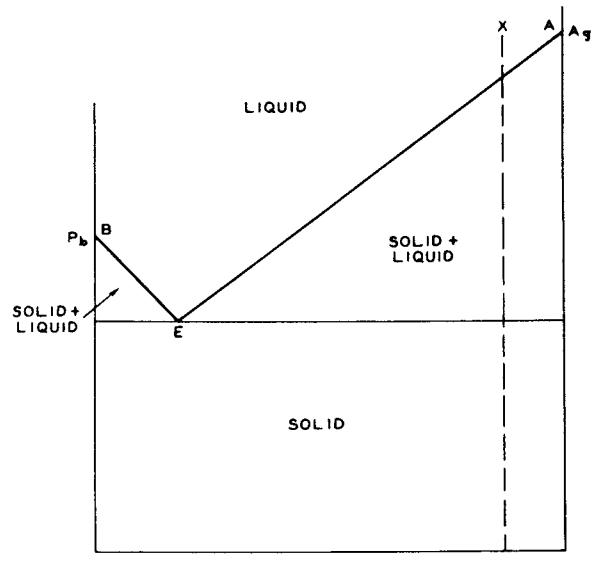


Fig. 7

because it is an intermediate case. This type occurs when the two metals are only partially soluble in one another in the solid state. Silver and copper is an example of this. At the eutectic temperature, each metal is soluble in the other to the extent of about 5 per cent. At lower temperatures, the solubility usually decreases but not always.

When such an alloy freezes, the metal which separates from the liquid is not a pure metal but is a solid solution containing some of the other constituent; and the eutectic is composed of two solid solutions simultaneously precipitated.

The maximum solubility of copper in iron is 3.4 per cent, whereas the solubility of copper in nickel is 100 per cent. To us, these are important facts because they permit us to copper-braze steel parts easily and to clean copper-plated button stems. If the solubility of copper in steel were greater, the copper would not flow over the surface into the joints which we wish to seal but would soak into the metal like water into a blotter. If the solubility were zero, the copper would not wet the surface, and consequently, would not spread over it. Even so, if a heavy coating of copper is placed on steel and heated above the melting point of copper, the surplus runs off the surface or collects in puddles. To avoid this action on the button-stem header where a heavy layer of copper is required,

the steel is first plated with a layer of nickel heavy enough to absorb all the copper and prevent the puddles which would otherwise form. The nickel resists oxidation during stem-making better than copper alone, but it could not be cleaned up later if the copper were not present. During stem-making, the copper is more rapidly oxidized at the surface of the metal, and is easily reduced to metal during the hydrogen cleaning operation which leaves a bright copper-colored surface.

It may be noted that molybdenum cannot be brazed in hydrogen with copper, because copper is insoluble in molybdenum and there is no tendency toward wetting of molybdenum by copper.

Ordinary solder, a lead-tin alloy, is a similar case and brings up an important point. With the exception of the 65% Sn - 35% Pb alloy which has the eutectic composition, no other composition of these alloys has a freezing point. They have freezing ranges and are pasty above 181°C. This means that in soldering a joint, it is necessary to heat it above the pasty range (above 240°C for a 40-60 solder) but the finished joint should never be heated above 180°C, because liquid metal is present above 180°C and weakens the joint. The wiped joints made on lead pipes by plumbers were only possible because they used a solder low in tin which had a long pasty range that permitted molding the joint by wiping.

Lecture 8

METALS FOR VACUUM-TUBE CONSTRUCTION

S. Umbreit

NICKEL

In the manufacture of receiving tubes, the most important metal is nickel, although with the advent of the metal tube, iron and steel have become important factors. If nickel is superseded by steel, it will be largely because of the latter's lower cost, for in nearly every respect the properties of nickel are superior to those of iron or steel for our purposes.

Nickel is used largely because it is a very ductile metal and can be drawn or formed into the complicated shapes that we require. This property of working quality is due to the ability of nickel to work-harden considerably, whereas annealed steel on working first softens and then work-hardens slowly. This is indicated by the stress-strain diagram shown in Fig. 1. The working property may be demonstrated by bending

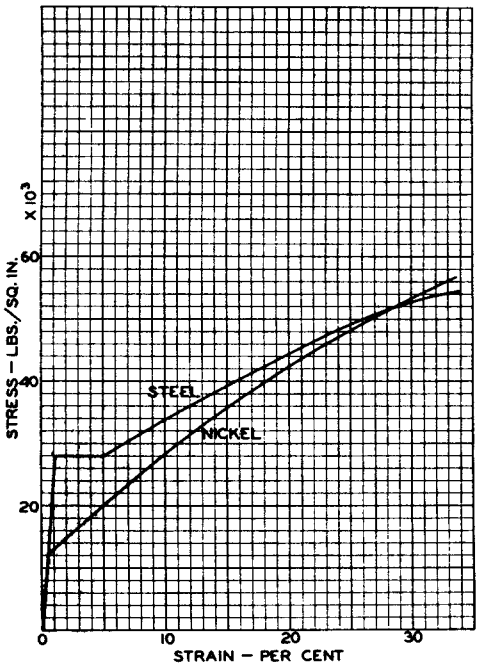


Fig. 1

a piece of nickel wire or strip and a piece of steel both of which have been annealed at 900°C. The nickel will bend with a smooth curvature of long radius and the steel will bend sharply with a short radius. This localized bend is called a "break" and is the result of the lack of work-hardness.

Nickel does not have a transformation point as does iron, and what carbon is retained in the metal does not form a hard carbide. It does not rust in moist atmosphere, or oxidize readily when heated in air. Its oxide is readily reducible by hydrogen even though it contains substantial amounts of water vapor or a few tenths of a per cent oxygen.

Nickel has a fairly high recrystallization temperature and will withstand better than iron "creep" which is the slow deformation that occurs with moderate loads at high temperature. The vapor pressure of nickel is low enough to permit high-frequency heating to about 1000°C in a good vacuum; iron vaporizes somewhat more readily.

Sulphur has a detrimental effect on nickel because these elements unite readily to form the sulphide. This sulphide has a low melting point which causes the grains of nickel to be separated by films of the brittle sulphide. In the melting of nickel, oxygen and sulphur are removed by the addition of a small amount of magnesium which raises the ductility of the nickel and permits hot-working of the metal.

Nickel is used as the base metal for filaments and cathodes coated with barium-strontium oxide mixtures. In this case, the composition of the metal is important. Filaments are heated by the passage of current so that the electrical resistance at the operating temperature is important. Because the filaments are draped over hooks that apply tension which keeps the filaments taut and straight when hot, the hot tensile or creep strength should also be high. Most burn-outs of nickel filaments are not the result of melting the metal, but the result of too much tension. At the operating temperature, the tension applied by the hook exceeds the creep strength of the filament which is eventually pulled apart. When the lighted filament is broken, an arc is formed; the arc fuses the fractured ends and gives the appearance of a burn-out. Cobalt is often added to nickel to raise both its hot tensile strength and hot resistance; a high resistance means a larger conductor which is inherently stronger.

The way in which impurities in nickel affect cathode emission is a complicated matter. If reducing agents remain after the deoxidation of the molten nickel, as they always do, the emission is higher than when they are absent. The greater the amount of reducing-agent residual, the higher the emission. This effect is particularly noticeable when these nickels are compared with electrolytic nickel which has poor initial emission. The reducing agents that may be present are silicon, manganese, titanium, aluminum, magnesium, and vanadium. These impurities have other effects. A few tenths of one per cent of manganese will increase the radiating ability of the cathode enough to lower its operating temperature 20°C. Titanium reacts with the coating to produce a black compound that reduces the temperature 60° to 80°C. The effect of impurities in nickel is

important and is receiving more and more attention. We should soon be able to say which impurities are most important and which have little or no effect.

Nickel is also used for supports for grids because it has sufficient ductility to permit the cutting of a notch into which the grid wire is laid and then peened securely. The thermal conductivity of nickel is good so that the grid wires can dispose of their heat and operate fairly cool.

The cross wire or grid wire usually is of nickel alloy, such as manganese- or chromium-nickel or even a molybdenum-iron-nickel; these are cheaper than molybdenum. When grid bombardment is used on grids of nickel alloys, the temperature should be carefully controlled because their melting points are lower than that of nickel and much lower than that of molybdenum.

The leads and supports in tubes are also made of nickel and here the stiffness and annealing characteristics are important; impurities are not so important, except insofar as they affect these characteristics or the thermal conductivity. When stems are made, the nickel leads become so hot at the point where they enter the glass that they are partially annealed. If the annealing is complete, the lead may be too soft to support the attached part properly. An instance of the effect of impurities on nickel for lead wire was experienced recently when a glass stem was made with one hard and one soft plate lead, but after stem making, the hard lead was less stiff than the soft lead. The hard lead was much purer than ordinary grade A nickel and, consequently, annealed more readily and resulted in a softer lead.

The thermal conductivity of the nickel leads is important. If one wishes to avoid electrolysis in a rectifier stem, he might do it by making the lead of a poorer conductor than nickel, such as manganese-nickel or ferro-nickel. But, if the anode is so hot that reverse emission from the plate is a factor, a pure nickel lead might better be used to conduct more heat away from the plate. The resultant lead would be a nice adjustment of size and composition to give the proper thermal and electrical conductivity and stiffness.

Nickel is particularly adapted to plate making because of its ductility, the absence of any transformation from one phase to another and its behavior with carbon. Compared with steel, it has a greater ductility which permits the more complicated shapes to be made with less difficulty, as mentioned previously.

Most plates are carbonized. Carbonization is accomplished by the batch process which is impractical with steel. If steel were carbonized under the conditions used for nickel, the carbon would dissolve in the iron to a much greater extent forming iron carbide, which makes cutlery as hard as it is. It is apparent that any distortion or warping of the plate in this hardened state could not be corrected because the resiliency and rigidity would be too great. If the temperature of carbonizing were reduced to a point where the iron carbide does not form, the character of the carbon deposit would be incorrect.

Moreover, when iron is heated above 910°C, there is a change from the body-centered-cubic arrangement to the face-centered-cubic arrangement which is accompanied by a change in volume. Such changes, especially in thin sections result in distortion or warping, and repeated excursions from one condition to another would result in a badly distorted plate.

To sum up the advantages of nickel would be to list its high ductility, freedom from rust or oxidation, ease of bright annealing, freedom from transformations involving a change in volume, low solubility of carbon and absence of carbide, easy weldability, good electron emission from oxide-coated surfaces, low vapor pressure, and good strength and stiffness.

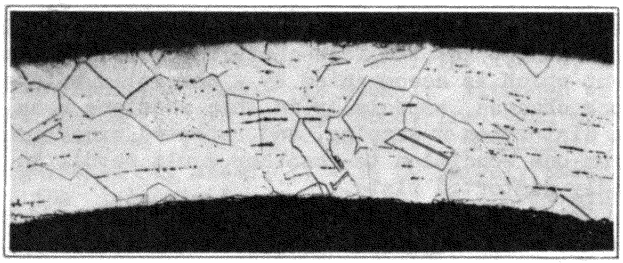
IRON AND STEEL

Iron and steel suffer in comparison with nickel when properties are concerned but when cost is considered they are much cheaper than nickel not only because the price per pound is less but also because the density is less. Iron has a higher melting point than nickel and retains its magnetism up to a higher temperature (770°C); this may be an advantage or a disadvantage depending on whether magnetic shielding is desired or not. Much has been written about the gas content and permeability of iron to hydrogen but the data are conflicting. However, it appears that nickel is more impervious at low temperature and that iron is more impervious at high temperature, possibly only when in the gamma state. Since carbon must have much to do with this property, the discrepancies in the data may arise from variable carbon contents. Knoll. and Espe hazard the guess that the diffusion occurs in the carbon segregated in the grain boundaries.

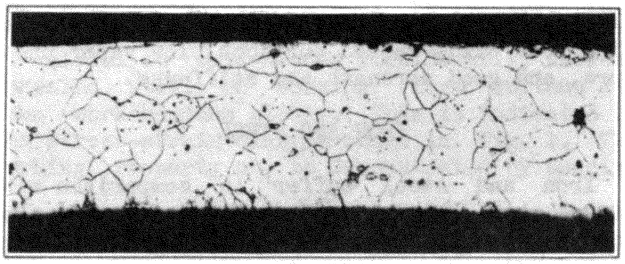
Very pure iron has a lower vapor pressure than nickel while steel has a higher pressure. Consequently, it is necessary to make a distinction between iron and steel. When steel is nickel plated, a combination of the desirable properties of nickel and of steel can be obtained. At present, the main difficulty is in obtaining strip steel with a heavy plate of nickel evenly distributed. Much nickel has been replaced by nickel-plated steel and in the future more will be.

Fig. 2 shows the microstructure of nickel strip and iron strips. The top section is that of nickel which is characterized by the short, fine, dark lines running in the direction of rolling. Some crystals have parallel sides, a condition which is characteristic of rolled and annealed nickel. The next in order are "Svea" and "Armco" iron which have much the same appearance. The dark particles are inclusions of slag

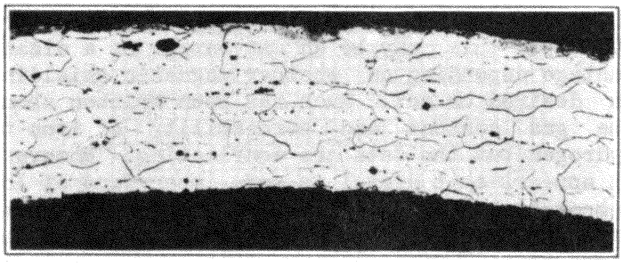
or oxide. There appears to be little carbon present in these metals. The bottom strip is Swedish iron which is distinguished by heavy black lines



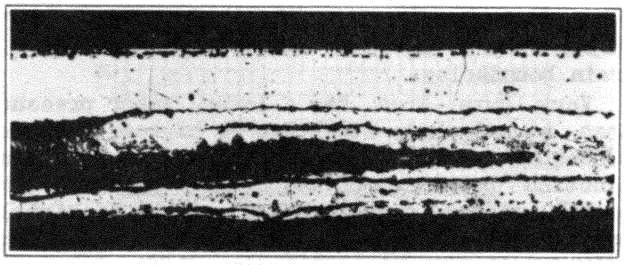
NICKEL



SVEA IRON



ARMCO IRON



SWEDISH IRON

Fig. 2 - Strip materials.

of slag. Needless to say, this metal cannot be used for the fabrication of parts due to lack of ductility. Furthermore, when such a strip is fired in hydrogen, blisters are formed where slag inclusions are near the surface.

TUNGSTEN

Tungsten and molybdenum are metals whose melting points are so high that no satisfactory refractories are available for crucibles in which to melt them. With the high melting points of these metals, there are several other properties that are particularly valuable for our work, such as low vapor pressure, low gas content, low coefficient of expansion, high tensile strength particularly at high temperature, low specific heat, etc. However, it is more practical to take up each metal individually, and compare it with other metals, and discuss its merits and demerits.

Tungsten, because it is prepared by the costly process of sintering the tungsten powder into coherent form without contact with other materials, is one of the purest metals commercially available. The coherent metal is prepared by reducing the purified oxide with hydrogen. The powder so produced is pressed into a bar or slug which is consolidated in an atmosphere of purified hydrogen by passing a current through it such that the temperature is raised nearly to the melting temperature.

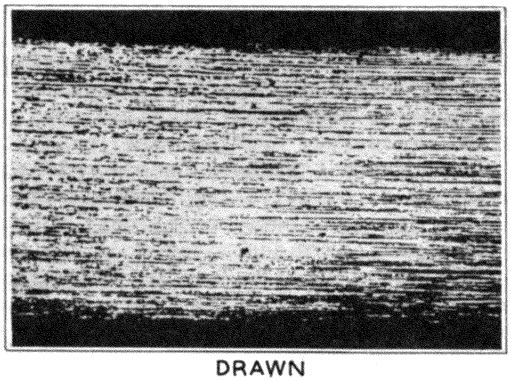
Unthoriated filaments usually contain no more than 0.02 per cent impurities. The sintering process has advantages over the melting process; consolidation of the metals begins at the center of the slug so that impurities have an opportunity to escape. However, the greatest advantage of this method is that we can add infusible oxides, like thoria, and still maintain the oxide particles finely divided and widely dispersed. If this were attempted in a molten metal, differences in density and surface tension would result in segregation.

We use two kinds of tungsten. Thoriated tungsten, containing 1.5 to 2.0 per cent thorium oxide finely divided and uniformly distributed throughout the wire, is used for filaments in power tubes because of its high electron emission. More thoria can be added to tungsten but drawing becomes difficult and impractical. Life tests indicate that the more thoria present, the better the life although, at the end of life, about 80 per cent of the original thoria remains in the filament.

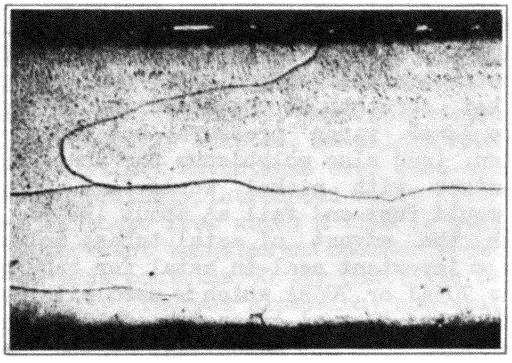
The other kind of tungsten is called non-sag tungsten. It is essentially pure tungsten. This metal is produced by adding to tungsten powder limited amounts of silica and alkalis which vaporize during the heating operations of manufacture, and leave only minute traces that greatly influence the shape of the grains produced when the filament is heated. The mechanism of this grain formation is not well understood, but the grains are quite large, and overlap one another for considerable distances. This structure is quite stiff at high temperatures and when coiled into a lamp filament does not sag during life as would a thoriated or pure tungsten filament. Non-sag wire is used for heater filaments, hooks, and supports where hot strength is desired, although our temperatures are not so high as to produce or require the normal overlapping grains.

Fig. 3 shows the fibrous structure of drawn wire and the crystalline structure developed by burning a tungsten filament.

The tensile strength of tungsten is very high, especially when it is in fine wires in the cold-drawn state. Moreover, this high tensile strength is well maintained at high temperature although it falls off with increasing temperature. Below room temperature, the tensile strength is lower because of the glass-like characteristics of the



DRAWN



ANNEALED

Fig. 3 - Tungsten wire.

intergranular cement. Drawn tungsten at room temperature is pliable but has no elongation when broken in tension. Annealed or equiaxed wire is very brittle and cannot be formed or bent at room temperature, but if the temperature is raised to between 300° and 400°C, the wire becomes quite ductile and can be bent or formed easily. If the temperature is raised to between 1500° and 1700°C, the grains in some lots of non-sag wire begin to coalesce or grow, although in appearance they are still fibrous; this grain growth results in the brittle wire which gives so much trouble in heater winding and forming. It is believed that brittleness is primarily caused by this action rather than by the formation of carbide from the graphite left on the surface of the wire after wire-drawing. All drawn-tungsten wire before annealing is in a severely cold-worked condition, which varies with the conditions of wire-drawing for larger sizes. Variations in amount of cold work cannot be measured but they have a decided effect upon the annealing temperature, and cause some wire to become brittle and other wire not brittle, in spite of the close control in manufacture. Very small amounts of impurities must also have a large effect on the annealing temperature. Much time and money has been spent by the lamp companies in an effort to control the working qualities of tungsten, but in the twenty-five years, the results are none too satisfactory.

Tungsten possesses the highest modulus of elasticity of all metals and it does not decrease greatly, even at 1000°C. A high elastic modulus means great stiffness or strength, and requires a larger load to produce a given deflection than does a lower one. This is particularly important for hook or support material. The elastic modulus of tungsten is about twice that of nickel at room temperature, and at 500°C it is about three times as great. It may be noted here that the mechanical properties of tungsten, as of other metals, are affected by the grain size and rate of loading in tensile testing, especially at elevated temperatures.

The thermal properties of tungsten, such as vapor pressure, rate of evaporation, heat conductivity, thermal expansion, specific resistance, electron emission, etc., have been thoroughly and carefully determined by workers here and abroad because of the importance of tungsten to the lamp industry. The electrical resistance of tungsten is about one-half that of nickel, and three times that of copper at room temperature. The resistance rises rapidly and this rise in resistance gives us an easy means of measuring the operating temperature of our tungsten heaters or filaments. This high coefficient of resistance of tungsten permits a surge of current when a tungsten heater is lit which reduces somewhat the heating time from what it might be if some other metal were used.

The low thermal expansion of tungsten, about $4.6 \times 10^{-6}/°C$, is the lowest of all metals. The thermal expansion of graphite is about 3.0, and that of silica is 0.5×10^{-6}. The expansion of tungsten matches that of Nonex glass so closely and the adhesion of the oxide to tungsten is great enough to permit the making of a good glass-to-metal seal, with the result that tungsten is extensively used for leads in power and transmitting tubes utilizing Nonex or hard-glass seals. Furthermore, its high electrical conductivity is a decided advantage over other metals used in seals especially where high-frequency currents are carried.

Dr. Lederer has mentioned the fact that tungsten, when heated in an atmosphere of hydrocarbon gas, forms two carbides which aid the emission and emission life of thoriated filaments. W_2C is more stable and, consequently, is the one found after burning. Carburizing is used for two reasons: to produce thorium from its oxide more readily, and to retain the thorium on the tungsten surface for a longer period. It is particularly fortunate that we can produce thorium slowly and continuously from finely divided

thoria, because this process assures that the filament will have a constantly replenished surface of thorium atoms. If this were not the case, an alloy of thorium and tungsten would very rapidly lose its thorium and become depleted, because the thorium would reach the surface much more rapidly than a monatomic film could evaporate, with the result that a polyatomic layer would be formed and this would evaporate more rapidly than a monatomic layer. When a carburized tungsten filament is burned, the tungsten carbide disappears at a greater rate than can be accounted for by the reaction between thoria and the carbide. This disappearance must result from the decomposition of the carbide and the vaporization of carbon. A filament in a 50-watt tube operated for 500 hours at normal voltage would contain no carbide at the hottest part, although the emission would show no serious diminution. Apparently, the effect of the carbide remains after its disappearance. The rate at which the carbide disappears is a function of not only the getter action but also the amount of magnesium present.

MOLYBDENUM

In less pronounced form, molybdenum has the characteristic properties of tungsten and consequently, is used less widely in the lamp industry. However, in the drawn state, it possesses a respectable amount of ductility and elongation; even after prolonged heating at elevated temperature, the wire is not brittle. For this reason alone, it has applications where it is preferable to tungsten. We are using molybdenum for support hooks, support rods, grid wire, filaments, plates, seal-in-wire, getter tabs, and furnace windings for high-temperature work. In power and transmitting tubes, much molybdenum is used where nickel is used in receiving tubes.

Because of its superior working properties, drawing, rolling, and forming of molybdenum can be performed at room temperature, whereas tungsten must be worked hot. In the cold-drawn state, its properties are similar to those of tungsten; when annealed, molybdenum behaves like the ordinary ductile metal.

At one time, the greatest use of molybdenum in receiving tubes was for grid wires, but now it has been largely replaced by cheaper nickel alloys which also have some other advantages. Molybdenum is an excellent grid wire for several reasons. Its thermal conductivity is high so that heat is transmitted to the support rods readily, it has good hot strength so that it does not distort when subjected to grid bombardment, it possesses ample ductility for making grids which will retain their shapes, its thermal expansion is low so that there is less likelihood of buckling in the tube if the grid is bombarded, and the melting point is high enough to permit thorough degassing. Six or seven years ago, molybdenum was used almost exclusively for grid wires but now its use in receiving tubes is limited to a few flat grids where other materials can not be substituted. There are several reasons for this change to nickel alloys. We found that manganese nickel, because of its contact potential or lower grid emission, enabled us to obtain greater power output from tubes like the 45, 46, and 47. Then, too, more receiving types were made with cylindrical grids which do not require the hot strength of molybdenum and so nickel alloys were cheaper and equally satisfactory. Furthermore, molybdenum oxidizes readily and its highest oxide is distinctly volatile even at atmospheric pressures. During the exhaust of a tube, a molybdenum grid may become oxidized and be subjected to positive ion bombardment if gas is present. Under these conditions, molybdenum or its lower oxides are deposited upon the cathode which then loses its emission. For these reasons, little molybdenum is now used for grid wires. It has been replaced by manganese nickel, chrome nickel, and even steel.

Molybdenum is also used in making support hooks and leads or tips for tungsten filaments, where the superior mechanical properties of molybdenum are required. In power tubes, molybdenum tips are used between the tungsten filament and the nickel lead not only for the sake of the added hot strength at the operating temperature and the lower vapor pressure, but also because tungsten (and also molybdenum for that matter) in contact with nickel would form a eutectic which would fuse and fail at about 1300°C.

With the advent of metal tubes, molybdenum became an important seal-in metal for hard glass, such as 705AJ or 705BA which is used with Fernico eyelets. The thermal expansion of molybdenum is linear and does not match that of the glass as well as does that of Fernico or Kovar, but the much better electrical conductivity of molybdenum gives it a distinct advantage. The electrical resistance of Fernico is high. Because of the small size of the eyelets and short path in the glass between eyelet and lead, the latter must be kept as small as possible and be made of a metal having good conductivity.

When we started to make metal tubes with molybdenum seals, we felt it necessary to fire the nickel-molybdenum-nickel leads in hydrogen for one hour at 1100°C in order to remove from the surface the carbon that gave gassy seals. This treatment was disadvantageous and even detrimental. At 1100°, the nickel became exceedingly soft and sintered the leads together so that it required a large number of girls to separate them and to straighten them by rolling. Also, some lots of molybdenum would recrystallize near the nickel welds and become so brittle that the leads were useless. We soon found that the firing of these leads was unnecessary if the wire before weldmaking was properly cleaned by etching away the outer surface.

Molybdenum support wires are used in power tubes because of their greater stiffness par-

ticularly at elevated temperature, and also because they are readily degassed. Molybdenum is also used for plates because of its high melting point, low gas content, and its capability for forming into useful shapes. However, its thermal emissivity or ability to radiate heat is quite poor as compared with graphite or carbonized nickel. The thermal emissivity of molybdenum plates can be improved somewhat by sandblasting them.

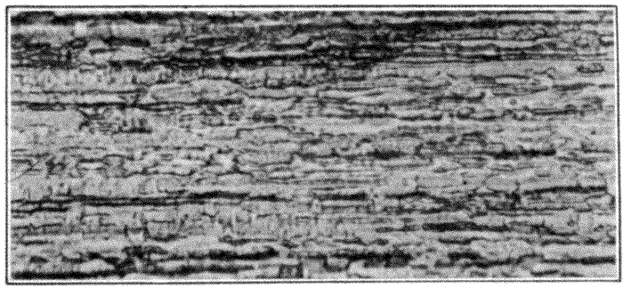
Molybdenum getter tabs are used when barium is to be flashed from barium-aluminum getter, because aluminum does not alloy with molybdenum as it does with nickel and a much higher flashing temperature can be reached.

In the manufacture of coiled-coil filaments for present-day lamps, molybdenum has become important as a mandrel material because it is soluble in dilute nitric acid while tungsten is not. During the process of making these filaments, very fine tungsten wire is wound on a molybdenum mandrel. This wire and mandrel are then wound on another molybdenum mandrel. After the wires are fired at elevated temperature to set the tungsten, the mandrels are dissolved in nitric acid. The coiled-coil filament is a more efficient light source because of the concentration of the filament into a small space.

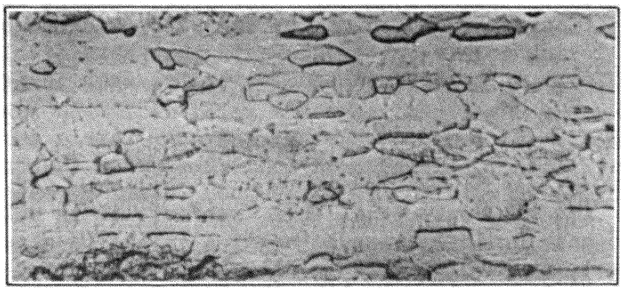
As a heater for cathodes, pure molybdenum wire is not particularly adapted, but alloys with tungsten have distinct advantages over pure tungsten and were, accordingly, introduced about five years ago. One of these from which H wire is made contains 20 per cent tungsten and another from which J wire is made contains 50 per cent. The melting point, density, thermal expansion, and stiffness of these alloys are intermediate between those of tungsten and molybdenum. H wire can be used for heaters having cores that act as supports, but coreless heaters are made with J wire which is stiffer when hot and can support the insulation without sagging. The advantage of using these alloys arises chiefly from the better mechanical properties resulting from the addition of molybdenum which gives ductility and freedom from brittleness. The factory shrinkage, especially in wire drawing and cleaning, has been reduced considerably by the use of H and J wires. Heaters made from these wires have some other properties that are not understood because they have not been thoroughly studied. When these heaters are lighted intermittently, the heater current rises above normal by an amount proportional to the percentage of molybdenum in the heater wire. This increase is not experienced with tungsten heaters. Noise and hum with H and J wires are somewhat lower than with tungsten.

Fig. 4 shows the effect of annealing tungsten, molybdenum-tungsten alloys, and molybdenum for a period of 30 minutes at 1800°C. The upper section for tungsten shows the elongated grains persisting; the center section for molybdenum-tungsten alloy shows the recrystallization is complete; and the lower section for molybdenum shows that grain growth has occurred.

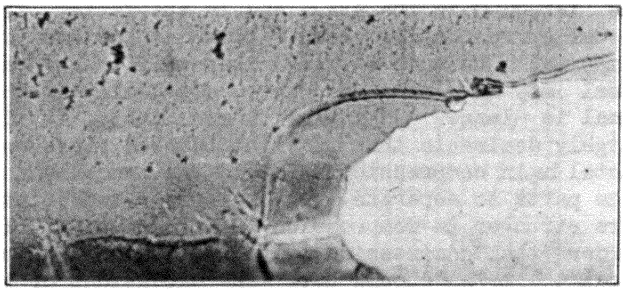
In some power tubes, molybdenum with 20 to 25 per cent tungsten is used for grid wire where more strength is required than is obtainable with molybdenum.



TUNGSTEN



MOLYBDENUM-TUNGSTEN ALLOYS



MOLYBDENUM

Fig. 4 - Effect of annealing W, Mo-W, and Mo for period of 30 minutes at 1800°C.

In general, molybdenum is useful principally because it combines the high temperature properties of tungsten with the ductility found in the ferrous and softer metals.

COPPER

Copper is used in vacuum tubes where good electrical or thermal conductivity, easy reduction of the oxide, or softness are required. The vapor pressure of copper is rather high so that the temperature of copper parts in a vacuum tube should not exceed 400° to 500°C. With the exception of aluminum, copper is more impervious to hydrogen than any other metal. Its limited sol-

ubility in iron and its suitable melting point permit it to be used as a solder for joining iron parts as mentioned previously.

Copper is readily oxidized and forms two oxides, the red (cuprous), and the black (cupric). Unfortunately, the cuprous oxide is soluble in copper and remains in the grain boundaries where it causes brittleness if the copper is heated under reducing conditions.

Copper is used for lead wires because of its flexibility, electrical conductivity, and ease of soldering. For water-cooled anodes in transmitting tubes, it is particularly useful because of its ductility, high thermal conductivity, and imperviousness to gas. In recent years, we have been using copper because of its high thermal conductivity for grid support wire in tubes where grid emission is important. The disadvantages of copper for this purpose are the lack of stiffness and difficulty of welding. The stiffness of copper can be improved somewhat by alloying copper with elements that have little effect on the thermal conductivity, such as Ag, Cd, Cr, Sn.

GLASS-TO-METAL SEALS

At the outset of this subject, it is well to make a distinction between a metal-in-glass seal, such as is made in all glass tubes and the glass-in-metal seal like the eyelet-to-glass or insert-to-glass seal of the metal tubes. The difference is that in the former case, the wire is imbedded in the glass and in the latter, the metal surrounds the glass. This distinction is important because a metal that will make one seal may not make the other, especially if the seal is large. In making these seals, it is highly desirable that the joint between glass and metal be in compression to avoid the tendency for the parts to separate. All materials and joints are stronger in compression than in tension. In order that compression can be maintained in the seal, the metal should have less expansion, or to express it more exactly, less contraction than the glass in a metal-in-glass seal; and in a glass-in-metal seal, the glass should have the lower contraction. The earliest seal-in metal was platinum, because its expansion was approximately that of glass. Platinum forms no oxides to complicate the seal and its electrical conductivity is good, being about equal to that of nickel. Platinum was used exclusively until about 25 years ago when dumet was developed by Eldred and Fink.

It happens that the thermal expansion of iron-nickel alloys varies widely with composition and that the expansion of some other alloys matches that of glasses. The nickel alloys never made good metal-in-glass seals; although with 48 per cent Ni, the expansion was about that of platinum but the electrical resistance was high and the adhesion of the glass was poor. Eldred's idea was to cover the iron-nickel alloy with a layer of copper which carried most of the current and gave a better bond with the glass. When the copper was added, the expansion was increased and as a result the nickel content of the alloy was reduced to 42 per cent. Unfortunately, such a compound wire which we call dumet, does not have a simple expansion characteristic. The expansion along the wire is about 25 per cent less than the radial expansion. The expansion in length is 7.6 x 10^{-6} /°C and the radial expansion is 10.25 x 10^{-6} /°C. Accordingly, it is necessary to keep these seals small by using a small wire of short length. In order to improve the adhesion between glass and dumet and to protect the latter from excessive oxidation during sealing-in, the wire is covered with a layer of borax.

Chrome iron containing 27 per cent chromium has very good adhesion to glass and an expansion of 10.5 to 11.0 x 10^{-6} /°C which is higher than that of the soft glasses. This makes it suitable for glass-in-metal seals. Its electrical resistance is high so that it is not desirable as a conductor for a metal-in-glass seal. More chromium might be added to iron to reduce the expansion but such metal, with present mill practice, is unworkable. Iron with 17 per cent chromium has about the same expansion as iron with 27 per cent chromium, but the expansion increases slightly faster at the higher temperatures, and every seal made with it and No.12 glass cracks.

In making chrome-iron-to-glass seals, it is our practice to weld the chrome-iron ring into the steel header or shell and then bright-anneal the latter. This procedure cleans the steel shell but covers the chrome iron with chromium oxide. The glass is then sealed into the insert or ring by pressing a mass of pasty glass into it. The seal is very strong. Under comparable conditions, seals with nickel-iron are much weaker and frequently fail. Fig. 5 shows the expansion curves of a soft glass and of the metals which are used with it. Soft glasses have annealing temperatures of about 430°C, and it is desirable that the expansion of the glass and of the metal be the same up to 430°, if the seal is to be free from strain. This is seldom possible because the expansions of glasses and of metals do not match too well, and the problem is to keep the deviation as low as possible.

Since the match between the metal and glass is not good and often in the wrong direction, the seal may be subjected to alternate tension and compression. Furthermore, in the manufacture of metal tubes where the glass and metal are hardly ever at the same temperature until the tube is finished, it seems somewhat useless to attempt an exact match in the expansion characteristics. Due to these differences in temperature, the strength of the bond between the glass and the metal is the limiting factor. This accounts for the failure of the 17 per cent chrome-iron and of the 50 per cent nickel-iron when sealed to soft glasses.

We believe that the strength of the bond is

governed entirely by the adhesion of the metallic oxide to the metal which in turn is a function of the thickness of the oxide. During the stem making, the metal is oxidized and 27 per cent chrome iron does not oxidize so readily especially if it has been preoxidized in a partially burnt gas atmosphere which oxidizes the chromium readily but not the iron. Iron with 17 per cent chromium oxidizes more readily, and copper or nickel-iron

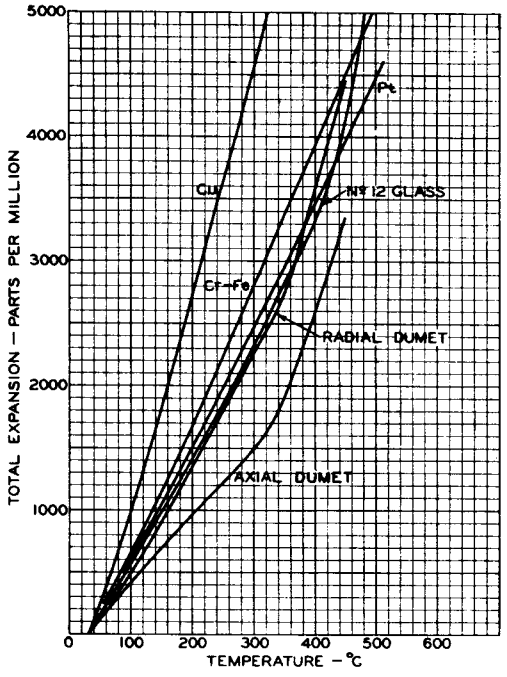


Fig. 5

produce large quantities of oxides unless specially protected from oxidation. The glass is a solution of a number of oxides and has a tendency to dissolve the oxides of the seal-in metals but the rate of solution is too slow to accomplish the dissolving during the period of sealing-in. Consequently, the seal consists of metal to metallic oxide to glass in which the adhesion at each interface must be strong enough to resist the stress introduced by the temperature gradients. When the layer of oxide becomes heavy, its thermal expansion which differs from that of the metal exerts enough force to separate the oxide from the metal. This separation results in fissures because the oxides are brittle and lack flexibility.

With hard glasses like Nonex, tungsten is quite satisfactory but it must be protected from excessive oxidation by beading the wire before sealing-in. This operation of beading controls the amount of oxide formed by restricting the convection of air around the wire. The oxide of tungsten is quite volatile so that the accumulation is negligible, but erosion of the surface should be avoided. Tungsten has the virtue of high electrical conductivity, but it is too brittle and quite expensive.

With the development of the metal tube, Fernico and molybdenum were used with the hard glass called G705BA, or No.704. This glass has an expansion somewhat greater than that of Nonex, and so molybdenum was found suitable for the metal-in-glass seal. The glass-in-metal seal was made with eyelets of Fernico, or Kovar. These cobalt-nickel-iron alloys contain 18 per cent Co, 28 per cent Ni, and small amounts of the impurities ordinarily present in steel.

Cobalt, when substituted for nickel, has the property of reducing the thermal expansion of the corresponding nickel-iron alloy, and of thus permitting it to be used with hard glasses which have lower expansion than soft glasses. Furthermore, adhesion is improved by the presence of cobalt in the joint, whether it is in the glass or in the metal.

Fig. 6 shows the thermal expansion of the hard glasses and of the metals which are ordinarily used with them. The best seals are those in which the expansions of the glass and metal are identical below the annealing temperature of the glass. Above this temperature, the strains caused by inequalities of expansion would disappear in time. Because this condition of identical expansion is very difficult to attain, a compromise is made by matching the average thermal expansion between room temperature and the annealing temperature of the glass.

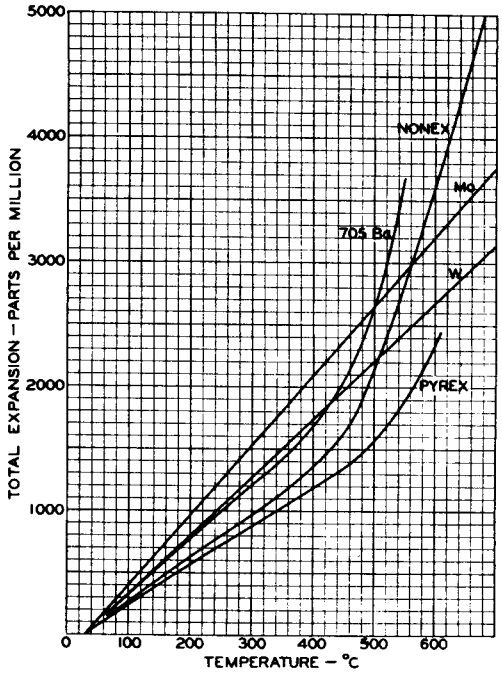


Fig. 6

With larger seals, this compromise matching is not sufficient and so Fernico was developed to give a more nearly perfect match with glass. Fernico is a distinct improvement but the match

with glass is not exact and various heats of Fernico differ in expansion. The reason for this is that small amounts of some impurities or slight variations in the major constituents have a large effect on the thermal expansion. Carbon has the greatest effect and its amount is difficult to control because it is easily introduced or lost during heating operations. Variations in annealing and amount of cold work have an effect on the expansion of Fernico, but these are not important except in large seals.

Lecture 9

GLASS AND ITS PROPERTIES

G. R. Shaw and C. A. Jacoby

DEFINITION

Various definitions of glass have been given by different authorities but probably the most logical one for our purpose is by Littleton and Morey, who state that glass is an industrial material obtained by melting together various inorganic oxides (or compounds that yield oxides by decomposition during the melting process) and cooling the resulting solution so that crystallization does not take place.

COMPOSITION

The various physical properties of glass are determined primarily by the chemical composition of it. The glasses commonly used for radio tubes may be divided generally into soft and hard glasses depending upon the temperature required for working them.

In the soft-glass group, we have the soda-potash-lead glasses identified as No.1, No.12, and No.816 (814KW), together with the soda-lime glass identified as No.8. The soda-potash-lead glasses are used for tube stems where dumet lead wires are sealed into the glass stem press. They may also be used for chrome-iron seals, as in button stems. The soda-lime glass is used for bulbs of radio tubes which are sealed to lead-glass stems. This glass is also used for some special tubes, such as gas phototubes. In this case, it is used because it is not attacked by the caesium.

In the hard-glass group, we have the lead-borosilicate glass No.772 (702P), and the boro-silicate glasses No.774 (726MX) and No.704 (705BA). The lead-borosilicate glass has the trade name Nonex and is used for tube stems where tungsten lead wires are sealed into the glass stem press. It is also used for bulbs. The borosilicate glass 774 (726MX) is a chemical-resistant glass and has the trade name Pyrex. It is used for transmitting and cathode-ray tube bulbs. The 704 (705BA) glass is used with Fernico metal seals and molybdenum lead-wire seals in metal tubes.

All of these glasses contain, and have as their base, silica which is obtained from sand. The amount ranges from 57 to 72 per cent in the soft glasses and from 72 to 81 per cent in the hard glasses. Another important constituent common to all the glasses is sodium oxide. The amount ranges from 3 to 4 per cent in the No.816 glass and is about 17 per cent in the No.8 glass. Of all the constituents customarily present in glass, sodium oxide is the one which does most to decrease the resistance of the glass.

All of these glasses contain some potassium oxide. The amount ranges from that present as an impurity to 10 per cent in some of the soda-potash-lead glasses. It serves approximately the same purpose as sodium oxide in respect to lowering the viscosity and raising the coefficient of expansion, but gives a substantially higher resistance than does an equivalent amount of sodium oxide.

The hard glasses are essentially borosilicates. Pyrex contains about 81 per cent of SiO_2 and somewhat over 10 per cent of B_2O_3. This composition serves as the base for the rest of the series, each one being suitably modified to accomplish the purposes for which it is intended. Nonex, for example, has a percentage of the silica replaced by lead oxide and the result is a glass in which the resistance is much improved, even though the viscosity range is lowered somewhat.

In general, the chemical compositions of the different glasses may be represented by the following equations, where x, y, and z are variable percentages of the various constituents.

Lead Glasses

$$x\ SiO_2 + y\ PbO + z\ Na_2O + \ldots.$$

Lime Glasses

$$x\ SiO_2 + y\ CaO + z\ Na_2O + \ldots.$$

Borosilicate Glasses

$$x\ SiO_2 + y\ B_2O_3 + z\ Na_2O + \ldots.$$

CONSTITUTION

Recent speculations on the constitution of glass based on x-ray evidence and studies of glass by means of x-rays indicate that the glassy state is distinguished by a continuous, randomly oriented network of atoms. Picture diagrams of soda-silica glass and lead-oxide-silica glass are shown in Figs. 1[1] and 2[2], respectively.

PRODUCTION OF GLASS PARTS

Lime-glass bulbs in large quantities are produced largely on the Corning, continuous-ribbon-type, bulb-blowing machine. Glass flows from a melting tank to a forehearth, thence through a

---

[1] B. E. Warren and A. D. Loring, "X-ray Diffraction Study of the Structure of Soda-Silica Glass," Jour. Amer. Ceramic Soc., Vol. 18, No.9, p. 275; 1935.

[2] George J. Bair, "The Constitution of Lead-Oxide Silica Glass," I. Atomic Arrangement, Jour. Amer. Ceramic Soc., Vol. 19, No.12, p. 346; 1936.

vertical orifice, and down between two rollers which deliver a flat ribbon of glass to a horizontal ribbon conveyor. As the glass ribbon progresses, the glass sags through holes spaced about 3.9 inches on centers in the plates of the conveyor. Blowheads traveling on top of the ribbon conveyor blow air down into the holes and form partially blown bulbs. At the same time,

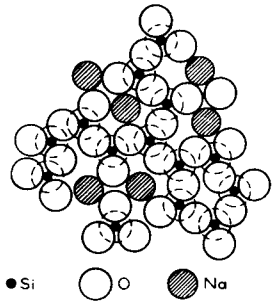


Fig. 1 - Schematic arrangement of atoms in a soda silica glass.
(Courtesy of American Ceramic Society)

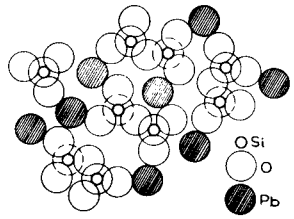


Fig. 2 - Atomic arrangement in lead-oxide silica glass.
(Courtesy of American Ceramic Society)

molds traveling below meet and close around the partially blown bulbs into which air continues to be blown while the molds revolve around the bulb to prevent mold marks. In the final operation, the blowheads leave the ribbon, the molds open up, and the bulb is exposed to be cracked away from the ribbon by a hammer mechanism. The unused glass or cullet is returned to the tank for remelting while the bulbs are transferred by conveyor belt to an annealing oven. A more detailed description has been given by F. W. Preston.[3]

Lead-glass tubing is drawn continuously on the Danner-type, tube-drawing machine. Molten glass flows from a forehearth over a weir and down to a large, revolving, tapered mandrel set at an angle. Air is blown through the hollow mandrel as it revolves and the hollow cylindrical form of glass is drawn down to the proper size as it cools into a continuous piece of tubing. The diameter and wall thickness are controlled by the rate of flow of glass, speed of mandrel, rate of draw, and air pressure. At the end of the drawing operation, the glass is cut into lengths or sticks by an automatic cutting wheel. The tubing is then gauged, weighed, and sorted before being wrapped and packed in bundles.

The glass-melting process consists of a fusion of all batch particles into a homogeneous mass. For closed pots, a temperature of 1400°C is the practical limit, but open pots having no crown to be considered may be somewhat hotter. In tanks, temperatures go as high as 1500°C or slightly more, with occasional instances of the more refractory glasses requiring even higher temperatures. The degree of fluidity attained is dependent on the method of fabrication to be employed.

The following tabulation shows the approximate working temperatures for soft glass and hard glass during the manufacturing operations of flare making, stem making, and sealing-in.

Operation	Working Temperature °C	
	Soft Glass	Hard Glass
Flare Making	800	1000
Stem Making	1000	1250
Sealing-In	1050	1300

At one time, it was believed that the physical properties of glasses were affected by the thermal treatment of the glass in the molten condition. Later, it was discovered that dissolved clays of the melting container caused the formation of a glass of a new composition and consequently of new physical properties.

Glass will crystallize in time below some definite temperature called the devitrification or liquidus temperature. If fabrication can be accomplished above this temperature, then there is slight danger from devitrification during cooling, since the cooling of the fabricated article is always rapid until the annealing zone is reached. Commercial glasses, of course, will not devitrify in the annealing zone in the time required to anneal. The glasses more fluid at the liquidus temperature are more easily devitrified.

PHYSICAL PROPERTIES

The chart shown in Table I is for general reference and shows the principal physical properties of the glasses used in radio-tube manufacture. From the columns for the refractive index and density, it will be noted that the refractive index is higher in the case of the heavier glasses.

A. Viscosity

The viscosity of a glass is of prime importance because it determines the working range of temperature during the fabrication of the article to be made. Viscosity is that property of a mater-

[3] F. W. Preston, "New Lamps for Old," Glass Industry, Vol. 12, No.8, pp. 159-165; August, 1931.

Table I — PHYSICAL PROPERTIES OF GLASS

Glass	Type	Use	Density g/cm³	Softening Point °C	Annealing Point °C	Strain Point °C	Expansion Coefficient/°C 0° - 310°C	Resistance Megohms/cm³ 250°C	Refractive Index
1	Lead	Stem tubing	2.85	626	425	389	90 x 10^{-7}	1190	1.542
8	Lime	Bulbs	2.47	696	510	475	92 "	2.26	1.512
12	Lead	Stem tubing	3.04	630	431	395	87 "	11900	1.557
814KW (816)	Lead-barium	Stem tubing	2.99	630	433	398	92 "	36300	1.553
726MX (774) (Pyrex)	Hard	Bulbs	2.23	818	553	510	32 "	263	1.474
702P (772) (Nonex)	Hard	Stem tubing	2.35	756	521	486	36 "	1170	1.487
705BA (704)	Hard	Seals with Fernico	2.24	697	484	450	46 "	5300	1.480

ial or mixture which resists flowing when a force is exerted upon it. The unit of viscosity is the poise. It is defined as the force exerted tangentially on one square centimeter of either of two horizontal planes one centimeter apart, which will move the one plane at the rate of 1 cm/sec. with reference to the other, the space between the two planes being filled with the viscous liquid. Expressing this definition by an equation, we have

$$p = \frac{F}{VA}$$

where,

p = viscosity in poises
F = force in dynes
A = surface in sq.cm.
V = velocity in cm/sec.

1) Fabrication — At the fabrication point, or the time when the glass is drawn from the pot or tank, its viscosity should be 10^5 poises. For purposes of comparison, the viscosities of some other substances at room temperature are:

Substance	Viscosity Poises
Water	0.01
Glycerine	10
Glucose	1000

2) Softening Point — Glass has no definite melting point but the softening point of all glasses is designated in this country as the temperature corresponding to a definite viscosity of 4.5 x 10^7 poises. As a check on the manufacturing operations in the glass works, the following test has been installed. The softening point is determined as that temperature in degrees Centigrade at which a glass fiber (diam. = 0.55 to 0.75 mm ± 0.01 mm; length = 23.5 cm) will elongate under its own weight in a special furnace at the rate of 1 mm/min. Commercial variations permitted in the value of this constant are: ± 5°C for glasses having a coefficient of expansion per °C above 50 x 10^{-7}, and ± 10°C for glasses below 50 x 10^{-7}.

3) Annealing Point — The annealing point is designated as that temperature corresponding to a viscosity of 2.5 x 10^{13} poises. It is also defined as the temperature at which 90 per cent of the strain in the glass is removed in 15 minutes in a glass sample 1/4 inch in thickness. This temperature is sometimes called the upper annealing temperature.

4) Strain Point — The strain point is designated as that temperature corresponding to a viscosity of 4.0 x 10^{14} poises. It is also defined as the temperature at which 90 per cent of the strain is removed in 4 hours in a glass sam-

ple 1/4 inch in thickness. This temperature is sometimes called the lower annealing temperature below which permanent strain is not introduced at ordinary rates of cooling and above which annealed glass cannot be reheated without introducing permanent strain and requiring a reannealing to remove the strain.

5) Viscosity Curves — Fig. 3 shows the viscosity curves of glasses No.1, No.8, No.12, Nonex, and Pyrex. The temperatures at the fabricating point of 10^5 poises for each of the glasses are indicated on the curves as are also the temp-

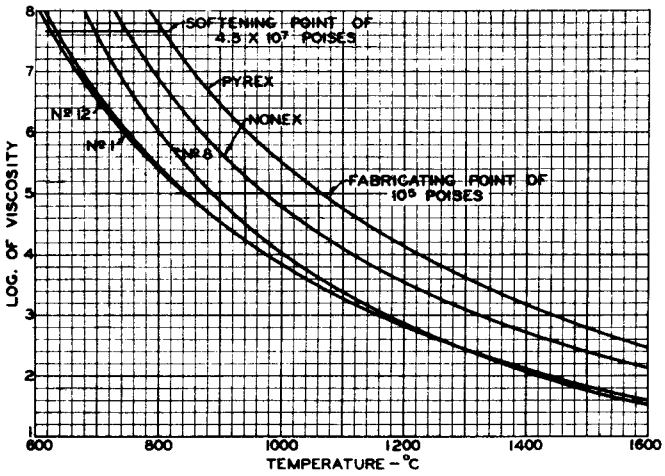


Fig. 3 - Viscosity curves of several glasses.

eratures at the softening point of 4.5×10^7 poises. Fig. 4 gives graphical data on a continuation of the viscosity of the various glasses down into the lower temperature ranges. These curves are logarithmic and are of the form

$$\log p = \frac{A}{T} + B \text{ , or } p = B\,\varepsilon^{\frac{A}{T}}$$

where,

p = viscosity in poises
T = absolute temperature in degrees
A and B = constants.

6) Annealing — The term annealing as used in connection with glass is not the same term as is used in connection with metals which lose their hardness when heated. When glass is annealed, the internal stress is removed without the glass becoming soft. Technically, annealing consists of the removal of a portion of the stress by a heating schedule and then cooling the glass at a certain rate so as to allow a desired amount of stress to remain in the glass at room temperature.

The terms stress and strain are frequently used when annealing is discussed. Stress is the force per unit of area. Strain is a relative deformation resulting from a stress. The relation between stress and strain is given by the equation

$$F = SE$$

where,

F = the stress in kg/cm^2
S = the strain in cm/cm
E = the modulus of elasticity in kg/cm^2

The study of strains in glass involves the theories of elasticity and heat conduction. When a slab of glass of certain thickness is heated, a temperature gradient is set up by the conduc-

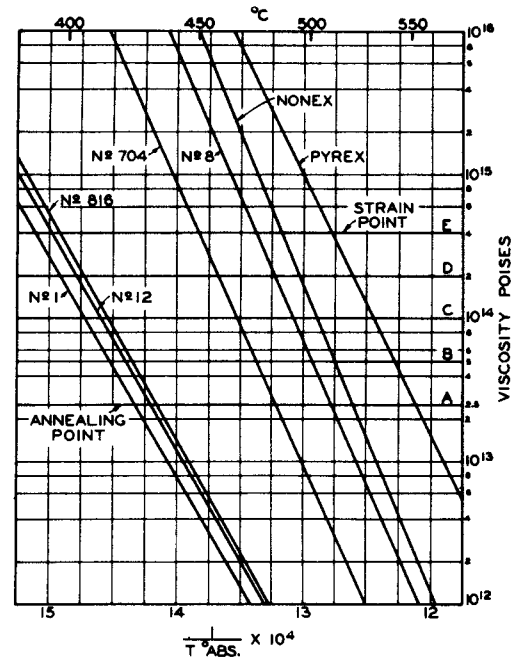


Fig. 4 - Viscosity of glasses between annealing and strain points. The intersection of lines A, B, C, D, and E with the viscosity curves for the various glasses indicate the temperature intervals at which the viscosities are doubled, as shown in the tabulation immediately preceding Table II.

tion of heat from the outer to the inner layers through the glass. The outer layers are hottest and any movement is resisted by the inner layers of the glass. The result is a compression in the outer layers and a tension in the middle. If, now, the temperature is allowed to become uniform throughout, then the thermal gradient, and hence the stress, will disappear. When the glass is cooled, the reverse condition is true. There is a tension in the outer layers and a compression in the middle. Strain introduced in the glass by removal of the temperature gradient is called ordinary or permanent strain. Strain introduced by viscous yielding of the glass when it is beginning to cool is called reverse or temporary strain. Glass may receive a permanent strain when cooled from a high temperature down to a room temperature.

Fig. 5 shows a piece of glass 2 centimeters thick cooled from a high temperature, i.e., above the annealing range. A permanent strain is introduced when the temperature gradient is smoothed out as the glass cools to room temperature. If the amount of permanent strain exceeds the strength of the glass, a fracture will occur. The object of annealing, then, is to reduce the permanent strain in the glass to a safe limit.

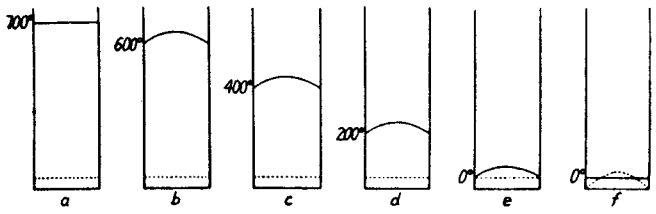


Fig. 5 - Diagram showing in several stages how a slab of glass cooled at a uniform rate from a sufficiently high temperature receives an internal strain only when the cooling is stopped. The abscissae represent distances in a direction normal to the surface, and the ordinates represent temperature (full lines) or strain (dotted lines). (From Jour. Franklin Inst., Vol. 190, No.5, page 602; 1920).

If the glass is heated to some moderate temperature at the lower end of the annealing range, but without thermal gradient, a reverse or temporary strain is introduced at the beginning of the cooling which will exactly balance the ordinary or permanent strain set up when the temperature gradient is smoothed out as the glass cools to room temperature, as shown in Fig. 6. The resultant strain is zero. Glass cooled from some intermediate temperature will acquire internal strain, the amount of which will lie between the amounts received in the previous two cases discussed.

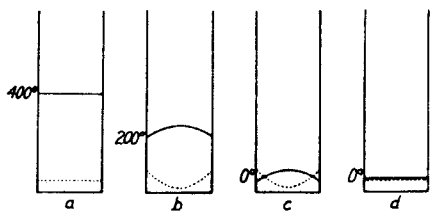


Fig. 6 - Similar to Fig. 5 except that the glass is cooled from a temperature below the annealing range, in which case no permanent strain is introduced. (From Jour. Franklin Inst., Vol. 190, No.5, page 603; 1920).

In general, the strain remaining in a piece of glass is equal and opposite in sign to the reverse strain lost by viscous yielding of the glass in the early stages of the cooling process.

The standard annealing point has been defined as the temperature for annealing a piece of glass 1/4 inch thick in 15 minutes. Doubling the thickness decreases the annealing temperature by twice the temperature interval necessary to double the viscosity. Doubling the thickness increases the annealing time by 4, or the annealing time is proportional to the square of the thickness of the glass. It will be noted that what is called the annealing point is really the point at which the viscosity is 2.5×10^{13} poises. Physically, this represents the point at which there is sufficient mobility within the glass structure to permit the movement necessary to relieve any strains set up within a definite period of time (15 minutes for a glass thickness of 1/4 inch). Since the rate of release of strain increases so rapidly above this point, the same care does not need to be taken in the region above as in the region below. When the viscosity increases to 4×10^{14} poises, or 16 times at the annealing point, the time required for the release of any strains has increased by the factor of 16, or by 4 hours.

The temperature intervals for doubling the viscosities of the different glasses are shown in Fig. 4. The temperature interval to double the viscosity is

$$T = \frac{\text{Annealing Point} - \text{Strain Point}}{4}$$

Example: Temperature interval to double viscosity for No.12 glass is

$$T = \frac{431^{\circ}C - 395^{\circ}C}{4} = \frac{36}{4} = 9^{\circ}C$$

The temperature intervals for the different glasses are listed in the following tabulation.

Glass	Temperature Interval(°C)
No.1	9
No.12	9
No.816	8.75
No.704	8.5
No.8	8.75
Nonex	8.75
Pyrex	10.75

The annealing temperatures and times for the various glasses of different thickness are indicated in Table II.

The initial cooling rate for annealing a piece of glass is derived from the fundamental

Table II

Glass Thickness		Annealing Time	Annealing Temperature of Various Glasses °C						
inches	mm	min : sec	No.1	No.12	No.816	No.704	No.8	Nonex	Pyrex
.50	12.70	60 : 00	407	413	415	467	492	503	531
.25	6.35	15 : 00	425	431	433	484	510	521	553
.125	3.17	3 : 45	443	449	451	501	528	539	575
.0625	1.58	0 : 56	461	467	467	518	544	555	596
.0312	0.79	0 : 14	479	485	485	535	562	573	618

equations by Adams and Williamson.[4] The cooling rate h_o in °C per minute is given by

$$h_o = \frac{BF}{C} = \frac{\Delta N_s}{C}$$

where,

B = birefringence constant
F = stress in kg/cm²
ΔN_s = final stress in optical units
= 2.5 millimicrons per centimeter of glass thickness
C = constant

$$= \frac{\alpha B a^2}{6\kappa\left(\frac{1}{6R} + \frac{2}{9K}\right)} \times 10^7$$

$$= 3.2\ a^2 \text{ for Pyrex}$$

where,

α = coefficient of expansion/°C
κ = thermal diffusivity
$$= \frac{\text{heat conductivity}}{\text{density} \times \text{specific heat}}$$
$$= \frac{K}{d\ s}\ (\text{cm}^2/\text{sec})$$
R = modulus of rigidity
K = modulus of compressibility
a = semi-thickness of glass

If we assume that a piece of Pyrex glass has a semi-thickness of 0.5 centimeter,

$$C = 3.2 \times 0.25 = 0.8$$

Substituting, we find that the initial cooling rate for the piece of Pyrex 1 centimeter thick is

$$h_o = 2.5/0.8 = 3.12°C/\text{min.}$$

After the initial cooling rate, the rate of cooling is given by

$$h_x = \frac{h_o}{2}\left(1 + 2^{\frac{T_o - T_x}{20}}\right)$$

where,

h_x = new cooling rate in °C/min.
T_x = new temperature in °C
h_o = original cooling rate in °C/min.
T_o = original temperature in °C

Assuming temperature intervals of 20°C, we can calculate the following values for h_x. The actual time of cooling per interval is obtained by dividing the temperature interval by the rate of cooling for that interval.

Cooling Rate (°C/min)		Time Interval (min)
h_o	$= 3.12$	6.41
$h_{20} = \frac{h_o}{2}(1 + 2) = 1.5\ h_o$	$= 4.68$	4.27
$h_{40} = \frac{h_o}{2}(1 + 4) = 2.5\ h_o$	$= 7.80$	2.56
$h_{60} = \frac{h_o}{2}(1 + 8) = 4.5\ h_o$	$= 14.05$	1.42
$h_{80} = \frac{h_o}{2}(1 + 16) = 8.5\ h_o$	$= 26.50$	0.75
$h_{100} = \frac{h_o}{2}(1 + 32) = 16.5\ h_o$	$= 51.50$	0.38
$h_{120} = \frac{h_o}{2}(1 + 64) = 32.5\ h_o$	$= 101.40$	0.19

[4] Adams and Williamson, "Annealing of Glass," Jour. Franklin Institute, Vol. 190, No.5, pp. 597-631; No.6, pp. 835-870; 1920.

L. H. Adams, "The Annealing of Glass as a Physical Problem," Jour. Franklin Institute, Vol. 216, No.1, pp. 39-71; 1933.

The maximum cooling rate for Pyrex is given by the equation

$$h_p = \frac{200}{C} = \frac{200}{0.8} = 250°C/min.$$

The cooling rates are inversely proportional to the square of the glass thickness, and may be expressed by the formula

$$h = \frac{\Delta N_s}{C} = \frac{\Delta N_s}{3.2\ a^2}$$

The cooling rates for several thicknesses of Pyrex glass at temperature intervals of 20° are shown in Table III.

Fig. 8 shows a typical annealing curve of the rotary electric annealer as used in our plant for annealing glass stems, and is illustrative of the type of heating and cooling required in such practice.

B. Coefficient of Expansion

The expansion curves of the various glasses are shown in Fig. 9. The expansion is usually given for the range between 0° and 310°C, but consideration must be given to the entire curve when it is being used to select a glass for sealing to a given metal or ceramic.

The glasses used in radio-tube manufacture may be classified into two groups according to whether their coefficients of expansion per °C are above

Table III

Pyrex Glass Thickness (mm)	Cooling Rate (°C/min) — Temperature Intervals of 20°C						
	h_o	h_{20}	h_{40}	h_{60}	h_{80}	h_{100}	h_{120}
20	0.78	1.17	1.95	3.51	6.62	12.87	23.30
10	3.12	4.68	7.80	14.05	26.5	51.50	101.40
5	12.48	18.72	31.20	56.20	106.0	206.0	-
2.5	49.92	74.88	124.8	224.8	-	-	-
1.25	199.68	299.5	-	-	-	-	-

Fig. 7 shows a typical annealing curve for Pyrex glass 1/4 inch thick plotted from Corning data. The glass is heated and cooled from one side only, as in a hollow cylindrical tube. In this case, the heating time and cooling time are doubled to allow for heating and cooling from the outside surface.

or below 50 x 10^{-7}. The lead and the lime glasses fall under the first group with a range of from 87 x 10^{-7} to 92 x 10^{-7}. The borosilicate glasses fall under the second group with a range of from 32 x 10^{-7} to 46 x 10^{-7}. The variation permitted in the expansion of a given glass is ± 2 x 10^{-7} for glasses above 50 x 10^{-7} and ± 1 x 10^{-7} for

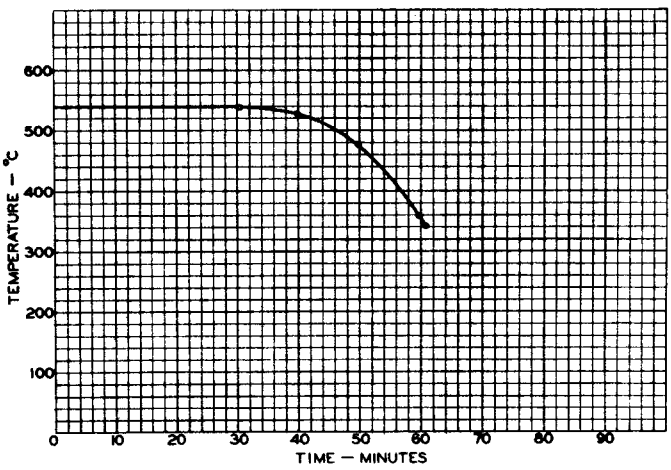


Fig. 7 - Typical annealing curve for 1/4"-thick Pyrex glass.
(After Corning data)

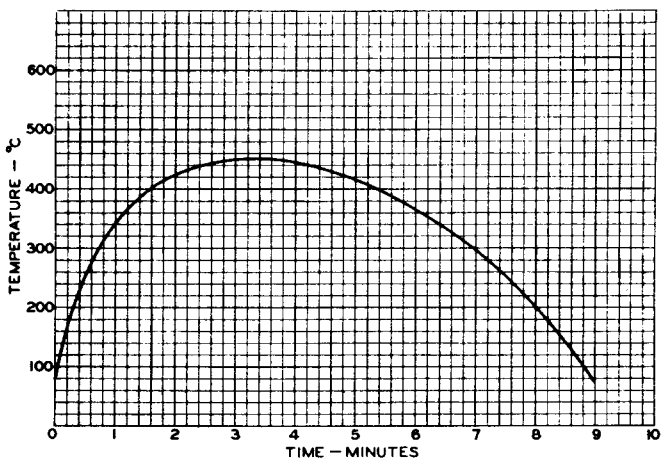


Fig. 8 - Typical annealing curve for glass stems.

glasses below 50 x 10^{-7}.

In the design of glass-metal seals, the selection of the proper glass to match the metal depends on the type of seal to be made. In general, it is important to remember that the glass should always be left in a state of compression when the

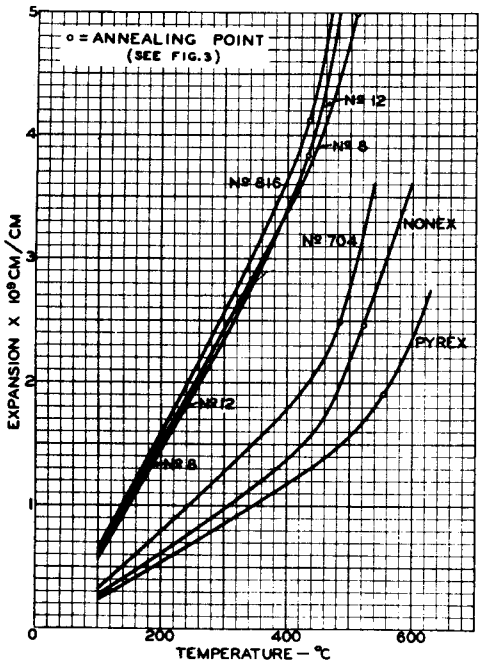


Fig. 9 - Expansion curves of various glasses.

seal is cooled down to room temperature, provided the safe compressive strength is not exceeded which might cause a failure of the glass seal. The points marked on the expansion curves in Fig. 9 are the standard annealing points given in Table I.

C. Thermal Endurance of Glass

Thermal endurance is that property of glass by which it can withstand sudden changes of temperature without breaking. Denoting this property by F, we have

$$F = \frac{P}{\alpha E}\sqrt{\frac{K}{d\,s}}$$

where,

P = tensile and compressive strength in kg/cm²
α = coefficient of expansion/°C
E = elastic coefficient (Young's modulus) in kg/cm²
K = thermal conductivity in cal/cm/°C/sec.
$\frac{K}{ds}$ = thermal diffusivity in cm²/sec.

Each of these factors will now be discussed.

1) Tensile and Compressive Strength — When stresses exceed the tensile or the compressive strength of the glass, fracture will occur. When sudden cooling takes place, the tension in the surface layers is the greater stress. Sudden heating causes a high surface compression. The safety of glass depends upon its tensile strength because the compressive strength of glass is always greater than the tensile strength. It is evident that glass will more readily withstand sudden heating than sudden cooling. Values for tensile strength and compression strength are not well coordinated, but for purposes of any ordinary calculations, the following values may be used.

Glass	Tensile Strength		Compressive Strength	
	lbs./in.^{2}	kg/cm^2	lbs./in.^{2}	kg/cm^2
Hard	10000	700	170000	1200
Soft	6000	420	114000	800

An inspection of the figures shows immediately why compressive strains are more desirable than tensional strains.

2) Coefficient of Expansion — Sudden cooling of glass causes the surface layers to rapidly contract and throws them into a state of tension, while at the same time the interior is thrown into a state of compression as previously outlined. The reverse is true when the glass is suddenly heated. This action is best explained by the fact that any movement of the outer layers is resisted by the inner layers of the glass. Consequently, there exists a neutral zone where there is zero stress between the inner and outer layers during heating or cooling. The greater the rate of thermal expansion or contraction, the more intensely will glass be stressed. Glasses with relatively low thermal expansion will show the highest thermal endurance. The thermal endurance of the various glasses is inversely proportional to their coefficients of expansion by actual test and is so indicated in the above formula.

3) Elastic Coefficient (Young's Modulus) — Low elasticity favors thermal endurance since the stress set up by local strains will be less than in glass with a high elastic coefficient. Values for elasticity are as follows:

Glass	Elastic Coefficient	
	lbs/in.^{2} x 10^{-6}	kg/cm^2 x 10^{-6}
Hard	9.23	0.650
Soft	8.88	0.625

4) Thermal Conductivity — High thermal conductivity favors thermal endurance since heat suddenly applied locally is rapidly transferred through the glass and stresses due to unequal expansion are minimized.

5) Thermal Diffusivity — Thermal diffusivity denotes the rate at which the interior of the glass is heated or cooled from the outside. It depends on the thermal conductivity (K), the density (d), and the specific heat (s).

The thermal endurance depends upon the square root of the diffusivity. A glass having four times the diffusivity of another glass will consequently have twice the thermal endurance of the other glass. The higher the diffusivity, the higher the thermal endurance, everything else being equal.

A list showing the thermal endurance of the various glasses in percentage of that of Pyrex is given below:

Glass	Thermal Endurance %
Pyrex	100.0
Nonex	89.5
No.705 BA	71.7
No.8	27.7
No.12	20.6
No.1	20.5
No.814 KW	19.3

D. Weathering

One of the sources of failure of glass is the weathering which occurs in the case of exposure to the atmosphere. The atmosphere always contains moisture and gases, such as carbon dioxide. In cities, small amounts of ammonia, sulpheretted hydrogen, and sulpherous and sulphuric acids are also usually present in the atmosphere. After more or less prolonged exposure to the atmosphere, the surface of window and similar types of glass is roughened and becomes coated with an iridescent or opalescent coating. With optical glass, atmospheric weathering is of extremely great importance. The formation of a slight film on lenses is often disastrous in its results on expensive and delicate optical combinations. The matter is further complicated by the fact that glass of relatively poor chemical stability has to be used sometimes in order to obtain other desired properties.

The nature of weathering and chemical attack is worth considering. Moisture in contact with glass is first of all adsorbed and then absorbed. The resulting hydration is succeeded by the hydrolysis of the soluble silicates forming part of the complex silicate mixture of which glass consists. In a soda-lime glass, such as No.8 lime-glass bulbs, sodium metasilicate is the most soluble silicate present and hydrolysis chiefly results in the liberation of sodium hydroxide and silicic acid according to the equation:

$$Na_2SiO_3 + (x + 2)\ H_2O = H_2SiO_3 \cdot xH_2O + 2NaOH$$

The caustic soda passes into solution while the silicic acid (which contains an uncertain number of molecules of water of hydration) is also to some extent soluble, and again is partly mechanically removed, often giving the water a turbid appearance. The action of the water is not always on the surface, for it can be shown that in many cases the water penetrates to an appreciable depth below the surface, which, on subsequent heating and drying, shows a large number of minute cracks giving the glass a dull appearance.

The attack of a moist atmosphere upon glass results in the absorption of appreciable amounts of water, and it is found that glasses best resisting the attack of water condense the least moisture upon their surfaces. The condensed water attacks the glass, breaks down the silicates, and leaves a solution of caustic alkali on the surface of the glass as previously explained. When the air contains carbon dioxide, a further reaction occurs between the gas and the alkali, and results in the formation of a carbonate. In the case of soda glasses, sodium carbonate will result according to the equation:

$$2NaOH + CO_2 = Na_2CO_3 + H_2O$$

The sodium carbonate so produced, when appreciable attack occurs, crystallizes out in fine needles upon the surface and the white deposit results in the dimming of the glass. By rubbing off the crystals with a dry cloth, the surface of the glass may be badly scratched, but the deposit can be removed without damage by washing with water. Washing, however, will not remove the surface layer of silica obtained by the hydrolysis, so that a filmed glass surface can never be satisfactorily cleaned with water alone.

Atmospheric attack on potash glasses results in the formation of potassium carbonate, a substance so deliquescent that it does not crystallize out, but leaves a moist, greasy deposit upon the glass when the surface is badly attacked. The attack of moist air on alkali-silicate glass is intensified by heating and effects have been made visible in 12 hours at 80°C in a moist atmosphere. Glasses high in lead-oxide or barium-oxide content are susceptible to attack by acid vapors in moist atmospheres and show filming and dulling of the surface.

ELECTRICAL PROPERTIES

A. Electrical Resistance

The resistance of the various glasses is shown in Fig. 10. It will be noted that the resistance of No.8 lime glass is very low in comparison with the resistances of the other glasses, in fact, it is only about 5 per cent of the resistance of Pyrex at a temperature of 250°C.

The resistance depends upon the chemical composition and glasses which contain the least sodium generally have the highest resistance because, in the range of chemical compositions, sodium has the highest conductivity.

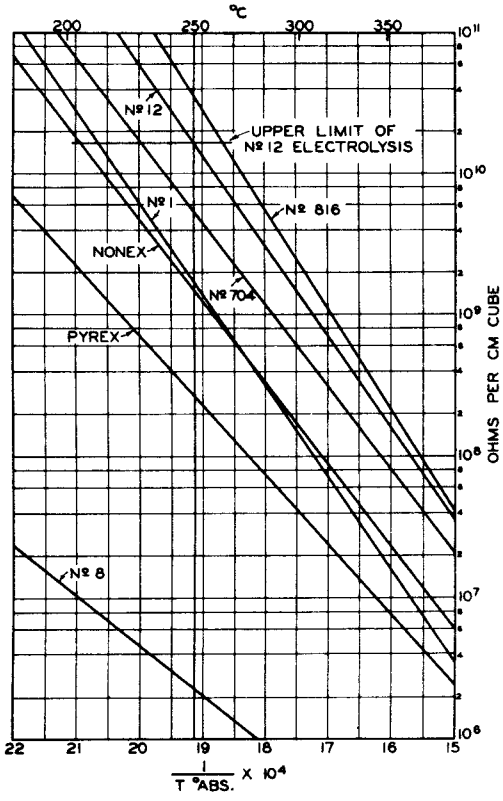


Fig. 10 - Resistance of various glasses.

The specific resistance of glass is given in terms of ohms per cm cube over the range of operating temperatures used in radio tubes. The maximum resistance of each glass at a given temperature is not limited in manufacture. The minimum is minus 10 per cent for Nos.12, 704, and 816 glasses and minus 25 per cent for all other glasses. It will be noted that Nonex, although having a relatively high softening point and low expansivity, is not any better than No.1 lead glass as far as resistance is concerned.

The temperature of glass also determines its electrical resistance as shown on the graphs which are straight lines according to the equation

$$\log R = \frac{A}{T} + B \text{ , or } R = B\,\varepsilon^{\frac{A}{T}}$$

where,

R = specific resistance in ohms/cm^3
T = absolute temperature in °C
A and B = constants

The condition of strain in the glass also determines its resistance. It has been found that a well-annealed glass has from 2.5 to 3 times the resistance of an unannealed sample.

The presence of electrolysis in glass is another factor affecting its resistance. If a high d-c potential difference is applied to two electrodes embedded in glass, a current will flow between them. The value of the current is a function of the geometry, the composition, the condition, and the temperature of the glass. Conduction is ionic (not electronic), and as a consequence, electrolysis with the accompanying decomposition is always present under such conditions. Faraday's law is obeyed and positive metal is deposited at the cathode and oxygen is liberated at the anode. The main portion of the current is carried by the positive sodium ion, because it is more mobile than the large negative silicate ion. The sodium ion migrates to the cathode and is deposited. However, as soon as it is deposited, its charge is neutralized and it becomes a sodium atom. Being very reactive, it practically always reacts with the glass in some way. In the case of the lead glasses used for stems, the sodium atom reduces the lead oxide to lead with the formation of the lead deposit around the lead wires. This condition is familiar to all who have had any connection with studies of stem electrolysis. Eventually such action results in a leak along the lead wire.

Electrolysis varies with the operating temperature of the glass for the reason that the glass resistance changes with temperature. It has been found that electrolysis is the most troublesome in rectifier tubes which have the highest stem operating temperatures. The approximate stem operating temperature in the type 5Z3 is 250°C; in the type 80, 212°C; and in the type 5Z4, 198°C.

Referring again to the resistance curves in Fig. 10, it will be noted that the horizontal line termed the upper limit of electrolysis of No.12 glass passes through the resistance curve of this glass at 250°C. This is the limiting operating temperature for rectifier-tube stems. A type 5Z3 stem of No.12 glass has a resistance of 6000 megohms and an electrolytic current of 0.05 microampere at 300 volts d.c. From the geometry of the stem involving the thickness of stem press, the length of the wire-glass seal, and space between lead wires, the specific resistance can be calculated.

A comparison of resistance for the different glasses shown in Fig. 10 would indicate that No. 816 (814KW) glass has about three times the resistance of No.12 glass at an operating temperature of 250°C. Therefore, the expected electrolysis using No.816 glass would be about 32 per cent of that with No.12 glass under the same conditions.

The advantage of No.816 glass over No.12 glass as far as electrical resistance is concerned, is not so great as would at first appear from the graphs in Fig. 10, because at the safe limit of electrolysis the No.816 glass only has a margin

of about 13°C over No.12 glass. In many instances the tube stems made of No.816 glass have shown as bad electrolysis as those made of No.12 glass due to the fact that the stems were very poorly annealed.

All of the resistance values in the curves in Fig. 10 were measured on well-annealed glass. As stated previously, a poorly annealed glass has from 33 to 40 per cent of the resistance of well-annealed glass. It will be seen that the reduction in electrolysis expected when No.816 glass is substituted for No.12 glass would be nullified by a poor annealing of the No.816 glass. The resistance curves of the different glasses would be considerably lowered in the case of poorly annealed glasses.

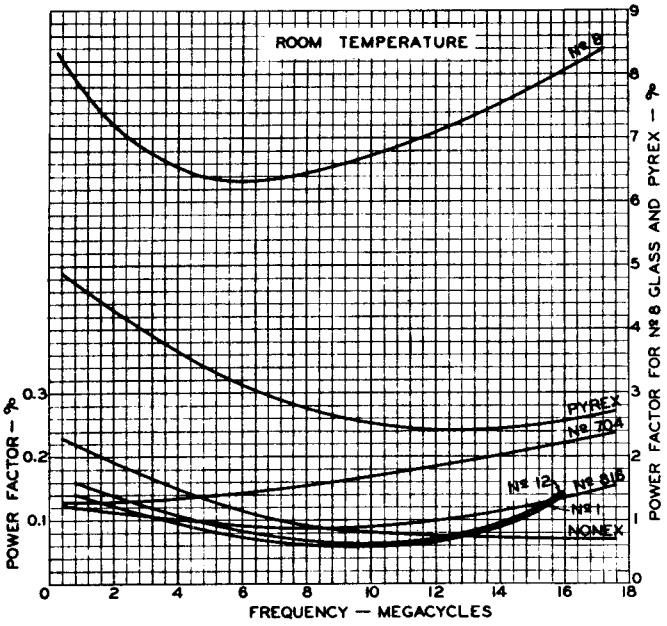


Fig. 11 - Effect of frequency on power factor of glasses at room temperature.

The factors to be considered for minimum electrolysis are: high electrical-resistance glass; glass well annealed; low operating temperature; and maximum lead-wire spacing.

B. Dielectric Properties

Intimately connected with the electrical resistance of a glass are its properties as a dielectric. The dielectric constant will range in the usual soft glasses between 6 and 8. Hopkinson[5] found that the dielectric constant for a series of lead glasses with varying quantities of lead followed the general relation

$$k/d = 2.2$$

The dielectric constant is somewhat dependent on frequency and temperature but not in such a manner as to be serious. Variation with temperature at high frequencies is less marked than at low frequencies. In the case of the dielectric loss, however, there is a different story. A number of measurements made by P. A. Richards and others in this laboratory on the glasses used in radio-tube manufacture are shown in Fig. 11. These curves show the effect of frequency on the power factor of glasses at room temperature. It will be seen that in all cases except Nonex glass the curves go through a minimum when the measurements are carried to sufficiently high frequencies. The effect of raising the temperature is very marked in increasing the power factor in all cases, but unfortunately, there is not sufficient systematic data available to present in curve form the information on our glasses. Measurements on other glasses indicate that a threefold increase in power factor may be expected from a rise in temperature of 100°C.

BIBLIOGRAPHY

E. M. Guyer, "The Electrical Behavior of Glass at Room Temperature," Jour. Amer. Ceramic Soc., Vol. 16, No.12, pp. 607-618; 1933.

Hodkin and Cousen, "Textbook of Glass Technology." (Second Edition) 1929, D. Van Nostrand Co., Inc.

H. R. Kiehl, "The Electrical Conductivity of Glass," Parts I and II, Physics, Vol. 5, No.12, pp. 363-369 and 370-373; 1934. Part III, Vol. 7, No.1, pp. 20-25; 1936.

H. R. Lillie, "Stress Release in Glass - A Phenomenon Involving Viscosity as a Variable with Time," Jour. Amer. Ceramic Soc., Vol. 19, No.2, pp. 45-54; 1936.

H. R. Lillie, "Viscosity of Glass between the Strain Point and Melting Temperature," Jour. Amer. Ceramic Soc., Vol. 14, No.7, pp. 502-511; 1931.

H. R. Lillie, "Viscosity-Time-Temperature Relations in Glass at Annealing Temperature," Jour., Amer. Ceramic Soc., Vol. 16, No.12, pp. 619-631; 1933.

J. T. Littleton, "The Physical Processes Occurring in the Melting and Cooling of Glass," Jour. Amer. Ceramic Soc., Vol.17, No.3, pp.43-49; 1934.

[5] J. Hopkinson, "Electrostatic Capacity of Glass," Phil. Trans. of Roy. Soc. of London, Vol. 168, pp. 17-23 and plate 2; 1878.

J. Hopkinson, "Electrostatic Capacity of Glass," Proc. of Roy. Soc. of London, Vol. 31, pp. 148-149; 1880.

J. Hopkinson, "Dielectric Capacity of Liquids," Proc. of Roy. Soc. of London, Vol. 31, pp. 347-348; 1881.

J. T. Littleton, "The Effect of Heat Treatment on the Physical Properties of Glass," Bulletin of Amer. Ceramic Soc., Vol. 15, No.9, pp. 306-311; 1936.

Littleton and Morey, "The Electrical Properties of Glass." (First Edition) 1933, John Wiley & Sons.

Littleton and Wetmore, "The Electrical Conductivity of Glass in the Annealing Zone as a Function of Time and Temperature," Jour. Amer. Ceramic Soc., Vol. 19, No.9, pp. 243-245; 1936.

George W. Morey, "Constitution of Glass," Jour. Amer. Ceramic Soc., Vol. 17, No.11, pp. 315-328; 1934.

George W. Morey, "Annealing of Pyrex Chemical Resistant Glass," Ind. and Eng. Chem. (Indust. Edition), Vol. 27, No.8, pp. 966-971; 1934.

Morey and Littleton, "Phase Equilibrium Relationships Determining Glass Composition," "Critical Temperature in Silicate Glasses," Ind. and Eng. Chem. (Indust. Ed.), Vol. 25, No.7, pp. 742-755; 1933.

W. E. S. Turner, "The Chemical Durability of Glass," Bulletin of Amer. Ceramic Soc., Vol. 14, No.5, pp. 165-169; 1935.

Lecture 10

CONSTRUCTION TRENDS IN RADIO TUBES

N. R. Smith

The structural design of radio tubes is a very comprehensive subject. To go into the detailed mechanical design of the parts and electrodes entering into the fabrication of vacuum tubes would involve a great many formulae, tabulations and design factors that would require much more space and time than has been apportioned to this subject. Consequently, this lecture will be devoted to a short historical review of the various mechanical constructions used and a description of the various types of electrodes now available. A comparison of the present with the earlier constructions will serve to indicate the improvements made and will provide a more complete understanding of the problems involved in structural design.

EDISON EFFECT

The phenomenon which has played so important a part in the development of the radio tube was first noted by Thomas A. Edison in 1881. With an experimental device similar to that shown in Fig. 1, Edison discovered that when an electrode was interposed between the legs of an incandescent filament and was connected to the positive end of the filament, a current would flow between that electrode and the positive filament terminal. This phenomenon was called the "Edison effect," and a patent was issued covering it in 1884. The parts used in this experimental device were the same as were then being used in the manufacture of carbon-filament incandescent lamps which had been developed by Edison during the previous two years. It is interesting to note, that for all practical considerations, the fundamental design of enclosure suggested in this device fifty-five years ago has been retained on many tubes manufactured today, and until the advent of the dome-top bulb in 1932, was the most prevalent design used by American tube manufacturers.

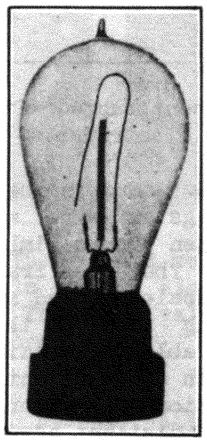


Fig. 1 - Edison's experimental lamp with an electrode interposed between legs of the incandescent filament and connected to the positive end of the filament.

FLEMING VALVE

For the succeeding twenty-four years, following Edison's discovery, no practical application was made of the "Edison effect." The first commercial application was made by Professor J. A. Fleming in 1905 when he developed the two-electrode tube, "Fleming valve," and suggested that it be used for the rectification of alternating currents. One type of Fleming valve is shown in Fig. 2. The similarity of this tube to several modern rectifiers is apparent.

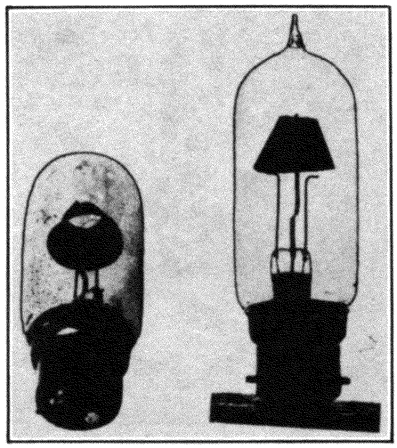


Fig. 2 - One type of Fleming valve.

DeFOREST AUDION

The following year, 1906, Dr. Lee DeForest added a third electrode, a grid, between the cathode and plate. This grid provided a means for controlling the current passing through the tube. In these first tubes, such as is illustrated in Fig. 3, Dr. DeForest made use of two standard lamp flanges sealed into opposite ends of a G-16-1/2 type bulb. The filament was supported from one flange and the leads were brought out to a standard mogul screw base. The grid and plate were supported from the opposite flange. The 25-mil nickel grid lead was formed inside the tube into a zigzag shape with spaces between adjacent zigzags varying between 3/64" and 3/32". The plate consisted of a piece of 5-mil nickel ribbon, 9/16" square, scribed and sheared out by

hand. This was probably the first instance of a double-ended construction and of an attempt to align the electrodes by setting the glass seal.

Although Dr. DeForest demonstrated the practicability of the Audion (the name which he applied to his three-electrode tube) as a detector and amplifier of radio signals in various instal-

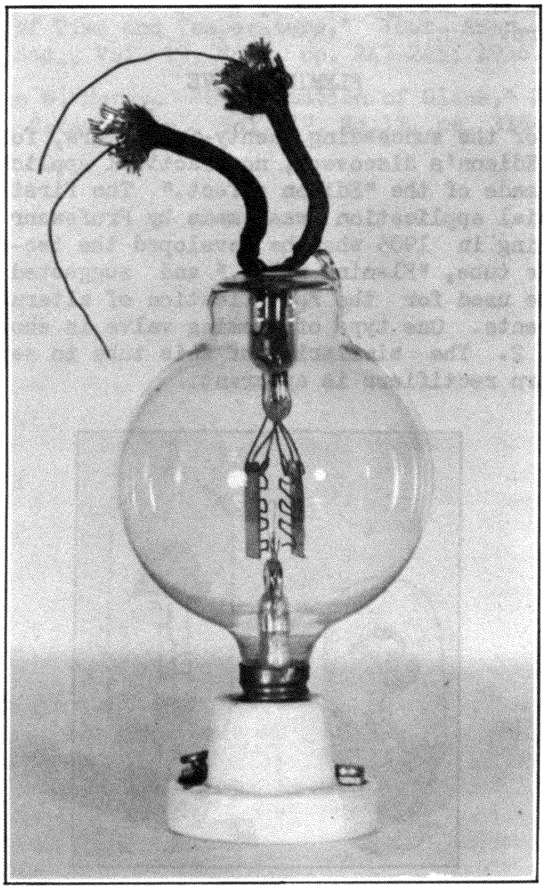


Fig. 3 - Early type of DeForest triode.

lations made during 1907, little interest was manifested in the development. This may have been due, in part, to the fact that both the Fleming valve and DeForest Audion were definitely limited in useful application by the small plate voltages which had to be used to prevent excessive ionization. The output of the early Audions did not exceed six milliwatts. This limitation resulted from a lack of knowledge on how to degas the elements of the tube and obtain a high vacuum. Improvements in exhaust technique were retarded by the theory, prevalent at the time, that conduction of current through the tube depended on gas ionization and that a small gas content, therefore, was essential to successful tube operation.

EXHAUST DEVELOPMENT AND FILAMENTARY-ELECTRODE TUBES

The most important advance in radio-tube design following DeForest's development was contributed by Dr. Irving Langmuir, W. C. White, and Dr. Saul Dushman. During their investigation of exhaust technique extending over a period of four years from 1909 to 1913, many improvements in exhaust technique were made. For example, the degasification of the parts was improved by the use of filamentary electrodes. A tube of this construction is shown in Fig. 4. Two leads were brought out from each of the electrodes to permit degasification by heating with a current which could be passed through each electrode. The grid consisted of tungsten wire wound on a glass frame which also included the filament suspended from

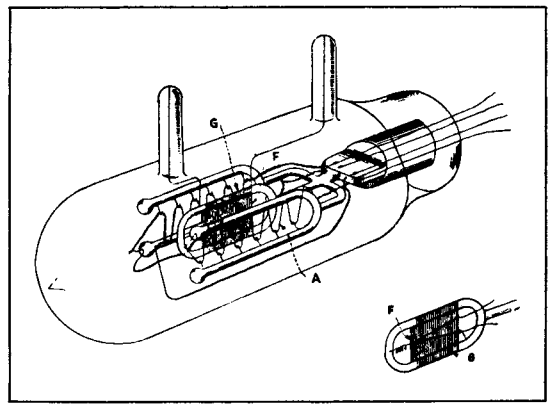


Fig. 4 - Tube constructed with filamentary electrodes.

tungsten hooks imbedded in the glass. The plate consisted of tungsten wire wound in a zigzag fashion on tungsten hooks imbedded in U-shaped glass supports. The glass frames for both grid and plate were sealed to and supported from the glass stem. Using tubes of this construction, Dr. Langmuir was able to obtain outputs as high as one ampere with anode potentials of about 200 volts. Continuation of the work soon led to the discovery that a distinct advantage was gained by prebaking parts in a vacuum prior to assembly of the tube, and by the use of electron bombardment for degasification of the electrodes. Using these artifices, Dr. Langmuir was able, late in 1913, to produce a triode which would modulate as much as two kilowatts at 10000 volts and a two-electrode tube capable of rectifying as much as 100 milliamperes at 100000 volts.

Experiments made during the course of this investigation led to the discovery by Dr. Langmuir, of the 3/2-power formula which has become a fundamental law for the design of vacuum tubes.

DUMET WIRE

The development of "dumet wire" by C. G. Fink

and its application to lamp-manufacturing practice in 1912 was one of the major improvements in constructional procedure. The dumet wire used in making the vacuum-tight seal to glass had much better thermal and electrical characteristics than the more expensive platinum press leads which it replaced. The initial production of radio tubes in 1920 used this material which still remains the standard material for seals to soft glass.

WEAGANT VALVE

Another interesting construction developed in 1913 was the Weagant valve shown in Fig. 5. In this tube the conventional grid consisting of a lattice interposed between cathode and plate was replaced by a grid external to the vacuum chamber. This external grid was in the form of a cylindrical electrode concentrically aligned

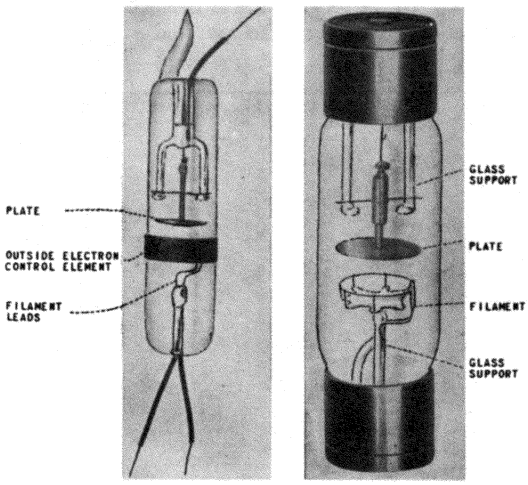


Fig. 5 - (left) The Weagant valve with the outside electrostatic control element in position. (right) The same valve with the control element removed.

with the cathode stream. Although a tube with this construction was somewhat less expensive to manufacture, its low sensitivity and its limitations as an amplifier when compared with the DeForest type made it unsatisfactory for commercial development.

TELEPHONE REPEATER TUBES WITH OXIDE-COATED FILAMENTS

In 1912 the Western Electric Company became interested in the Audion for use as a repeater or booster in telephonic communication. A tube similar to that shown in Fig. 6 was designed and placed in service in 1915 to establish the first transcontinental telephone service. The electrodes were enclosed in a G-18 bulb and the leads brought out to a molded composition four-pin base. The stem had a formed-cane glass support sealed to one end of the press. Various tungsten-wire supports were imbedded in the cane and served to hold the electrodes in correct relationship. The filament was platinum coated with barium oxide. This filament gave much better emission than the pure tungsten filaments then commonly used. Although the oxide-coated filament was developed by the Western Electric Company in 1914, it was not utilized by other manufacturers until the Westinghouse Electric and Manufacturing Company announced the WD-11 in 1920. The grid used in the repeater tube consisted of parallel tungsten wires welded across nickel side-rods. The grids were made in strip and folded to form the two sides. Of especial interest to the engineer concerned with end loss is the filament mounting. The ends of the filament were clamped into the ends of 25-mil copper leads which were sealed

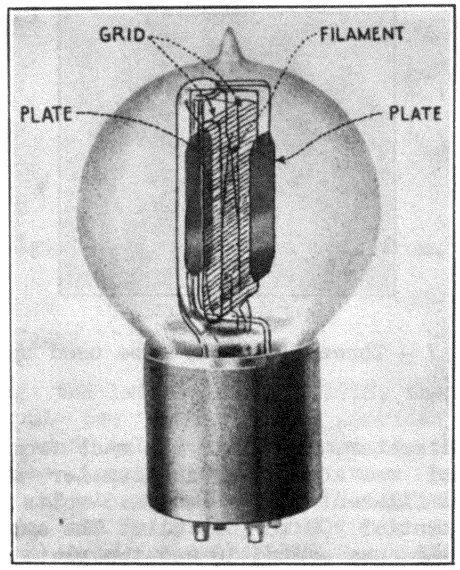


Fig. 6 - Early type of Western Electric triode.

into the glass stem press. However, tubes of this early design were fundamentally satisfactory and with slight modifications are still produced.

TUBES DEVELOPED IN EUROPE FOR WAR SERVICE

The World War served in 1914 to stimulate further development of the radio tube. The value of this device for use in both telegraphic and telephonic communication under front-line conditions was recognized and so the French began development of simple, rugged tubes which could be made in quantity with uniform characteristics. One tube used extensively during the latter years of the War is shown in Fig. 7. The tube had a G-16-1/2 bulb and a molded composition base with long pins. The electrodes were mounted horizontally. This type of construction has been retained abroad especially in Holland and Germany,

and seems to be preferred to the vertical mounting which predominates in American-made tubes. The filament was a straight tungsten wire about 0.83" long mounted along the axis of a spirally wound grid and cylindrical plate. The grid was made of 10-mil nickel wire wound in the form of a helix 0.7" long with 11 or 12 turns having a

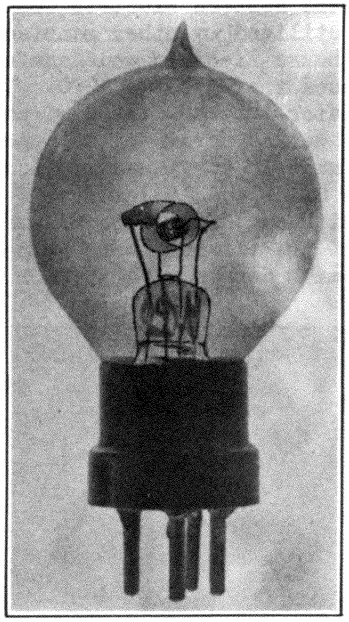


Fig. 7 - Three-electrode tube used by the French Army.

160-mil diameter. The plate was made from nickel ribbon and was about 0.4" in diameter and 0.6" long. The filament voltage was 4.5 volts and the plate potential 300 to 400 volts. The amplification factor was around 10 and the plate resistance was about 24000 ohms. The French used this tube as detector, amplifier, and oscillator; they preferred to sacrifice the advantages of specialized tube designs for the sake of simplifying construction and manufacturing problems.

TUBES DEVELOPED IN U. S. FOR WAR SERVICE

When the United States entered the War in 1917, a demand for large quantity production was made upon the manufacturing companies. Having had no commercial outlet prior to this time and consequently no factory experience, the companies were faced with an extremely difficult problem.

The Western Electric Company had been making tubes for telephone service and were in a better position to supply tubes on short demand than were any of the other manufacturers. Two of the first tubes supplied by Western Electric and known under the Signal Corps designations as VT-1 and VT-2 are shown in Fig. 8. These tubes are interesting inasmuch as they incorporated cold-rolled steel parts which are now being reintroduced to radio tube manufacture. The oxide-coated, platinum-ribbon filament developed by Western Electric was used. The plate consisted of two sheet-steel elements mounted on either side of the filament. They were clamped around the stem press at the lower end and supported by a lava spacer at the top. The grid also consisted of two stamped steel elements welded to lead wires at the bottom and held at the top by wires imbedded in the lava spacer. In the lava spacer was imbedded a wire used to support the apex of the filament. Except for the grid, the whole assembly was very rugged and could stand considerable abuse.

However, these tubes were not suitable for adaptation to the high production equipment of the lamp industry. Therefore, new designs adaptable to lamp manufacturing procedure were necessary. The new designs were conceived by lamp engineers who so coordinated the procedures for manufacturing lamps and tubes that succeeding developments in both industries followed along together for a considerable period.

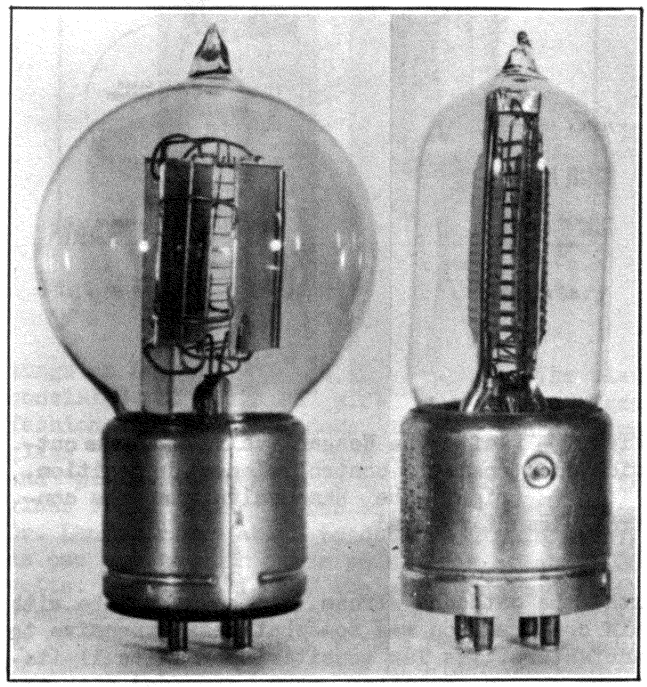


Fig. 8 - (left) Signal Corps tube type VT-1. (right) Signal Corps tube type VT-2.

TUBES FOR AMATEUR USE

At the close of the War and with the restoration of amateur licenses, there was a limited demand for tubes for amateur use. This demand was supplied by various concerns manufacturing tubes under the DeForest patents. Two of the most prominent types were the Myers Audion of Canadian manufacture, and the Audiotron (DeForest

Audion) distributed by the Audiotron Sales Organization. The former was sold in the United States under the name Rac-Audion.

Myers Audion and Rac-Audion

The Myers Audion, shown in Fig. 9, had the electrodes mounted on leads extending through glass beads at either end of the mount assembly. The beads were positioned by two, heavy side-rods

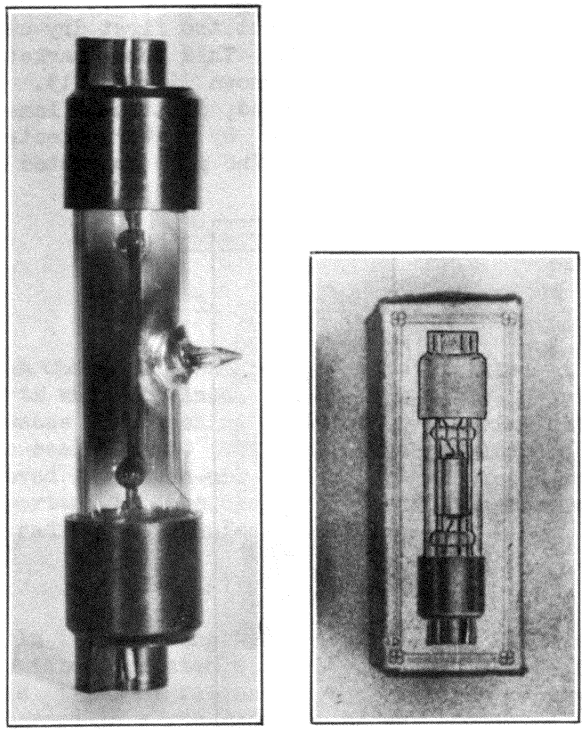


Fig. 9 - Myers Audion.

extending the full length of the tube and sealed into the glass press. The vacuum seal was made by pressing the ends of the tubular enclosure against the leads. This construction was representative of one of the earliest attempts to obtain unit-mount assembly.

Audiotron

The Audiotron, shown in Fig. 10, was probably the predominant tube used by amateurs in the United States from 1918 to 1921. It was a hand-made tube using standard lamp flanges sealed into the ends of a tubular bulb. One flange supported two filaments having separate external connections while the opposite flange supported the grid and plate. The grid was simply a continuation of the inner copper lead formed into a helix. The plate was made of sheet aluminum formed into a cylinder which was allowed to expand against the walls of the bulb to hold it in position. The plate connection was made by threading the inner copper lead through a hole in one end of the plate and bending the lead so that it clamped firmly onto the plate material. Comparison of this construction with the high-precision tubes of today illustrates most clearly the many improvements made in radio tube manufacture during the past eighteen years.

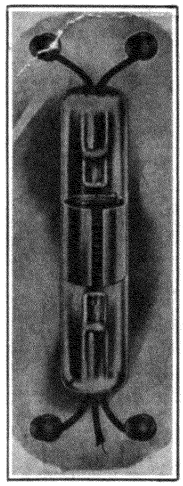


Fig. 10 - The DeForest Audiotron.

Marconi Tubes

During the latter part of 1918, the Marconi Company put two tubes on the American market. One was a soft tube (detector) and the other was a hard tube (amplifier). They were both constructed as shown in Fig. 11. These tubes were similar to the French type "R" tubes developed during the early part of the War. Their sale however, was of short duration due to the formation of the Radio Corporation of America.

RADIO CORPORATION OF AMERICA

The Radio Corporation of America was formed on October 17, 1919. It acquired rights from the Marconi Wireless Telegraph Company of America and from the General Electric Company and subsequently from the American Telephone and Telegraph Company and the Westinghouse Electric and Manufacturing Company, which enabled the new organization to take the first steps in the development of radio for public entertainment.

TUBES FOR BROADCAST RECEPTION

A little more than a year after the formation of the Radio Corporation of America, broadcasting was begun from Westinghouse station KDKA. This occurred in November, 1920. The following month (December, 1920), the first RCA tubes were announced. They were the UV-200 and the UV-201.

UV-200 and UV-201

The UV-200 was a soft or detector tube, while the UV-201 was a hard or amplifier tube. The

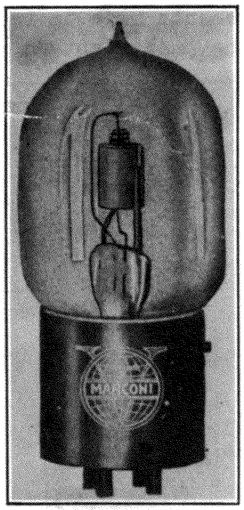


Fig. 11 - Marconi Tube.

UV-201 is shown in Fig. 12. Both tubes were identical in construction and used an ST-14 bulb, Navy metal-shell base, pure tungsten filament, molybdenum grid with nickel side-rods, nickel plate, small tungsten filament hook, and nickel

Fig. 12 - Radiotron UV-201.

inner leads. The only difference between the two types was that the UV-200 contained about 50 microns of argon. The grid wire was shortly changed to nickel, but when this proved too soft for use on the newly developed automatic grid machines, the wire was changed to Nichrome. Nichrome and molybdenum have since been most generally used for this purpose. These tubes were almost identical in construction to the VT-13, an improved design of the VT-11 which was made by the General Electric Company for the Signal Corps in 1918. They employed one of the two general types of construction used for all receiving tubes developed up to 1927. The second type of construction was used in the WD-11.

WD-11

The Westinghouse Company prior to its association with RCA brought out the first dry-cell tube, the WD-11, in 1920. This tube marketed later in 1920 by RCA, is shown in Fig. 13. It used a straight oxide-coated, platinum filament similar to that developed by Western Electric several years previously. The grid consisted of

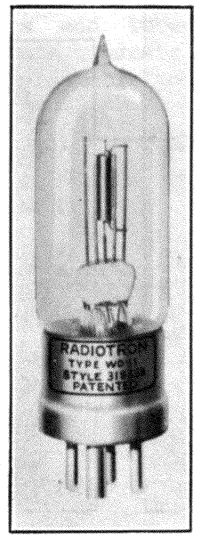


Fig. 13 - Radiotron WD-11.

a helix welded to a single side-rod which was formed above and below the plate to weld to an inner lead extending alongside the plate. The plate was a section of metal tubing welded between side-rods extending from the glass press.

Peanut Tubes

The announcement of the UV-200, UV-201, and WD-11 in 1920 and the rapid development of commercial broadcasting was followed by a wave of amateur set building. A great many tubes and other devices were put on the market for use by home set builders. Among these were the so-called "Peanut Tubes." The Welsh peanut tube, shown in Fig. 14, was one of the most popular of these small designs, and was constructed on the principle of the Weagant valve.

HIGH-FREQUENCY INDUCTION HEATER

An improvement in the processing of tubes was

made possible by the development of the high-frequency induction heater in 1920. By means of the high-frequency current it produced, the inner metal parts of a tube could be heated during exhaust to a higher and more uniform temperature

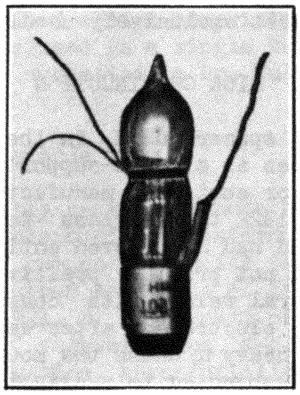


Fig. 14 - Welsh Peanut Tube.

with the result that better degasification of the parts was obtained. Also, this device provided a means for flashing a gettering material within the sealed tube. This procedure resulted in improved gas clean-up. These developments were important to the use of thoriated-tungsten wire in radio tube manufacture.

THORIATED-TUNGSTEN WIRE

In 1915, Irving Langmuir proposed the use of thoriated-tungsten wire for vacuum-tube cathodes. This type of filament, when properly activated and used in a gas-free tube, was found to give superior performance with only half the filament power used for pure tungsten. It appeared probable that suitable tubes could be made commercially with the thoriated wire provided production equipment could be improved to get the high vacuum required. However, during the latter part of 1920, quite by accident, some thoriated-tungsten wire became mixed with the pure tungsten wire used for the UV-201. Tubes made with the thoriated wire proved so satisfactory that immediate development of thoriated-tungsten filament tubes was considered. Consequently, design work on an improved UV-201 and on a new dry-cell type of tube was started early in 1921. This work resulted in the announcement of the UV-201-A and UV-199 in December, 1922. The design of the UV-199, shown in Fig. 15, followed closely that of the WD-11 except that a bead was added to facilitate support of the electrodes.

STRUCTURAL DEVELOPMENTS

Tipless Stem

In March, 1924, a most important structural change went into production. Through the invention of Messrs. L. E. Mitchell and A. J. White, the exhaust tube was removed from the top of the bulb and brought out through the stem tubing. This change was of particular value not only be-

Fig. 15 - Radiotron UV-199.

cause of the greater ease with which tubes could be handled on the exhaust equipment, but also because it eliminated a tubulating operation. Also, it paved the way for the later development of double-ended tubes.

Molded Base

During the fall of 1924, the molded-composition base superseded the brass-shell base for general use. A progression of other base changes started at this time which will be detailed later under the subject of bases.

Channelled Plates

The use of channels or ribs for strengthening and maintaining tolerances on small, sheet-metal parts was initiated in parts manufacturing in 1924.

POWER OUTPUT TUBES

The development of loudspeakers in 1925 created a demand for tubes capable of delivering additional power. The UX-112, UX-120, and UX-171 were developed to meet this need. The UX-112 was a medium-mu tube with oxide-coated filament, while the UX-120 and UX-171 were low-mu tubes with thoriated-tungsten filaments. The UX-120 was patterned after the UX-199 while the UX-112 and UX-171 followed the UX-201-A design. In each case, the filament current was doubled.

FULL-WAVE RECTIFIERS

Early in 1924 the development of B-eliminators created a demand for suitable rectifier tubes. Development was already under way on a full-wave

rectifier tube, to which the designation UV-213 was assigned in December, 1923. This tube used a pure tungsten filament and had a rated d-c output of 10 milliamperes at 90 - 100 volts across the filter input. The need for additional rectified power was increasing rapidly and so the UV-213 was redesigned early in 1924 to use thoriated-tungsten filaments and accordingly rerated at 65 milliamperes at 220 volts. Coincident with this change, the UV base was superseded by the UX base and the tube was designated as the UX-213. It was announced in September, 1925. The Raytheon BH (neon gas rectifier) with a rating of 85 milliamperes had preceded the UX-213 on the market and was being used by many B-eliminator manufacturers. The popularity of the BH rectifier, shown in Fig. 16, was shared with the QRS and Atwater Kent gaseous rectifiers. These three types were used extensively in early B-eliminator applications.

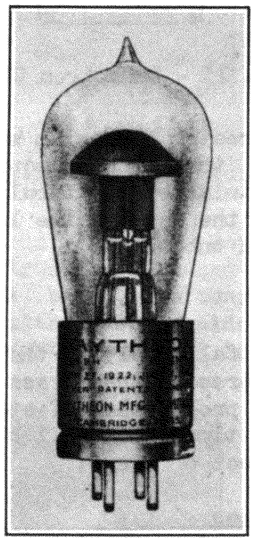


Fig. 16 - Raytheon BH Rectifier.

There was also another filament type of rectifier which deserves attention because of its unique construction. This tube, known as the UV-196, used separate oxide-coated filaments with a common anode to obtain full-wave rectification. It was manufactured exclusively for use in the Super-Ducon B-eliminator marketed by the Dubilier Condenser Corporation.

The next improvement in full-wave rectifier tubes came with the development of a-c tubes. A new rectifier was designed similar to the UX-213 but using oxide-coated filaments and having an output of 125 milliamperes with an applied a-c voltage of 300 volts. This tube was designated as the UX-280 and was announced in May, 1927. It has remained the most popular rectifier on the market to date.

OXIDE-COATED FILAMENTS

During the latter part of 1926, the 171 was revised to the 171-A by changing from thoriated-tungsten filament to oxide-coated filament requiring only half the filament current. During the first few months of 1927, the 112 was revised to the 112-A which likewise required only half its former filament current. Since then, oxide-coated filaments or cathodes for receiving types have been almost exclusively used.

MICA CONSTRUCTION

The lava spacer used in the early Western Electric tubes as a mount support and electrode spacer had, for cost and manufacturing reasons, given way by 1922 to the glass bead. However, the glass bead had not proved entirely satisfactory. It did not provide positive location and required several welds with consequent adjustment of the electrodes after welding. It was sufficiently heavy to cause the mount to shift if the tube was subjected to a slight impact. Late in 1926 the triangular mica spacer was adopted. The use of the mica, besides strengthening the mount, proved very beneficial in reducing microphonics in the UX-201-A.

DELAYED GETTER FLASH

Although several methods of applying getters to the tube mount had been tried since the development of the high-frequency induction heater in 1920, such as painting the parts with solutions of getter material, welding strips of getter directly to the parts, or fastening the getter to the parts by means of straps, the most popular method was to weld a strip of magnesium to a lower corner of the plate. Then, during the heat treatment of the plate on exhaust, the getter was flashed. By this method the temperature to which the plate was heated to degas it depended to some extent on the temperature at which the getter flashed. To improve the exhaust, the method was modified so that the high-frequency field was so directed that practically all the plate was heated to a high degassing temperature without flashing the getter, and then the field was shifted to flash the getter. It was found convenient to place the getter on a tab so located that it did not get hot enough to flash the getter during the degassing of the plate but it could be heated to flash the getter after the plate was degassed.

A-C TUBES

Raw A-C Type

The UX-226 announced in 1927 was designed for raw a-c operation of its filament and was identical in construction to the UX-201-A except that the tungsten filament was replaced by a heavier oxide-coated type designed to carry more current at lower voltage than the 201-A. Although the 226 was designed as a general-purpose type, it

did not perform as satisfactorily as desired when used as a detector. Another type, the indirectly heated cathode type, was then considered for this purpose.

Rectifier-Triode Type

Another construction to receive consideration was a single-ended design comprising a rectifier and triode enclosed in a single bulb. This arrangement was suggested by Dr. Hull of the General Electric Company.

Unipotential Cathode Type

The indirectly heated cathode had been suggested by Nicholson of the Western Electric Company in 1915, but no commercial application to receiving tubes had been made of the suggestion. As early as 1921, H. M. Freeman of the Westinghouse Company suggested an a-c unipotential cathode construction substantially like the indirectly heated cathode now in use. Tubes patterned after this suggestion were put on the market early in 1925 by McCullough and later by the Kellogg Company. An illustration of this type of construction is shown in Fig. 17.

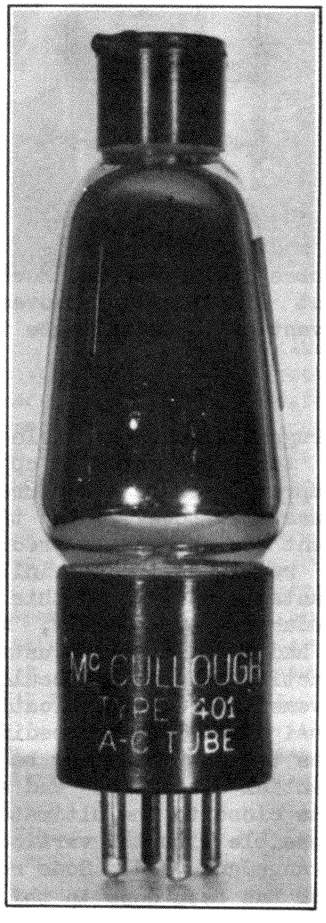


Fig. 17 - McCullough tube type with indirectly heated cathode.

The RCA finally adopted a single-ended construction proposed by the Westinghouse Company. This tube, using a new five-pin base, was designated as the UY-227, and was announced together with the UX-226 and UX-280 in May, 1927. In September, 1927, the first a-c set, the Radiola 17, using UX-226's as amplifiers and the UY-227 as detector, was announced.

THE UX-222 SCREEN-GRID TUBE
FOR DRY-CELL OPERATION

An investigation by Drs. Hull and Williams of the "shot effect" which had been causing difficulty in the Armstrong second-harmonic superheterodyne receivers put on the market in 1924 led to the announcement in 1926 of the screen-grid tube, known as the UX-222. The 222, illustrated in Fig. 18, was the first tube in the receiving

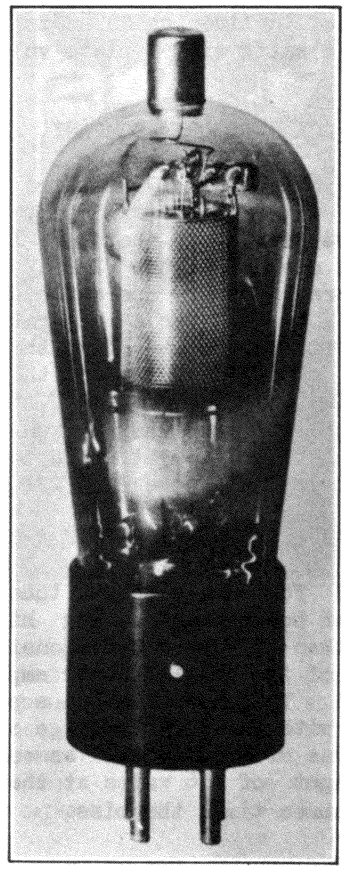


Fig. 18 - Radiotron UX-222.

group which made use of the vacancy left by removal of the exhaust tube from the top of the bulb. The control grid was brought out through a top tubulation and terminated in a small cap on top of the bulb. The screen grid was connected to the usual grid pin in the UX base. The shielding of the input circuit from the output circuit was very good. The 222 was a dry-cell tube operating with the same filament rating as the UX-120, and was announced in October, 1927.

THE UY-224 SCREEN-GRID TUBE FOR A-C OPERATION

Following closely the development of the UX-222, the UY-224 screen-grid tube was announced in 1928. It was of the unipotential-cathode type and had a 2.5-volt heater. The 224 used the five-pin base developed for the 227, and otherwise was similar to the UX-222 in construction. The 224, 226, 227, and 280 provided for the r-f detector, 1st a-f, and rectifier needs of a-c sets, but no power output tube was as yet available.

THE UX-245 POWER OUTPUT TRIODE

The requirement for a power output tube in a-c sets was fulfilled by the development of the UX-245, similar in all details to the UX-171-A, except that it used a heavier filament. The 245 was capable of delivering an undistorted power output of 1.6 watts with a plate voltage of 250 volts.

RCA RADIOTRON COMPANY

At the beginning of 1930, there were only about 15 RCA types of receiving tubes on the market. These included the 199, 120, and 122 for dry-cell operation; the 201-A, 112-A, and 171-A for storage-battery operation; and the 224, 226, 227, 245, and 280 for a-c operation.

However, with the formation of the RCA Radiotron Company early in 1930 and the consequent change in manufacturing procedure, the pace for development of new types was quickened. Shortly, the new organization brought out a two-volt battery line consisting of the 230, 231, and 232.

POWER AMPLIFIER PENTODE FOR A-C SETS

The demand for an a-c output tube capable of giving greater power and having higher power sensitivity than triodes was responsible for the introduction of the RCA-247 power amplifier pentode in 1931. This new tube had a power output of 2.7 watts with a plate voltage of 250 volts and a grid bias of -16.5 volts as compared with the 245's output of 1.6 watts at the same plate voltage and three times the bias.

AUTOMOBILE LINE OF TUBES

Following very shortly after the 247, a complete line of tubes designed especially for automobile service was announced by RCA. The line included the r-f amplifier type 236, the general-purpose type 237, and the power pentode type 238.

DOME-TOP BULBS

During 1932 the entire line of high-production tubes was redesigned for adaptation to the dome-top bulb. The replacement of the S-type bulb with the ST-type bulb (dome-top) constituted a major improvement in mechanical design inasmuch as it provided a means for obtaining greater tube strength with cheaper and less complex structures. In addition, the dome top permitted closer coupling of the internal to the external shielding and suggested the development of such tubes as the 57, 58 (shown in Fig. 19), 6C6, and 6D6.

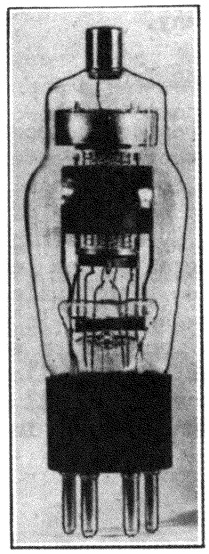
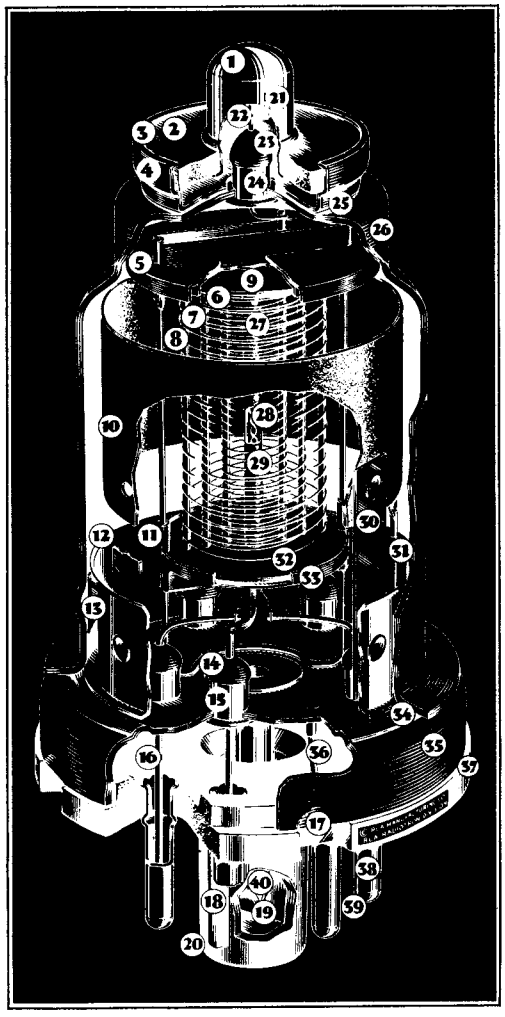


Fig. 19 - RCA-58.

Following 1932, new types were developed so rapidly that a discussion in this lecture of the individual tubes is impossible. However, no further important structural improvement was made until the advent of the metal tube in 1934.

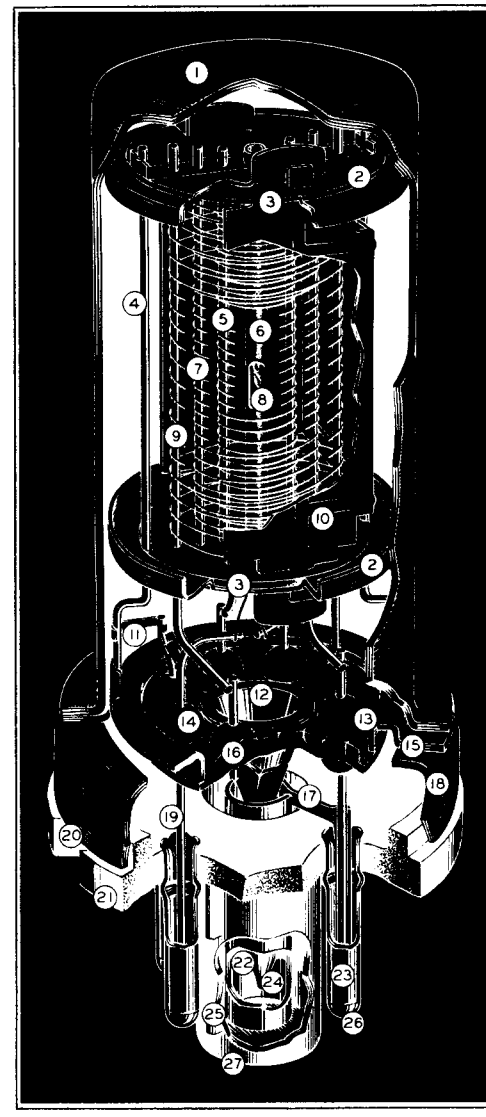
METAL TUBES

The more rugged and uniform enclosure provided by the metal tube presented an opportunity for obtaining improved mount structures at lower cost. Unfortunately, however, the time available for development was such as to preclude the possibility of realizing, in the initial designs, the many advantages offered by this new manufacturing procedure. Consequently, many possible improvements had to await the adjustment of manufacturing practice before they could be utilized. These improvements and simplification of metal-tube mount design have been proceeding gradually. On a few types the mount has now been reduced to only the essential electrodes and supports, and is probably as close to the ultimate simplification as is possible with the vertical, cylindrical-mount arrangement used. Some realization of the work which has been done in this respect can be gained by comparison of the first 6K7 design in metal shown in Fig. 20, with the present 6SK7 mount illustrated in Fig. 21.



1 - SOLDER
2 - CAP INSULATOR
3 - ROLLED LOCK
4 - CAP SUPPORT
5 - GRID LEAD SHIELD
6 - CONTROL GRID
7 - SCREEN
8 - SUPPRESSOR
9 - INSULATING SPACER
10 - PLATE
11 - MOUNT SUPPORT
12 - SUPPORT COLLAR
13 - GETTER TAB
14 - GLASS BEAD SEAL
15 - EYELET
16 - LEAD WIRE
17 - CRIMPED LOCK
18 - ALIGNING KEY
19 - PINCHED SEAL
20 - ALIGNING PLUG
21 - GRID CAP
22 - GRID LEAD WIRE
23 - GLASS BEAD SEAL
24 - EYELET
25 - BRAZED WELD
26 - VACUUM-TIGHT STEEL SHELL
27 - CATHODE
28 - HELICAL HEATER
29 - CATHODE COATING
30 - PLATE INSULATING SUPPORT
31 - PLATE LEAD CONNECTION
32 - INSULATING SPACER
33 - SPACER SHIELD
34 - SHELL-TO-HEADER SEAL WELD
35 - HEADER
36 - SHELL CONNECTION
37 - OCTAL BASE
38 - BASE PIN
39 - SOLDER
40 - EXHAUST TUBE

Fig. 20 - Early design of metal tube type 6K7.



1 - METAL ENVELOPE
2 - SPACER SHIELD
3 - INSULATING SPACER
4 - MOUNT SUPPORT
5 - CONTROL GRID
6 - COATED CATHODE
7 - SCREEN
8 - HEATER
9 - SUPPRESSOR
10 - PLATE
11 - BATALUM GETTER
12 - CONICAL STEM SHIELD
13 - HEADER
14 - GLASS SEAL
15 - HEADER INSERT
16 - GLASS-BUTTON STEM SEAL
17 - CYLINDRICAL BASE SHIELD
18 - HEADER SKIRT
19 - LEAD WIRE
20 - CRIMPED LOCK
21 - OCTAL BASE
22 - EXHAUST TUBE
23 - BASE PIN
24 - EXHAUST TIP
25 - ALIGNING KEY
26 - SOLDER
27 - ALIGNING PLUG

Fig. 21 - Simplified design of present-day metal tube type 6SK7 which supersedes type 6K7 in new equipment. (not to scale)

SUMMARY

As we briefly summarize the results of our inspection of the development of the radio tube, we are conscious of a gradual refinement in manufacturing technique with attendant improvements in mount design; yet, with the exception of metal tubes, there have been few fundamental changes in construction. The bulb, base, and type of stem have remained very similar to the first designs and the vertical-symmetrical mount has remained standard in the American industry.

Some of the more important developments contributing directly to improved construction can be listed as follows:

1) Development of dumet wire in 1912, primarily for lamps to provide a cheaper lead.

2) Adoption of the tipless stem in 1924 to improve manufacturing conditions and later to make the production of double-ended tubes practicable.

3) The use of a separate tab for the getter provided a convenient means for obtaining a delayed getter flash with improved evacuation of the tube.

4) Utilization in 1926 of mica for electrode spacers provided a lighter, cheaper, and more accurate means for locating the tube electrodes.

5) The development of the dome-top bulb in 1932 permitted a reduction in mount costs and at the same time provided stronger mounts.

6) The metal tube developed in 1934 provided a self-shielding structure of machine accuracy and permitted simplification and improvement in design.

In addition to the above developments, there has been a gradual but steady improvement in the design of parts with resultant economy in factory operations. Typical examples are: the use of lug-ear plates to eliminate unnecessary side-rods and assembly operations, and more efficient use of forming to obtain better mount support. Also, parts have been standardized to allow them to be used interchangeably on tubes having the same characteristics but not the same type of enclosures. The need for this standardization of mechanical design is apparent when one considers

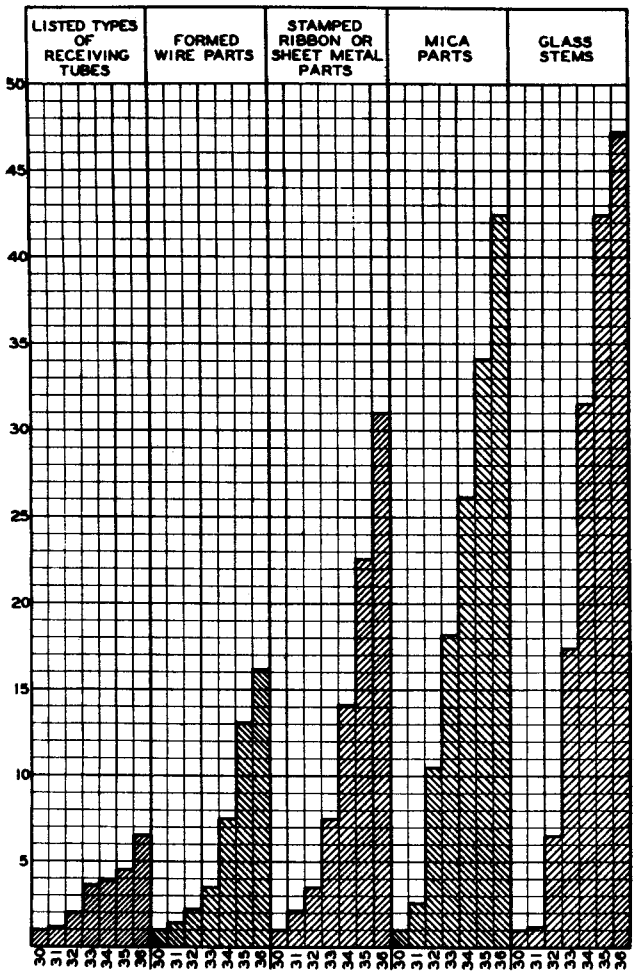


Fig. 22 - Chart showing comparative increases of tube types and parts, based on 1930.

the chart shown in Fig. 22, in which the number of receiving tubes announced have been tabulated for each year since 1920. As the number of types increase, the importance of standardization and mechanical design becomes of more concern and will, due to economic necessity, receive more and more concentrated attention.

Lectures 11, 12, 13, and 14

SPACE-CURRENT FLOW IN VACUUM-TUBE STRUCTURES

B. J. Thompson

A. DIODE THEORY

Ideal Case

The simplest vacuum tube is the diode. The behavior of multi-electrode tubes may be described most readily in terms of the behavior of a diode. For these reasons our treatment will start with the diode.

In the ideal diode, electrons are emitted from the cathode in unlimited numbers at zero velocity and a part of these are drawn over to the anode under the influence of the positive field established by its potential.

In Fig. 1, K represents the infinite plane cathode at zero potential and A the plane anode at a positive potential E_b spaced a distance d_{kp} from the cathode. Let us suppose first that no electrons are emitted from the cathode. The potential distribution will then be as represented by the line (a), the gradient at all points being E_b/d_{kp}. If now the cathode begins to emit a limited supply of electrons, all of these electrons will be drawn to the anode. The electrons move at a finite velocity and, therefore, there is a certain number of them in the space at all times. The field set up by the negative "space charge" of these electrons acts to depress the potential in the space below that of the first condition, increasing the field near the anode and decreasing it near the cathode. This condition is shown by line (b).

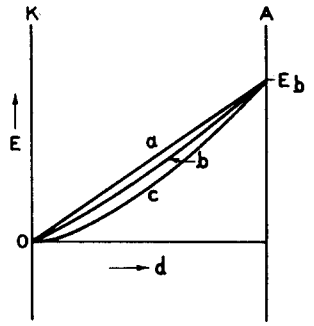


Fig. 1

If the rate of emission of electrons is continually increased, all of the emitted electrons will be drawn to the anode and the gradient at the cathode continually reduced until the gradient reaches zero. Since the electrons are assumed to be emitted with zero velocity, they can not move against a retarding field; therefore, there will be no increase in anode current with further increase in emission beyond this point. The condition of zero gradient at the cathode is represented by the line (c) in Fig. 1.

The mathematical analysis of the ideal parallel-plane case is quite simple. It will be presented here as an example of this type of analysis.

Poisson's equation in rectangular co-ordinates is

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} = -4\pi\rho \qquad (1)$$

Since there is no gradient in directions parallel to the cathode and anode, the equation becomes simply

$$\frac{\partial^2 E}{\partial x^2} = -4\pi\rho \qquad (2)$$

We may also write that

$$I = \rho v \qquad (3)$$

and

$$v = \left(\frac{2eE}{m}\right)^{\frac{1}{2}} \qquad (4)$$

where ρ is the space-charge density, E the potential at any point a distance x from the cathode, v the velocity of the electrons at x, I the current per unit area, and e and m the charge and mass of the electron.

On combining the last three equations, we obtain

$$\frac{d^2 E}{dx^2} = -4\pi \frac{I}{\left(\frac{2eE}{m}\right)^{\frac{1}{2}}} \qquad (5)$$

If we multiply both sides by dE/dx and integrate once, we obtain

$$\frac{1}{2}\left(\frac{dE}{dx}\right)^2 = \frac{8\pi I}{\left(\frac{2e}{m}\right)^{\frac{1}{2}}} E^{\frac{1}{2}} + F_o^2 \qquad (6)$$

where F_o is the field at the cathode. If we let F_o equal zero, a second integration gives us

$$E^{\frac{3}{4}}\Big|_0^{E_b} = 3(\pi I)^{\frac{1}{2}}\left(\frac{m}{2e}\right)^{\frac{1}{4}} x\Big|_0^{d_{kp}} \qquad (7)$$

or

$$I = \frac{1}{9\pi}\left(\frac{2e}{m}\right)^{\frac{1}{2}} \frac{E_b^{\frac{3}{2}}}{d_{kp}^2} = 2.334 \times 10^{-6} \frac{E_b^{\frac{3}{2}}}{d_{kp}^2} \qquad (8)$$

This is the well-known Langmuir-Child[1] equation for space-charge-limited current flow per unit area between parallel-plane electrodes. It means that for each square centimeter of cathode or anode area 2.334 microamperes of current will flow with one volt difference in potential and a distance of one centimeter between cathode and anode, and that a current of 233 microamperes per square centimeter will flow if the potential be raised to a little over 30 volts or the distance reduced to one millimeter.

The foregoing analysis is for parallel-plane electrodes. The case of concentric cylinders, of much practical interest, is very much less simple to analyze and, therefore, only the result will be presented here. Excellent analyses are available in the literature.[2]

The current in amperes per centimeter length of the concentric cylinders is given by the well-known Langmuir equation

$$I = 14.66 \times 10^{-6} \frac{E_b^{\frac{3}{2}}}{r_b \beta_b^2} \qquad (9)$$

where r_b is the radius of the anode and β_b^2 is a function depending on the ratio of anode radius to cathode radius. Tables and curves of β have been published.[3] It will be noted that the current again depends on the 3/2 power of the anode voltage; otherwise, the expressions at first glance do not appear very similar. Part of this difference is due to the fact that one expression is for current per unit area, while the other expression is for current per unit length.

It will be interesting to put the two expressions in similar form. Let us divide equation (9) by $2\pi r_b$. Equation (9) then becomes identical with equation (8) except for the presence of the term β_b^2 in the denominator and the fact that the distance r_b is measured from the axis of the cylindrical system. When the ratio of anode diameter to cathode diameter becomes very large, β_b^2 approaches unity and, of course, the distance between cathode and anode approaches r_b as a limit. At this limit, then equations (8) and (9) become identical, and we observe the interesting fact that the anode current flow per unit area is the same in a cylindrical system with fine-wire filaments as it would be in a parallel-plane system with the same distance between cathode and anode. This statement, of course, neglects the effect of initial velocity of emission.

At the other limit where the cathode and anode diameters approach each other the system is obviously essentially a parallel-plane one. The value of β_b^2 then changes rapidly and maintains such a value that $r_b^2 \beta_b^2$ is equal to d_{kp}^2.

The fact that the two expressions give identical results at the two limits of ratio of anode-to-cathode diameter should not lead one to suppose that the expressions are approximately identical for intermediate ratios. Where the anode diameter is from 4 to 20 times the cathode diameter, the current calculated from expression (8) is in excess of that indicated by expression (9) by very nearly 20 per cent. This is the maximum error that would result from the use of expression (8) for cylindrical structures.

The potential distribution between cathode and anode may be calculated most usefully from the expressions for current. From equation (8) we may write

$$\frac{E_b^{\frac{3}{2}}}{d_{kp}^2} = \frac{E^{\frac{3}{2}}}{x^2}$$

or

$$E = E_b \left(\frac{x}{d_{kp}}\right)^{\frac{4}{3}}$$

In other words, the potential between parallel planes varies as the four-thirds power of the distance from the cathode in the case of space-charge-limited currents.

The potential distribution between concentric cylinders is less simple. We may write from equation (9)

$$\frac{E_b^{\frac{3}{2}}}{r_b \beta_b^2} = \frac{E^{\frac{3}{2}}}{r \beta^2}$$

or

$$E = E_b \left(\frac{r\beta^2}{r_b \beta_b^2}\right)^{\frac{2}{3}}$$

where β^2 is taken for the ratio r/r_k. This expression is not analytical, the values of β and β_b being obtained from curves or tables.

Effects of Velocities of Emission

Electrons are emitted from a heated surface with a random distribution of velocities in all directions. The velocities which concern us in the present analysis are those normal to the surface of the cathode. This velocity distribution may be expressed most simply as follows:

$$\frac{n}{n_o} = \varepsilon^{-\frac{Ee}{kT}}$$

where n is the number of electrons out of the

---

[1] Langmuir and Compton, "Electrical Discharges in Gases," Part II, Review of Modern Physics, Vol. 3, No. 2, pp. 238-239; April, 1931.

[2] See Ref. 1, pp. 245-249.

[3] See Ref. 1, pp. 247-248.

total number n_0 which have a sufficient velocity to reach a plane electrode parallel to the cathode at a negative potential of E, T is the temperature of the cathode, and $\underline{k}$ is Boltzmann's constant. Expressed in terms of current this becomes

$$I = I_S \varepsilon^{-\frac{Ee}{kT}}$$

where I is the current reaching the negative electrode and I_S is the total emission current from the cathode. To carry this out experimentally, it is necessary that the collector electrode be placed so close to the cathode that space-charge effects do not cause a potential minimum in space.

We initially assumed that all electrons were emitted with zero velocity and that, therefore, the field at the cathode would not be negative. In the practical case where all electrons have finite velocities normal to the cathode, all of the emitted electrons must reach a positive anode parallel to the cathode unless at some point between cathode and anode a negative potential exists.

Fig. 2 shows the potential distribution between parallel-plane cathode and anode for successively higher values of emission. Line (a) represents the case where there is no emission

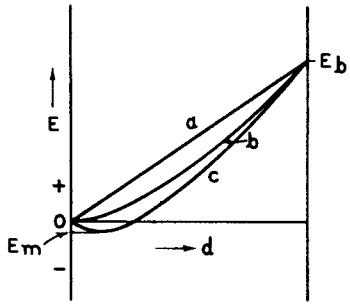


Fig. 2

— and, therefore, no space charge — with resulting constant potential gradient between cathode and anode. Line (b) shows the case where there is sufficient emission to reduce the gradient at the cathode just to zero. This is similar to the condition represented by (c) in Fig. 1 with the important difference that now all electrons pass over to the anode because of their finite velocities of emission.

Any further increase in cathode emission, however, will cause the potential near the cathode to become slightly negative as shown in line (c). In this case all electrons having velocities less than E_m are turned back to the cathode, while those electrons having greater velocities of emission pass on to the anode. Further increases in cathode emission cause the potential minimum to become more negative with the result that a larger fraction of the emitted electrons return to the cathode. For continued increase in cathode emission, however, there will always be some slight increase in anode current.

The results obtained from the simple analysis based on zero velocity of emission are obviously not applicable to this practical case if precision is desired. Since a greater maximum potential difference ($E_b + E_m$) is acting over a shorter effective distance ($d_{kp} - d_{km}$) and since the average velocity of electrons is higher because of their initial velocities and hence the space-charge effect of the electrons is less, it is obvious that the space-charge-limited current flow for a given anode potential is greater in the actual case than in the ideal.

Langmuir[4] has presented a complete analysis of the space-charge-limited current flow with initial velocities of emission. He has shown that a good approximation may be made by the use of equation (8) with a correction for the reduced effective distance and the increased effective potential. His equation is as follows:

$$I_b = 2.334 \times 10^{-6} \frac{(E_b - E_m)^{\frac{3}{2}}}{(d_{kp} - d_{km})^2} \times \left[1 + \frac{0.0247\, T^{\frac{1}{2}}}{(E_b - E_m)^{\frac{1}{2}}}\right] \quad (10)$$

where T is the cathode temperature in degrees Kelvin. I_b is in amperes per unit area. E_m is negative in sense. The value of d_{km} in centimeters may be calculated from the approximate expression

$$d_{km} \approx 0.0156 \left(\frac{1000}{I_b}\right)^{\frac{1}{2}} \left(\frac{T}{1000}\right)^{\frac{3}{4}}$$

The value of E_m is given by

$$-E_m = -\left(\frac{T}{5040}\right) \log_{10}\left(\frac{I_s}{I_b}\right)$$

More complete results of Langmuir's analysis are too cumbersome to be presented here. The use of equation (10) should lead to errors not greatly in excess of 2 per cent even under extreme conditions.

It is interesting to observe from Langmuir's calculation that in a practical case where the cathode temperature is 1000°K, the emission density greatly in excess of the anode current, and the anode current density 1 milliampere per square centimeter, that the distance from cathode to virtual cathode is approximately 0.016 centimeter

[4] See Ref. 1, pp. 239-244.

(0.006 inch). Thus, in modern close-spaced vacuum tubes the position of the virtual cathode can not be neglected.

The error involved in using equation (9) as compared with the exact solution for cylindrical structures is less than in the corresponding case of parallel planes. For a discussion of the effect of initial velocities in this case, the reader is referred to Langmuir and Compton.[5]

The potential distribution between parallel planes, taking into account initial velocities, may best be determined by the use of a plot presented by Langmuir and Compton.[6]

B. TRIODE THEORY

Triode Mu Formulas

The earliest analysis of the electric field existing between parallel planes with a parallel-wire screen interposed is that of Maxwell.[7] In this it is assumed that the spacings between the planes and the screen are large compared with the spacings between wires and that these in turn are large compared with the wire diameter. The result expressed in vacuum-tube terminology is

$$\mu = - \frac{2 \pi d_{gp}}{a \log_e \left(2 \sin \frac{\pi r}{a} \right)}$$

or

$$\mu = \frac{2 \pi d_{gp}}{a \log_e \frac{a}{2 \pi r}} \quad \left(\text{where } \frac{\pi r}{a} \text{ is small} \right)$$

In these expressions, d_{gp} is the distance from the center of the grid wires to the plate, $\underline{a}$ the spacing between grid wires ($a = 1/n$, where $\underline{n}$ is the number of wires per unit length), and $\underline{r}$ is the radius of the grid wires. It will be noted that the distance between grid and cathode does not appear.

This formula is in serious error when the spacing between grid wires is not large compared with the wire diameter, as is frequently the case. Because of this, Van der Bijl developed empirically the formula $\mu = C\, d_{gp}\, r\, n^2 + 1$, where C is equal to 160 for parallel planes. An obvious defect of this expression is that μ can never be less than unity.

The most generally useful and accurate formula for amplification constant which has been published is that developed by Vogdes and Elder.[8] This analysis assumes that the spacing between grid wires is small compared with the distances between the grid and the other electrodes. The development is as follows.

Fig. 3 represents the geometry of the vacuum tube. By means of a conformal transformation, this same geometry may be represented in different co-ordinates. In such a transformation,

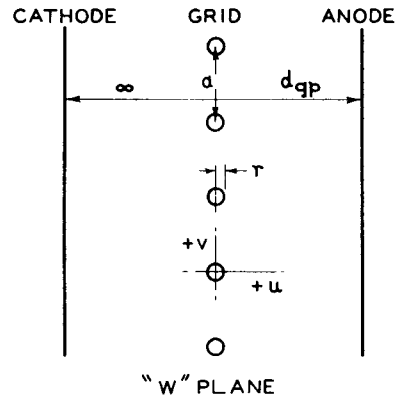


Fig. 3

equipotential surfaces and flux lines still cross at right angles and all laws of electricity still apply.

Suppose the geometry represented in the "w" plane in Fig. 3 be transformed to the "z" plane by the transformation

$$z = e^{2\pi n w}$$

Since $z = x + jy$

and $w = u + jv$

then $x + jy = e^{2\pi n u} \times e^{j 2\pi n v}$

$$= \rho\, e^{j\theta}$$

This transformation is represented in Fig. 4. The cathode is a point at the origin. The grid wires become a single figure intersecting the x-axis at $e^{-2\pi n r}$ and $e^{2\pi n r}$. The center of the grid wires is at $x = 1$. The anode is a circle about the origin of radius equal to

$$e^{2\pi n d_{gp}}.$$

The figure representing the grid wires is not a circle. If $\underline{r}$ is less than $a/2\pi$, however, it can be shown readily that the figure is essentially circular and it will be assumed, there-

---

[5] See Ref. 1, pp. 252-255.

[6] See Ref. 1, Fig. 42, p. 243.

[7] J. C. Maxwell, "Electricity and Magnetism," 3rd edition, 1904, Vol. 1, Section 203.

[8] Vogdes and Elder, "Formulas for the Amplification Constant for Three-Element Tubes," Physical Review, Vol. 24, p. 683; December, 1924.

fore, that such is the case. If the figure is a circle, its radius is

$$\frac{e^{2\pi nr} - e^{-2\pi nr}}{2} = \sinh 2\pi nr$$

and its center is located at

$$x = \frac{e^{2\pi nr} + e^{-2\pi nr}}{2} = \cosh 2\pi nr$$

In Fig. 3, if the anode were removed to infinity and a potential applied to the grid, the successive equipotential surfaces at greater

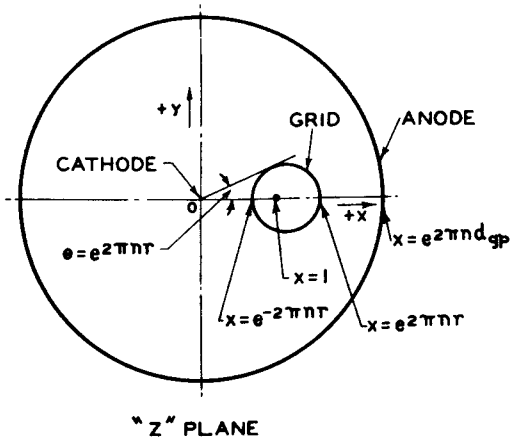


Fig. 4

distances from the grid would become more and more nearly planes until, at distances several times a, the surface could be regarded as essentially a plane. Therefore, under the limitations of our assumptions concerning relative spacings, the anode plane may be considered to be the equipotential surface due to the field of the grid alone. This is equivalent to saying that a circle of radius

$$e^{2\pi nd_{gp}} - \cosh 2\pi nr$$

drawn about the "center" of the grid wire in Fig. 4 does not differ materially from a circle of radius

$$e^{2\pi nd_{gp}}$$

drawn about the origin. The justification for this assumption maybe checked by considering the rather extreme case where $nd_{gp} = 0.50$ and $nr = 0.03$. Then

$$e^{2\pi nd_{gp}}$$

equals 23.1 and cosh $2\pi nr$ equals 1.02.

The convenient result of these assumptions is that a line charge placed at the "center" of the circular grid wire, Fig. 4, produces equipotential surfaces at the surface of the grid wires and at the anode, since the charge on the cathode located at minus infinity must be zero.

Let us place a charge $-Q$ at the "center" of the grid wire. The potentials E_k, E_g, and E_a of the cathode, grid, and anode become

$$E_k = C + 2Q \log \cosh 2\pi nr$$

$$E_g = C + 2Q \log \sinh 2\pi nr$$

$$E_a = C + 2Q\, 2\pi nd_{gp}$$

If the cathode potential be taken as zero,

$$E_g = 2Q \log \sinh 2\pi nr - 2Q \log \cosh 2\pi nr$$

$$= 2Q \log \tanh 2\pi nr$$

and

$$E_a = 2Q\, 2\pi nd_{gp} - 2Q \log \cosh 2\pi nr$$

Under these circumstances, the amplification constant may be defined as

$$\mu = -\frac{E_a}{E_g}$$

whence

$$\mu = \frac{\log \cosh 2\pi nr - 2\pi nd_{gp}}{\log \tanh 2\pi nr} \quad *$$

The assumptions made in this derivation invalidate the expression for use with relatively very close spacings between electrodes. The same type of analysis as that presented by Vogdes and Elder may be made to give more rigorous results. Mr. Bernard Salzberg of the Radiotron Laboratories has carried out such an analysis. It differs from that just presented chiefly in that an additional line charge is placed on the x-axis, Fig. 4, outside the anode at such a position as to make the true anode cylinder an equipotential surface. Therefore, the anode may be allowed to approach much more closely to the grid. This leads to an expression accurate for cases where the spacing between anode and grid is small compared with the wire spacing, though not when the wire spacing is small compared with the wire diameter. Mr. Salzberg's expression is

$$\mu = \frac{\log \cosh 2\pi nr - \dfrac{2\pi nd_{gp}}{\log(1 - e^{-4\pi nd_{gp}} \times \cosh^2 2\pi nr)}}{\log \tanh 2\pi nr -}$$

* Ref. 11-12-13-14, asterisk.

There is no obviously useful definition of amplification factor in the purely electrostatic case (no space charge) when the charge density induced on the cathode is not uniform. It is possible by extension of the analysis described above, however, to arrive at an expression for the charge distribution on the cathode when the spacing between cathode and grid is finite. Mr. Salzberg has carried out such an analysis. It departs from that of the cathode at infinity by considering the potentials in space produced by a line charge at the cathode in addition to the others.

The amplification-factor formulas here given may be applied to cylindrical tubes if $r_g \log(r_a/r_g)$ is substituted for d_{gp}, where r_g and r_a are the radii of the grid and anode, provided r/r_g is small.

Equivalent Potentials in Triodes

For most practical purposes in calculating the electric fields at cathode, anode, and the space between, except very near the grid, a potential may be assigned to the plane of the grid. In other words, it is assumed that an equipotential plane may be substituted for the grid without altering the electric fields. This would be true only when the grid wires are small and closely spaced in comparison with the spacings between grid and cathode and anode.

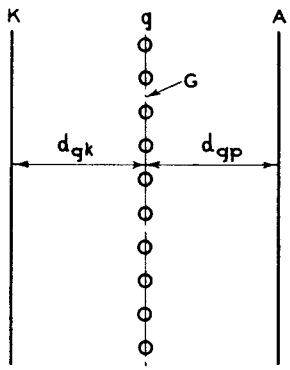


Fig. 5

The equivalent potential of the plane of the grid, E_G, may be derived in several ways. The most simple with which I am familiar is the following. The capacitance between anode and the equivalent plane G at the grid, Fig. 5, is

$$C_{pG} = \frac{1}{4\pi\, d_{gp}}$$

and the capacitance from cathode to G,

$$C_{kG} = \frac{1}{4\pi\, d_{kg}}$$

while, by definition,

$$C_{gG} = \mu\, C_{pG}$$

In the star network of capacitances, Fig. 6,

$$E_G = \frac{E_c C_{gG} + E_b C_{pG} + E_k C_{kG}}{C_{gG} + C_{kG} + C_{pG}}$$

Let us make E_k equal to zero. Then,

$$E_G = \frac{\mu\, E_c + E_b}{\mu + 1 + \dfrac{d_{gp}}{d_{gk}}}$$

or

$$E_G = \frac{E_c + \dfrac{E_b}{\mu}}{1 + \dfrac{1}{\mu} + \dfrac{d_{gp}}{d_{gk}\mu}}$$

The physical basis for this analysis is that the anode can influence the field at the cathode only by acting through the grid plane. By definition, the grid has μ times the influence of the anode. It is obvious that this reasoning implicitly assumes that amplification factor is proportional to grid-anode spacing, for we might just as well have called the cathode the anode. The quantity $d_{gp}/d_{gk}\mu$ is simply the reciprocal of the amplification factor of the grid with respect to the cathode.

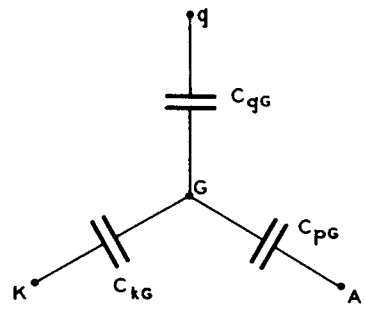


Fig. 6

We shall find it convenient to determine another equivalent-potential plane. The equivalent potential of the grid plane depends on grid and anode potentials and on grid-cathode and grid-anode spacings. Is there an equivalent plane the potential of which depends only on grid and anode potentials and grid-anode spacing?

In Fig. 7, E_G is the equivalent potential of the grid. If the constant potential gradient between grid and cathode extended past the grid,

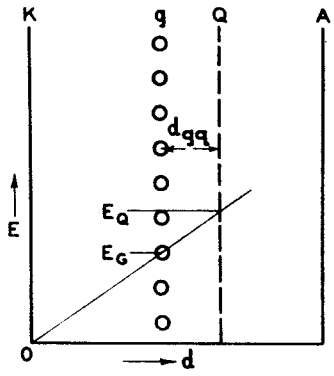


Fig. 7

the potential E at any point a distance x from the grid would be

$$E = E_G \left(1 + \frac{x}{d_{gk}}\right)$$

$$= \frac{\left(E_c + \frac{E_b}{\mu}\right)}{1 + \frac{1}{\mu} + \frac{d_{gp}}{d_{gk}\mu}} \cdot \left(1 + \frac{x}{d_{gk}}\right)$$

We wish to find a potential E_q at a distance d_{gq} from the grid which is independent of d_{gk}. At such a point the ratio

$$\frac{1 + \frac{x}{d_{gk}}}{1 + \frac{1}{\mu} + \frac{d_{gp}}{d_{gk}\mu}}$$

must be independent of d_{gk}. Obviously this means that

$$\frac{x}{d_{gk}} = \frac{\frac{d_{gp}}{d_{gk}\mu}}{1 + \frac{1}{\mu}}$$

or

$$x = \frac{d_{gp}}{\mu + 1} = d_{gq}$$

The potential E_q is given by

$$E_q = \frac{E_c + \frac{E_b}{\mu}}{1 + \frac{1}{\mu}}$$

Applications of this equivalent-potential plane will be given.

Interelectrode Capacitances in Triodes without Space Charge

The direct capacitance between grid and anode, Cgp, may be calculated readily from the expression for E_G, the equivalent potential of the plane of the grid.

The capacitance per unit area from anode to the equivalent plane of the grid is

$$C_{pG} = \frac{1}{4\pi\, d_{gp}}$$

Then

$$C_{gp} = C_{pG} \frac{dE_G}{dE_g}$$

$$= \frac{1}{4\pi\, d_{gp}} \left(\frac{1}{1 + \frac{1}{\mu} + \frac{d_{gp}}{d_{gk}\mu}}\right)$$

Similarly

$$C_{gk} = \frac{1}{4\pi\, d_{gk}} \left(\frac{1}{1 + \frac{1}{\mu} + \frac{d_{gp}}{d_{gk}\mu}}\right)$$

By definition

$$C_{pk} = \frac{C_{gk}}{\mu} = \frac{1}{4\pi\, d_{gk}} \left(\frac{1}{\mu + 1 + \frac{d_{gp}}{d_{gk}}}\right)$$

These derivations are for the parallel-plane case. The case of cylindrical electrodes may be treated in a similar fashion.

Amplification Factor in Multi-Grid Tubes

The analysis of multi-grid tubes may be readily carried out by use of the second expression for equivalent potential, E_q.

In Fig. 8, the Q plane is to be substituted for g_2 and A. Its potential is

$$E_q = \frac{E_{c_2} + \frac{E_b}{\mu_{g_2 p}}}{1 + \frac{1}{\mu_{g_2 p}}}$$

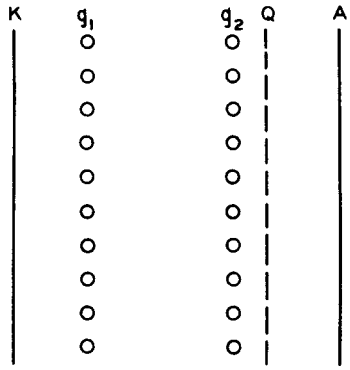


Fig. 8

and its distance from g_1 is

$$d_{g_1q} = d_{g_1g_2} + \frac{d_{g_2p}}{1 + \mu_{g_2p}}$$

We now have a triode and can calculate its μ. The simplest expression is

$$\mu_{g_1q} = \frac{\mu'_{g_1g_2}}{d_{g_1g_2}} d_{g_1q}$$

where $\mu'_{g_1g_2}$ is the amplification factor of g_1 with respect to a plane at g_2.

Now

$$\mu_{g_1g_2} = \mu_{g_2q} \frac{dE_{c_2}}{dE_q}$$

$$= \mu_{g_1q} \left(1 + \frac{1}{\mu_{g_2p}}\right)$$

$$= \mu'_{g_1g_2} \frac{d_{g_1q}}{d_{g_1g_2}} \left(1 + \frac{1}{\mu_{g_2p}}\right)$$

On substituting the expression for d_{g_1q} in this equation, one may reduce it to the following form by simple manipulation:

$$\mu_{g_1g_2} = \mu'_{g_1g_2} + \frac{\mu'_{g_1p}}{\mu_{g_2p}}$$

where μ'_{g_1p} is the value μ_{g_1p} would have if g_2 were removed.

Of course

$$\mu_{g_1p} = \mu_{g_1g_2}\mu_{g_2p}$$

whence

$$\mu_{g_1p} = \mu'_{g_1g_2}\mu_{g_2p} + \mu'_{g_1p}$$

The direct capacitance between g_1 and g_2 may also be determined readily, since

$$C_{g_1g_2} = C_{g_1q} \frac{dE_c}{dE_{c_2}}$$

Also, the capacitances between grids and anode or cathode may be determined in the same manner.

By an obvious extension of the method, amplification factors and capacitances may be determined in structures containing any number of grids.

Effects of Space Charge on Potential Distribution in Triodes

It was shown that in the absence of space charge a plane at a potential

$$E_q = \frac{E_c + \frac{E_b}{\mu}}{1 + \frac{1}{\mu}}$$

and at a distance

$$d_{kq} = d_{gk} + \frac{d_{gp}}{1 + \mu}$$

from the cathode could be substituted for the grid and anode without altering the potential distribution between grid and cathode, regardless of the position of the cathode. It follows that with space-charge-limited currents the error will be small in most cases if a triode be converted into an equivalent diode by the same substitution.

A more precise equivalent diode may be constructed by the use of the first expression for equivalent potential with a space-charge correction. Fig. 9 shows the potential distribution in a triode with space-charge-limited current. It is obvious that the field at the grid is the same as would exist without space charge if the cathode were at point h, determined by drawing a tangent to the potential curve at the grid. If it be assumed that the potential between cathode and grid varies as the four-thirds power of distance, d_{gh} is three-fourths of d_{gk}. Hence, we must modify the expression for E_G as follows:

$$E_G = \frac{E_c + \frac{E_b}{\mu}}{1 + \frac{1}{\mu} + \frac{4}{3}\frac{d_{gp}}{d_{gk}\mu}}$$

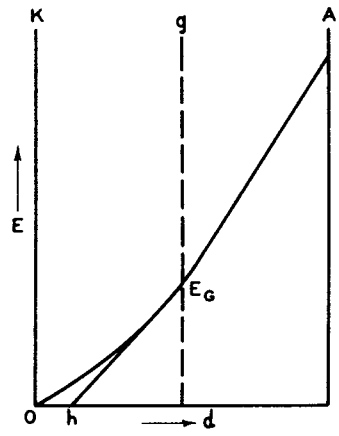


Fig. 9

The analysis of the current-voltage relationship of a triode may be made directly from the diode case by the use of this equivalent-diode expression. If in the equivalent diode the space current

$$I_b = f(E_G)$$

the cathode current (equal to the plate current with negative grid) is given directly. The transconductance, g_m, is found by taking the derivative of $f(E_G)$ with respect to E_c. The plate conductance, $1/r_p$ is found by taking the derivative of $f(E_G)$ with respect to E_b.

Electron Transit Time In Negative-Grid Triodes

The electron transit time in any electrode structure may be calculated readily if the potential distribution is known. In general

$$t = \int \frac{dx}{v} = \left(\frac{m}{2e}\right)^{\frac{1}{2}} \int \frac{dx}{E^{\frac{1}{2}}}$$

The calculation of transit time in the absence of space charge is obvious. In a parallel-plane diode with space-charge-limited current, the transit time from cathode to anode may be calculated if it be assumed that

$$E = E_b \left(\frac{x}{d_{kp}}\right)^{\frac{4}{3}}$$

whence

$$t = \left(\frac{m}{2e}\right)^{\frac{1}{2}} \frac{d_{kp}^{\frac{2}{3}}}{E_b^{\frac{1}{2}}} \int_0^{d_{kp}} x^{-\frac{2}{3}}\, dx$$

$$= \left(\frac{m}{2e}\right)^{\frac{1}{2}} \frac{3\, d_{kp}}{E_b^{\frac{1}{2}}}$$

$$= 5.05 \times 10^{-8} \frac{d_{kp}}{E_b^{\frac{1}{2}}}$$

where t is in seconds, d_{kp} in centimeters, and E_b in volts. In other words, the electron takes three times as long to pass from cathode to anode as if it had traveled at the final velocity the entire distance, and half again as long as if it had been uniformly accelerated.

The cylindrical analysis is not so simple but may be carried out as presented by W. R. Ferris.[9]

In the case of electron transits between grid and anode, the integration is carried out with the initial velocity of the electron corresponding to the equivalent potential of the grid.

9 W. R. Ferris, "Input Resistance of Vacuum Tubes as Ultra-High-Frequency Amplifiers," Proc. I.R.E., Vol. 24, p. 82; January, 1936.

Lecture 15 *

ELECTRON OPTICS

Part I. DETERMINATION OF ELECTRON TRAJECTORIES

V. K. Zworykin and G. A. Morton

INTRODUCTION

The term "electron optics" will be used to describe that class of problems which deals with the determination of electron trajectories. The expression originated as a consequence of the close analogy between optical arrangements and the corresponding electronic systems. It was found that this analogy not only had fundamental mathematical significance, but, in many cases, could be extended to practical devices. For example, it is possible to construct electron lenses which are capable of imaging an electron source. In many instances not only is the behavior of the two types of systems the same, but also many of the mathematical methods of optics can be applied to the corresponding electron problem. There are, it is true, many systems which in no way resemble those of conventional optics. However, there is a continuous transition between these and such as have a close optical analogue. Therefore, any attempt to subdivide the field on this basis results only in confusion.

The importance of electron optics is becoming increasingly apparent with the advance of electronics. For example, in the early vacuum tubes used in radio work little attention was paid to the exact paths of the electrons between the cathode and the plate. Recently, very real improvement in efficiency and performance has been achieved by the application of electron optics to tube design. In the design of the newer devices, such as the secondary-emission multiplier, the electron gun used in television tubes, and the image tubes, electron optics is essential.

The design problem usually encountered is one in which the two termini of the electron paths are specified and it is required to determine an electrode and magnetic coil configuration that will satisfy this demand. Unfortunately, a direct solution is still a good deal beyond our present mathematical means. It is not possible, except in very special cases, to determine from a given electron path the shape and potentials of the electrodes required to produce this path. In order to solve the above problem, it is necessary to assume an electrode configuration and then determine the resulting electron path. If this is not the required path the electrodes are changed and the trajectories recalculated. Usually this process does not have to be repeated very many times before the correct solution is reached, as the previously determined paths indicate the nature of the changes that must be made.

Restating the problem in the only form in which a solution is possible, it becomes: Given an electrode configuration and the potentials applied, determine the electron paths in the resulting potential field. Even this problem has no general solution, and often can be solved only by resorting to elaborate mathematical approximations, or to the use of mechanical and graphical methods. The solution can be divided into two distinct parts; namely, the determination of the potential field produced by the electrodes, and the calculation of the electron trajectories in this field. Essentially the same procedure is used when the electrons are guided by magnetic fields.

THE LAPLACE EQUATION

To determine the potential field produced by a given set of electrodes, it is necessary to solve the Laplace differential equation:

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2} = 0 \qquad (1)$$

with boundary conditions corresponding to the shapes and potentials of the electrodes. The solution of this equation gives the potential as a function of the coordinates, that is:

$$\varphi\ (x,y,z)$$

The electrostatic field can be found from this potential by differentiation with respect to the coordinates. Thus:

$$E_x = -\frac{\partial}{\partial x}\varphi(x,y,z)$$

$$E_y = -\frac{\partial}{\partial y}\varphi(x,y,z)$$

$$E_z = -\frac{\partial}{\partial z}\varphi(x,y,z)$$

From the original Laplace equation, which is satisfied by the potential function, it will be seen that the field must satisfy the differential equation:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0$$

---

* This lecture contains the same material as Chapter 3 of the book TELEVISION by Zworykin and Morton. Used by permission of John Wiley and Sons.

It should be noticed that this equation is like an equation of continuity and may be interpreted to mean that in any volume of free space within an electrode system as many electrostatic lines of force must leave as enter. Similar equations express corresponding laws obeyed by the flow of an incompressible fluid and by electric current in a conducting medium.

There is no general solution for the Laplace equation nor can any general method of attack be given. In certain special cases only can an analytic solution be obtained. Usually it is necessary to resort to series expansion or numerical integration in order to calculate a potential distribution. Both procedures are laborious in the extreme.

The simplest potential distribution is that between two infinite parallel plates, shown in

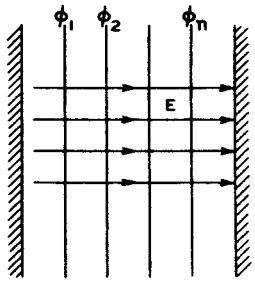


Fig. 1 - Potential and field between plates.

Fig. 1. Here the function that satisfies the differential equation is:

$$\varphi = \varphi_o + Ax$$

The field is found to be:

$$E_x = -\frac{\partial \varphi}{\partial x} = -A$$

$$E_y = E_z = -\frac{\partial \varphi}{\partial y} = 0$$

Other simple cases are:

Concentric spheres: $\varphi = -\frac{A}{r} + \varphi_o$

$$E_r = -\frac{A}{r^2}$$

Concentric cylinders: $\varphi = A \ln r + \varphi_o$

$$E_r = -\frac{A}{r}$$

Other cases, such as two separated spheres, a sphere and plane, a sphere between two planes, and the corresponding cylindrical systems can also be solved.

Problems involving cylindrical symmetry, such as illustrated in Fig. 2, are of considerable importance, since, as will be shown in the next

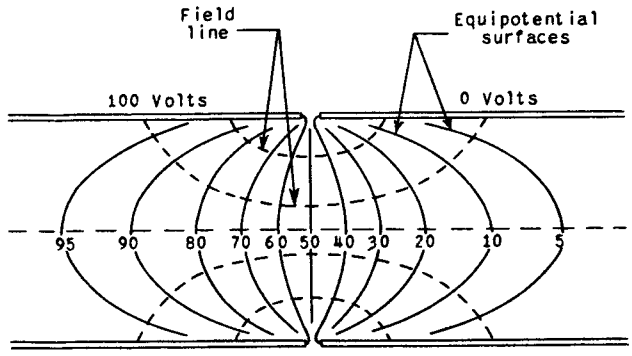


Fig. 2 - Potential distribution between two cylinders.

lecture, this symmetry is found in all electron lenses. When this symmetry is present, the Laplace equation is preferably expressed in cylindrical coordinates and then becomes:

$$\frac{\partial^2 \varphi}{\partial z^2} + \frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial \varphi}{\partial r}\right) = 0 \qquad (2)$$

For most boundary conditions, the solution of this differential equation is difficult, and no analytic solution is possible. A general method of attack is to consider the potential as a linear combination of functions in which the variables have been separated, thus:

$$\varphi(r,z) = \varphi_1 + \varphi_2 + \ldots\ldots + \varphi_k + \ldots\ldots$$

where :

$$\varphi_k(r,z) = F_k(z)\, G_k(r) \qquad (3)$$

When equation (3) is substituted in equation (2), the Laplace equation reduces to two ordinary differential equations:

$$\frac{1}{F}\frac{d^2F}{dz^2} = -k^2$$

$$\frac{1}{rG}\frac{d}{dr}\left(r\frac{dG}{dr}\right) = k^2$$

where k^2 is the separation parameter. The general solution of these equations can be written as:

$$F_k(z) = a\, e^{ikz} + b\, e^{-ikz} \qquad (4a)$$

$$G_k(r) = c\, J_o(ikr) + d\, N_o(ikr) \qquad (4b)$$

The solution of the Laplace equation then has the form:

$$\varphi(r,z) = \sum_k A_k F_k(z)\ G_k(r) \qquad (5)$$

Since k does not necessarily have discreet values, equation (5) may take the form of an integral:

$$\varphi(r,z) = \int A(k)\ F(z,k)\ G(r,k)\ dk \qquad (5a)$$

the integration being over the entire complex domain. The coefficient A(k) is determined from the boundary conditions by the usual methods of evaluating Fourier coefficients.

Another class of problem of considerable importance is that in which the potential expressed in Cartesian coordinates is a function of two of these coordinates only. This type of potential field is encountered wherever the electrode surfaces can be considered as generated by moving lines which remain parallel.

The Laplace equation in this case becomes:

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = 0 \qquad (6)$$

The solution of this equation can be approached in a variety of ways. One very useful method, for example, is that of conformal mapping. Although this equation can be solved more frequently than that for the three-dimensional case, often no analytic solution is possible. In this two-dimensional case, practical electrode configurations are usually quite complicated, so that the mathematical complexities even of an approximate solution are prohibitive.

ELECTROLYTIC POTENTIAL MAPPING

Because of the difficulties encountered in a mathematical solution of the Laplace equation, it is often expedient to resort to an electrolytic method of obtaining an equipotential map.

In essence, the method consists of immersing a large scale model of the electrode system being studied in a slightly conducting liquid and measuring the potential distribution throughout the liquid with a probe and bridge, potentials proportional to those to be used with the system being applied to the model. Fig. 3 shows diagrammatically an electrolytic plotting tank.

The tank used for this purpose is constructed of an insulating material so that the equipotential surfaces about the immersed electrodes meet the tank walls perpendicularly. This condition is necessary in order to reduce the influence of the walls of the tank upon the field to be plotted. The size of the tank is determined by the size of the electrode models which are to be studied, and this in turn is determined primarily by the accuracy desired. In order to reduce the boundary effects, the tank must be a good deal larger than the model.

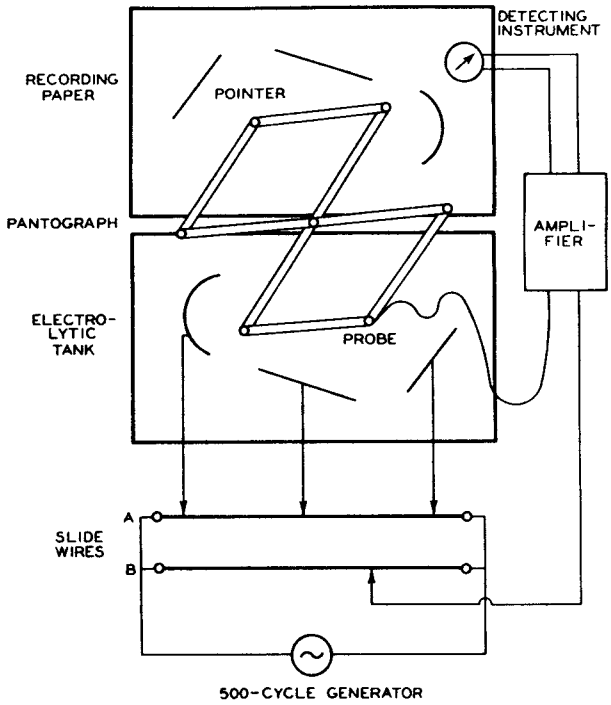


Fig. 3 - Diagram of potential plotting tank.

The electrolyte used in the tank is water to which a very small amount of soluble salt has been added. In most localities, ordinary city water contains a sufficient amount of these salts to make it amply conducting for the purpose.

Exact scale models of the electrodes, made of sheet metal and supported on insulating rods, are used in the tank. The supports should be reduced to a minimum so that they interfere with the potential distribution as little as possible. Almost every practical electrode configuration encountered in electron optics has mirror symmetry. The equipotential surfaces in space around the electrodes obviously must cross the plane of symmetry at right angles. Because of this requirement, as will become clear as the discussion proceeds, the models may be constructed so that they represent, to scale, half of each electrode bounded by the plane of symmetry. The model is placed in the tank in such a way that the free surface of the electrolyte coincides with the plane of symmetry. Although this restriction is not a fundamental limitation, nearly all practical plotting tanks are limited to use with electrode systems having this symmetry.

Upon the application of the proper potentials to the model electrodes, a current will flow through the electrolyte. Since it can be assumed that the electrolyte is an ohmic conductor, the

field strength at any point will be proportional to the current density. As was mentioned previously, the electric current behaves like an incompressible fluid so that the current density and hence the field strength obey the equation of continuity (i.e., their divergence is zero in the absence of sources or sinks). This is merely another way of saying that the potential throughout the electrolyte obeys the Laplace equation. Thus, the potential at any point in the liquid is proportional to the potential of a corresponding point in the actual electrode system.

The free surface of the electrolyte is a perfect insulating plane since no current can flow in the medium above the liquid. The equipotentials must intersect such a plane at right angles because there is zero vertical current flow, and hence the field vector normal to the surface is zero. For this reason it is possible to make use of models divided at their plane of symmetry.

The potential distribution over the plane of symmetry is measured by means of an exploring probe. This probe consists of a fine wire mounted so as to just break the surface of the liquid and is constrained to move in a horizontal plane. The potential of the probe is adjusted until zero current flows, and the potential is noted. This potential is that of the point where the probe touches the surface. For convenience, the probe is carried at the end of a pantograph linkage, so that the motion of the probe is reproduced by a stylus attached to the other end of the linkage. This stylus, or mapping pencil, moves over a plotting table. The arrangement will be clarified by reference to Fig. 3.

A photograph of a typical plotting tank is reproduced in Fig. 4. The tank itself is made of wood, coated on the inside with roofing cement to render it water-tight and shielded outside with sheet copper. It is 2-1/2 feet wide, 8 feet long, and 2-1/2 feet deep. Along one side is a table on which the mapping is done. Directly below the table are the potential dividers that supply the model electrodes and probe, and behind them the amplifier whose output is connected to the null-indicating meter. The probe is attached to the pantograph pivoted at the center of the edge of the tank nearest the mapping table. A shielded lead carries the current from the probe to the amplifier.

In the equipment illustrated in Fig. 4, the probe and electrodes are supplied from a 400-cycle oscillator, instead of with direct current. The use of the oscillator supply not only facilitates the determination of the null point, but also avoids the possibility of error due to polarization of the liquid at the probe or electrodes.

The usual procedure in mapping a potential distribution is to divide the voltage between the terminal electrodes into a convenient number of intervals, then to set the probe potential at each of these values in turn and map the path of the probe when it is moved in such a way that the current to it remains zero. The resulting map of the intersection of the equipotential surfaces corresponding to the voltage steps with the plane of symmetry of the electron optical system is the most convenient representation of a potential distribution for the determination of electron trajectories.

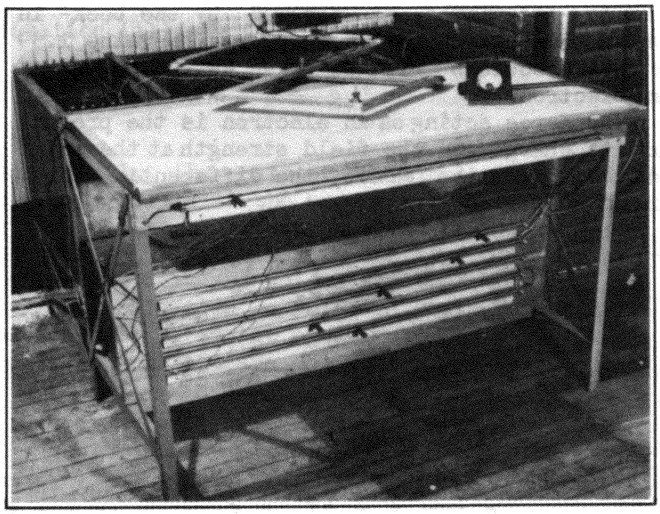


Fig. 4 - Potential plotting tank.

Often, in the consideration of electron lenses, to be taken up in the next lecture, it is necessary to know the axial distribution of the potential of the system, together with its first and second derivatives along the axis. The distribution can, of course, be found by direct measurement with the plotting tank. The slope of the distribution curve, plotted as a function of the axial coordinate, will give the first derivative. The second derivative can be found from the slope of the first derivative curve, but this method of obtaining it is very inaccurate. A more accurate determination can be made by means of the radius of curvature ρ of the equipotentials at the axis, the first derivative φ' and the relation:

$$\varphi''(z) = \frac{2\varphi'}{\rho}$$

THE MOTION OF AN ELECTRON IN A POTENTIAL FIELD

The potential distribution for a given electrode configuration having been obtained, the next step is the determination of the paths of electrons moving in this field.

In making this determination, it is convenient to consider the motion of an electron to be that of a charged particle of mass m obeying the laws of Newtonian mechanics, rather than to adopt the viewpoint of quantum physics and assume it to be a wave packet, as is necessary in the investigation of atomic phenomena. Furthermore, its mass

will be taken as constant and equal to 9.0×10^{-28} gram. Where electrons having extremely high velocities are to be considered, this assumption cannot be made, and it is necessary to correct for the increase in mass as dictated by relativity. Velocities where this correction is necessary are not encountered in the field of electronic television, as covered by the book[1] in which this lecture is one chapter. The value 1.59×10^{-19} coulomb will be taken as the charge of an electron.

The force acting on an electron is the product of its charge and the field strength at the point which it occupies. Hence, the differential equations of motion are:

$$m \frac{d^2x}{dt^2} = - eE_x = e \frac{\partial \varphi}{\partial x}$$

$$m \frac{d^2y}{dt^2} = - eE_y = e \frac{\partial \varphi}{\partial y}$$

$$m \frac{d^2z}{dt^2} = - eE_z = e \frac{\partial \varphi}{\partial z}$$

In principle, in order to determine the electron path, all that is necessary is to introduce the values of the potential into the above equations, solve, and eliminate time as a parameter. Actually, there is no general method of carrying out this process, and it is almost always necessary to apply mathematical approximations or graphical methods to obtain a solution.

A number of practical mathematical approximations and graphical methods have been developed for the purpose of facilitating the determination of electron paths when the potential field is known. These methods, when carefully applied, are capable of yielding a high degree of accuracy.

The general three-dimensional problem is extremely difficult even by approximate methods. Fortunately, configurations requiring the solution of this general problem are rarely encountered, at least at the present time.

Two classes of problems are of particular importance. They are:

1) Those involving electrode configurations in which the potential variation is confined to a plane.

2) Problems involving cylindrical symmetry.

The remainder of this lecture will treat the first class of problems, which are those involved in the design of the electron multiplier, deflecting plates, etc.

[1] V. K. Zworykin and G. A. Morton, "Television," John Wiley and Sons, Inc.

ELECTRON PATHS IN A TWO-DIMENSIONAL SYSTEM

Where the potential variation is confined to a plane, the Laplace equation, as has already been pointed out, involves two coordinates only:

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0$$

Similarly, the laws of motion become:

$$\left.\begin{aligned} \frac{d^2x}{dt^2} &= + \frac{e}{m} \frac{\partial \varphi}{\partial x} \\ \frac{d^2y}{dt^2} &= + \frac{e}{m} \frac{\partial \varphi}{\partial y} \end{aligned}\right\} \qquad (7)$$

This particular form of these laws is not very convenient in the present consideration, in that time enters the equations explicitly.

Taking the principle of least action as a starting point simplifies the treatment, but it should be noted that all the relations derived below can be derived directly from the force laws of equation (7).

The principle of least action states that any particle moving between two points in a potential field will follow a path such that the integral of the momentum over this path is an extreme, either maximum or minimum. Symbolically, this principle can be written as:

$$\delta \int_A^B mv\,ds = 0 \qquad (8)$$

An electron moving in a potential field has a kinetic energy just equal to the decrease in its potential energy during its motion. If the potential is set equal to zero at a point where the electron is at rest, the following relation applies:

$$\frac{mv^2}{2} = e\varphi$$

or

$$v = \sqrt{2 \frac{e}{m} \varphi}$$

The momentum in the action integral being represented by $\sqrt{2em\varphi}$, equation (8) becomes:

$$\delta \int_A^B \sqrt{\varphi}\, ds = 0 \qquad (9)$$

or

$$\delta \int_A^B \sqrt{\varphi} \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx = 0 \qquad (9a)$$

This is satisfied by a solution of the corresponding Euler differential equation:

$$\frac{d^2y}{dx^2} = \frac{1}{2\varphi}\left(\frac{\partial\varphi}{\partial y} - \frac{dy}{dx}\frac{\partial\varphi}{\partial x}\right)\left[1 + \left(\frac{dy}{dx}\right)^2\right] \qquad (10)$$

which is directly derivable from the variation principle. Where the numerical values of the potential as a function of x and y are known, it is possible to perform a point-by-point integration of this equation (e.g., by the method of differences) and thus determine the trajectories of electrons in this field. Although extremely laborious, this is probably the most accurate method of obtaining electron paths.

GRAPHICAL TRAJECTORY DETERMINATION

There are graphical methods for plotting electron paths on an equipotential map which are easy, rapid, and sufficiently accurate for most practical problems. Two of these are of sufficient importance as practical design tools to be worth discussing in detail.

The first is a circle method. Referring to Fig. 5, let it be assumed that an electron is

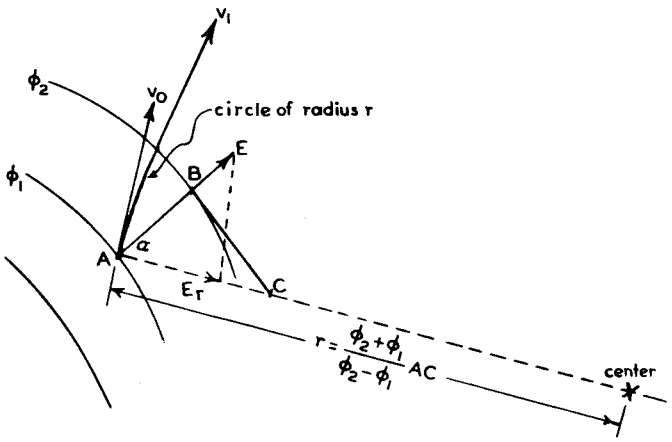


Fig. 5 – Circle method of graphical ray tracing.

moving in a potential field with the velocity indicated by the vector v_0. The magnitude of this velocity vector is:

$$v_0 = \sqrt{2\frac{e}{m}\varphi_1} \qquad (11)$$

The electric field, E, is normal to the equipotential φ_1 and has a value approximately equal to

$$\frac{\varphi_1 - \varphi_2}{d} = E$$

One component of the field lies along the direction of motion; the other, E_r, is at right angles to this direction. The latter exerts a radial force on the electron equal to eE_r giving rise to a centripetal acceleration:

$$\frac{v^2}{r} = \frac{eE_r}{m}$$

Solving for r and eliminating v with the aid of equation (11), it follows that

$$r = 2\frac{\varphi}{E_r} \qquad (12)$$

Accordingly, the path of the electron coincides approximately with the arc of a circle of radius r tangent to the vector v_0. Actually, this arc should be infinitesimal in length, but since the equipotentials are close together, it may for this approximation be extended to meet the next equipotential, φ_2. At φ_2 the velocity vector will be tangent to the arc and will have a magnitude:

$$v_1 = \sqrt{2\frac{e}{m}\varphi_2}$$

If the procedure at the successive equipotentials is repeated, the path of the electron can be mapped.

The radius and center of the arc can be found graphically to avoid the calculation indicated in equation (12). First, the approximate direction of the field vector, E, is obtained by dropping a perpendicular from A, the intersection of the path with φ_1, onto the equipotential, φ_2, cutting it at B. At right angles to the line AB, a line is extended until it meets the normal to the velocity vector at C. It is evident that E_r must lie along the normal to the velocity vector and that the center of the arc must also be located on this line. If the angle between E and E_r is α, then:

$$E_r = E\cos\alpha = \frac{\varphi_2 - \varphi_1}{AB}\cos\alpha$$

and

$$r = 2\frac{\varphi_2}{\varphi_2 - \varphi_1}\frac{AB}{\cos\alpha} = 2\frac{\varphi_2}{\varphi_2 - \varphi_1}AC \qquad (13)$$

Somewhat greater accuracy is obtained if φ_2 in the numerator of equation (13) is replaced by the mean potential, giving:

$$r = \frac{\varphi_2 + \varphi_1}{\varphi_2 - \varphi_1}AC \qquad (13a)$$

With the aid of this construction, path plotting can be carried out rapidly and accurately.

When the path in question starts from a surface of zero potential it is convenient to make

use of the fact that it issues normal to the surface and has an initial radius of curvature three times as large as that of the electrostatic field lines in the neighborhood of its point of origin.

The second graphical procedure is the parabola method. This method is advantageous where the curvature of the path is small. In such cases the radii are made awkwardly large when the circle method is used. Theoretically, if the process is carried to the limit and the separation between the equipotential lines is made to approach zero, either method gives the true path. In all practical cases tested, the accuracy of the two methods is about the same.

The parabola method is based upon the fact that an electron moving in a uniform field having a velocity component at right angles to the field follows a parabolic path. It utilizes the geometric principle illustrated in Fig. 6a, namely, that the tangent to a parabola at a point at an axial distance $\underline{x}$ from its vertex meets the axis at point C, located at -x, or at an axial distance 2x from the point of tangency.

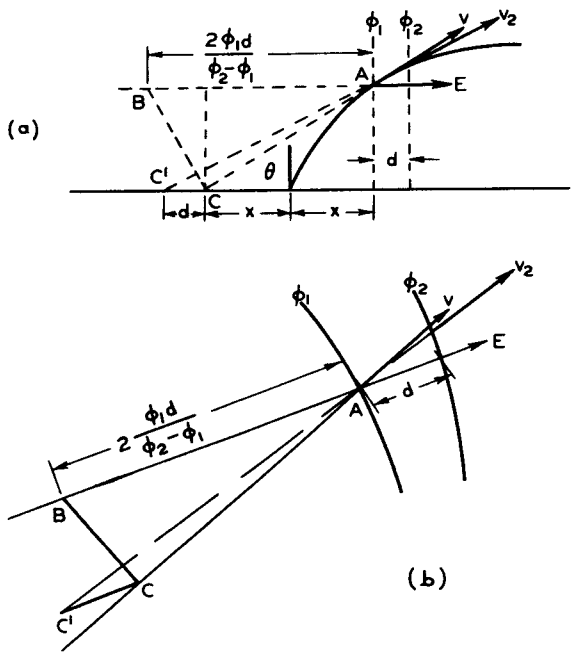


Fig. 6 - Parabola method of ray tracing.

Again referring to Fig. 6a, let it be assumed that an electron at point A is moving with a velocity $\underline{v}$, as indicated by the vector $\underline{v}$, and that its motion is due to a uniform field. It is possible to determine the parabola giving its motion as follows: The component of velocity due to the electron having fallen through the potential field to point A is v cos θ, where θ is the angle between the velocity and field vectors. The difference of potential φ^* between point A and the vertex of the parabola will, therefore, be:

$$\varphi^* = \frac{1}{2}\frac{m}{e} v^2 \cos^2 \theta \qquad (14)$$

If two equipotentials, φ_1 and φ_2, are separated by a distance $\underline{d}$, the field, E, can be expressed as

$$\frac{\varphi_2 - \varphi_1}{d}$$

The distance between the vertex of the parabola and the point A can consequently be written as:

$$x = \frac{\frac{1}{2}\frac{m}{e} v^2 d \cos^2 \theta}{\varphi_2 - \varphi_1}$$

$$= \frac{\varphi_1}{\varphi_2 - \varphi_1} d \cos^2 \theta \qquad (15)$$

Equation (15) indicates a simple construction which will locate the point C as shown in Fig. 6a. The field vector, E, is extended back a distance

$$\frac{2\,\varphi_1 d}{\varphi_2 - \varphi_1}$$

to B. From B a perpendicular is dropped to an extension of the velocity vector. This perpendicular will cut this vector at point C. Thus, this point can be determined from a knowledge of the vectors, v and E. Further, if a line parallel to E is drawn through C, and extended back a distance $\underline{d}$ to C', this new point must lie on the velocity vector for the point on the parabola where it intersects the equipotential φ_2. If a line is drawn through C' and A, it closely represents the path between the equipotentials φ_1 and φ_2.

To apply this construction to a general two-dimensional potential field, the procedure is as follows: Referring to Fig. 6b, the electron is at A, moving with a velocity and direction given by $\underline{v}$. From point A, a line E representing the field direction is drawn normal to the equipotential φ_1, and extends a distance $\underline{d}$ to φ_2. This line is drawn back from A a distance

$$2\,\frac{\varphi_1}{\varphi_2 - \varphi_1}\,d$$

to point B. A perpendicular is then dropped from B onto the prolongation of $\underline{v}$, locating point C. From C a line parallel to AB is drawn back a distance $\underline{d}$, locating the point C'. A line through C' and A locates the position of the electron on the equipotential φ_2, and gives the direction of its velocity vector.

As the curvature of the path decreases this method becomes increasingly accurate. It, there-

fore, is useful for determining the straighter portions of an electron trajectory, where, as has already been mentioned, the circle method becomes awkward because of the long radii involved.

The two plotting methods just discussed can be applied in any problem where the motion of the electron is confined to a plane; thus, they apply to any electrode configuration whose potential can be correctly mapped in a plotting tank of the type described.

THE RUBBER MODEL

By far the most convenient method of obtaining electron paths is by means of the rubber model. This model can be used in all problems where the potential can be expressed as a function of two rectangular coordinates, and where the electron path is confined to the plane of these coordinates. The accuracy which can be obtained is quite high, but not quite equal to that of a path carefully plotted by the graphical methods described.

A rubber membrane, stretched over a frame, is pressed down over a model of the electrode system which is made in such a way that its plan view corresponds to the geometrical configuration of the electrodes in the x - y plane while the height is proportional to the negative voltage on each electrode. The rubber is then no longer flat, nor does it follow the surfaces of the model electrodes, but rather it stretches over them in a series of mountains and valleys, touching only the top of every electrode. If care is taken that the membrane is in contact with the full length of the top edge of all the electrodes, the contour of its surface is found to correspond to an equipotential map of the electrode system.

A small sphere is placed at the point corresponding to the electron source and allowed to roll on the rubber. The horizontal projection of its path is a map of an electron trajectory in the electrode system under investigation.

The proof that the path of the sphere correctly represents the electron trajectory is divided into two parts. First, it is necessary to show that the height of every point on the rubber model is proportional to the potential existing in the electrode system. Second, it must be proved that a sphere rolling on such a surface follows a trajectory which represents the electron motion.

In order to show that the height z of the rubber surface represents the potential distribution, it is necessary to show that the surface obeys the differential equation:

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0 \qquad (16)$$

This may readily be demonstrated if two restrictions, which are more or less fulfilled in practice, are imposed on the rubber surface. These are: (1) that the slope of the surface be everywhere small, and (2) that the tension of the deformed rubber be uniform over the surface.

The most straightforward proof applies the principle of minimum energy. Since the energy in any region is proportional to the area, the area integral of the surface must be a minimum. Transforming this minimized integral into the Euler form leads to the differential equation required.

The physical significance of the shape assumed by the surface is more apparent if the following less rigorous demonstration is used. Fig. 7 shows an element of surface area, ds, together with the forces acting on it. It is obvious that the vector sum of these forces must be zero since the element is in equilibrium. In order to set up the conditions of equilibrium, the four forces F_1, F_2, F_3, and F_4 must each be resolved into their components along the coordinates. Considering first F_1, it is evident from the figure that:

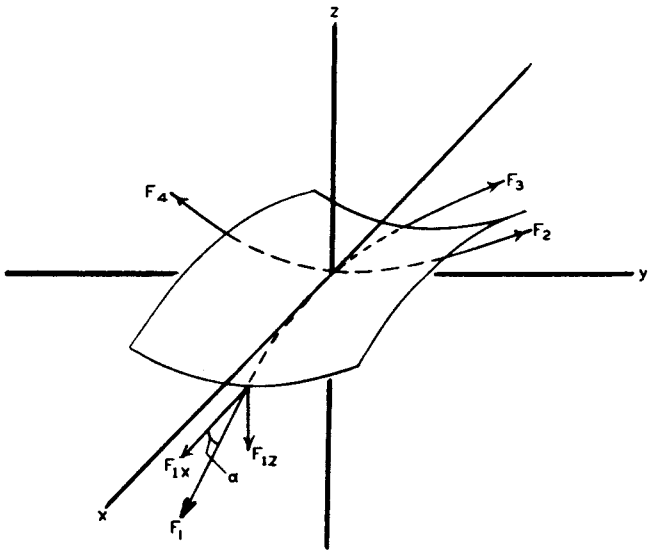


Fig. 7 - Forces acting on an element of stretched membrane.

$$F_{1x} = F_1 \cos\alpha = F_1 \frac{dx}{dx} \frac{1}{\sqrt{1 + \left(\frac{dz}{dx}\right)^2}} \qquad (17a)$$

$$F_{1z} = F_1 \sin\alpha = F_1 \frac{dz}{dx} \frac{1}{\sqrt{1 + \left(\frac{dz}{dx}\right)^2}} \qquad (17b)$$

where dz, dx, and α are as indicated. By the first restriction, dz/dx is small, so that

$$\frac{1}{\sqrt{1 + \left(\frac{dz}{dx}\right)^2}} \cong 1$$

and the components become:

$$F_{1_x} \cong F_1$$

$$F_{1_z} \cong F_1 \left(\frac{dz}{dx}\right)_1$$

Similarly

$$F_{3_x} \cong F_3$$

$$F_{3_z} \cong F_3 \left(\frac{dz}{dx}\right)_3$$

The two x components must be equal in magnitude and opposite in sign. Hence,

$$F_{3_x} \cong - F_{1_x}$$

$$F_3 \cong - F_1$$

and

$$F_{3_z} = - F_1 \left(\frac{dz}{dx}\right)_3$$

Summing the z components gives the upward force due to F_1 and F_3 which is:

$$F_{1_z} + F_{3_z} = F_1 \left[\left(\frac{dz}{dx}\right)_1 - \left(\frac{dz}{dx}\right)_3\right] = F_1 \frac{\partial^2 z}{\partial x^2} \Delta x \qquad (18)$$

where Δx is the length of the element in the x direction.

Applying the second restriction, that the force F_1 must equal the tension δ of the membrane times the width Δy of the element in the y direction, we find that equation (18) becomes:

$$F_{1_z} + F_{3_z} = \delta \frac{\partial^2 z}{\partial x^2} \Delta x \, \Delta y \qquad (18a)$$

In like manner, the vertical components of the other two forces are:

$$F_{2_z} + F_{4_z} = \delta \frac{\partial^2 z}{\partial y^2} \Delta y \, \Delta x$$

Finally, since the sum of all the z force components must be zero, and δ, Δx, and Δy are not zero, the equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0$$

must be true, and proves that z satisfies the Laplace equation for the boundary conditions determined by the electrode heights.

It may be mentioned at this point that the slope of the rubber surface is everywhere proportional to the field strength, and that the force exerted on any electrode by the rubber is proportional to the capacity of that electrode.

The next problem is to show that a body moving under the action of gravity on the rubber surface moves along a path which corresponds to the electron trajectory. For simplicity, let it be assumed that the body in question slides on the surface, and that its friction is negligible.

By the principle of least action, the action integral must be stationary, or

$$\delta \int_A^B mv\,ds = 0 \qquad (8)$$

Since, the system is conservative and hence the sum of the kinetic and potential energies must remain constant, the momentum mv can be found as follows:

$$KE + PE = \text{constant}$$

$$\frac{1}{2} mv^2 = - mgz$$

$$mv = \text{constant} \sqrt{z}$$

By substitution, equation (8) becomes:

$$\delta \int_A^B \sqrt{z}\, ds =$$

$$\delta \int_A^B \sqrt{z} \sqrt{1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2} dx = 0 \qquad (19)$$

It has already been assumed that the slope of the rubber is small so that $(dz/dx)^2$ can be neglected compared with unity. Therefore, the final form of the action integral is

$$\delta \int_A^B \sqrt{z} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx = 0 \qquad (19a)$$

Except for z replacing φ, equation (19a) is seen to be identical with equation (9a). In the previous derivation it was shown that the height z of the rubber is proportional to the potential φ. Therefore, the path of the body sliding on the rubber is geometrically similar to the corresponding electron trajectory.

If, instead of the motion of a sliding body that of a rolling sphere is considered, the conclusions are the same, provided the assumption is

made that the radius of curvature R of the sphere is small compared with the radius of curvature of the rubber. This is shown by deriving the total kinetic energy as follows:

Let $d\alpha$ be a small rotation of the sphere. Then the displacement, ds, of the center of the mass is given by:

$$ds = R\, d\alpha$$

The angular velocity in terms of the linear velocity is thus:

$$\omega = \frac{d\alpha}{dt} = \frac{1}{R}\frac{ds}{dt} = \frac{v}{R}$$

Next, the sum of the rotational and translational kinetic energies is expressed as follows:

$$KE = \frac{1}{2}\left(mv^2 + I\omega^2\right) = \frac{1}{2}\left(m + \frac{I}{R^2}\right)v^2$$

$$= \frac{1}{2}\, m^* v^2$$

where m^* is the effective mass.

As before, the total momentum can now be written as $\sqrt{z}$ times a constant. Hence equation (19a), expressing the principle of least action, is unchanged and the path of the rolling sphere indicates the desired path.

Certain assumptions were made in deriving the paths on the rubber model and the question might well arise as to how closely these assumptions must be fulfilled in order not to introduce serious errors in the final results. As the result of a large number of tests on the model, the indications are that even if the slopes become as great as 30° to 45°, and the tension in the rubber very far from uniform, the paths obtained will be sufficiently accurate for all practical purposes. In fact, the presence of friction, which has been neglected in the above derivation (except the implied friction required to produce rolling), makes it advisable to use fairly steep slopes.

A practical form of the rubber model is shown in Fig. 8. Ordinary surgical rubber is stretched over a square frame, which is about 3 feet on a side. Usually the electrode models are made from soft metal strips, either lead or aluminum, which are cut to the correct height to represent the potential, and bent to conform with the electrode shape. The table supporting the electrodes is built of welded angle iron and has a plate-glass top. The glass top permits illumination from below and as a result greatly facilitates the placing of the electrodes. Since it is often necessary to press the rubber down to make it come in contact with the more positive electrodes, the table is equipped with movable side arms to which can be clamped top electrode models.

It has been found convenient to use 3/16-inch steel ball bearings for the spheres. These have an advantage over glass spheres in that they can be held at the top of the cathode electrode with a small electromagnet and released when desired, without any danger of deflecting their course, merely by cutting off the current to the magnet.

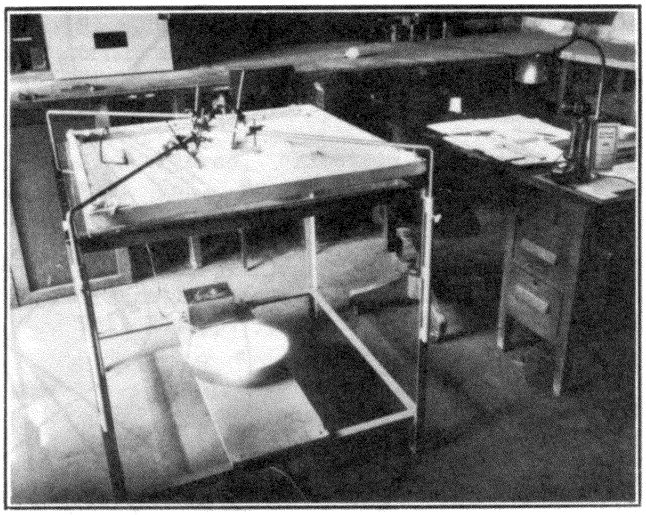


Fig. 8 - Rubber model for the study of electron paths.

Where a permanent record of the path is desired, a time exposure for the duration of the motion of the sphere can be made. In this case it is better to use black rubber for the model, and to illuminate it from above. Furthermore, with a pulsating light source such as a 60 cycle cold arc, the paths will appear as dotted lines. The spacing between the dots is a measure of the velocity of the electron.

Numerical values for the accuracy that can be attained with the rubber model are difficult to give since the error depends upon the electron path. Measurements of the parabolic paths obtained on such a model were made by P. H. J. A. Kleynen, who found an error of 1 per cent in the height of the parabola, and one of 7 per cent in the separation between the arms of the parabola when the sphere returned to its initial potential.

Results obtained with the model when used in connection with the design of the electrostatic multiplier shown in Figs. 9 and 10 indicate its

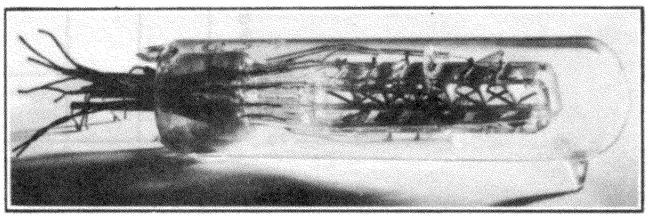


Fig. 9 - An experimental electrostatic secondary-emission multiplier.

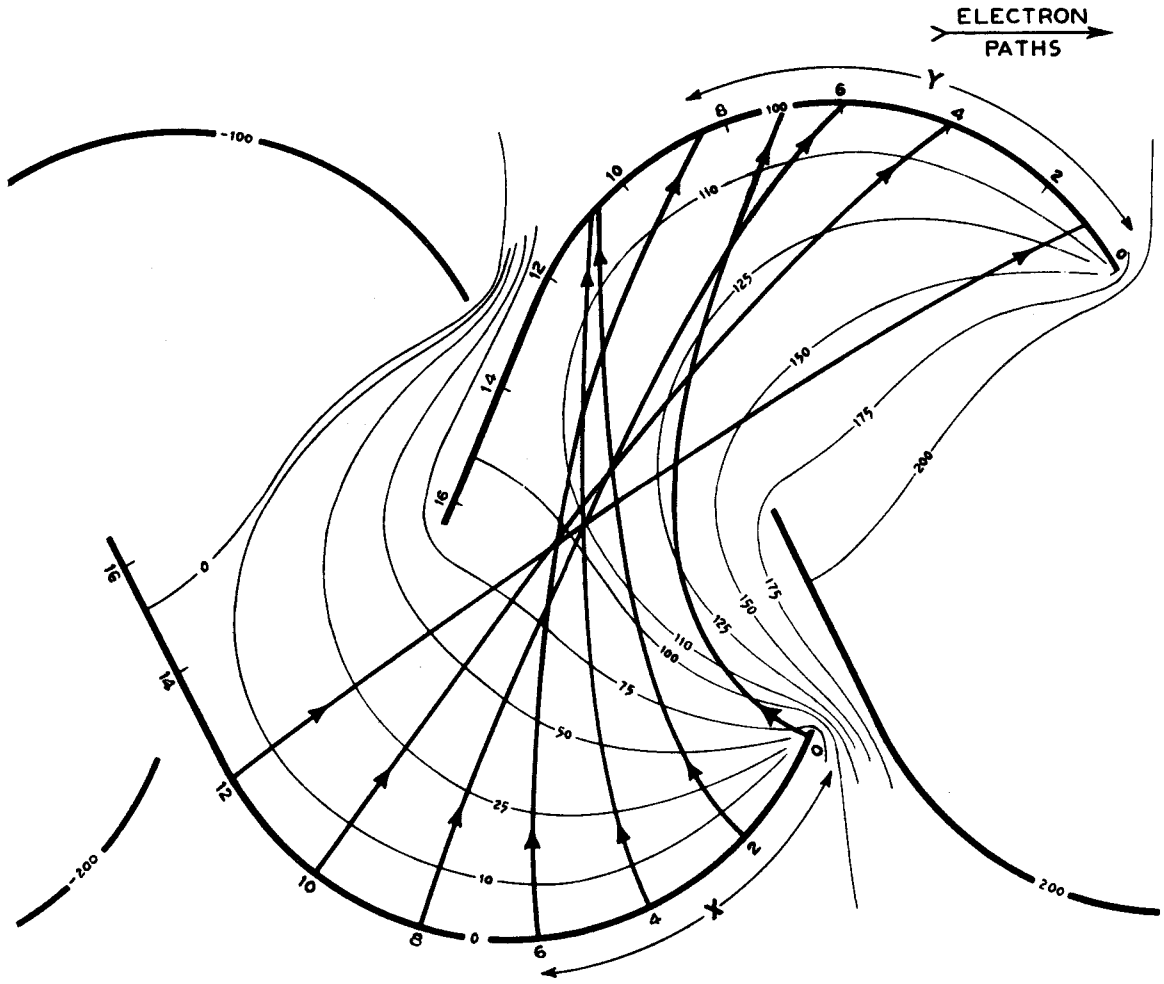


Fig. 10 - Electron paths in a multiplier of the type shown in Fig. 9 as determined from potential plot.

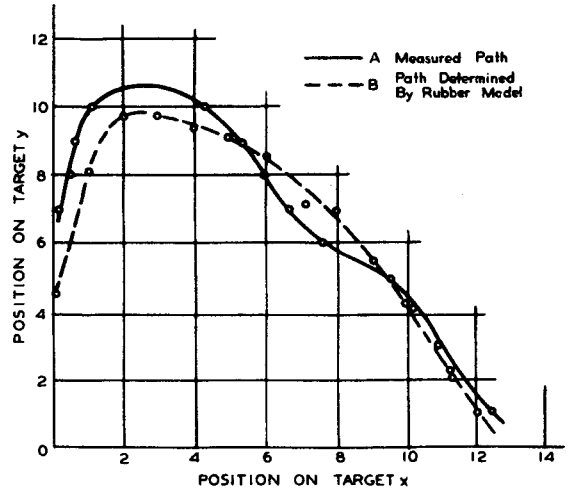


Fig. 11 - Curves showing terminii of electron trajectories in the multiplier as determined by direct measurement and from a rubber model.

excellence. Plots were made of the initial and terminal points of a number of electrons, as indicated by the model, and again with an actual electron tube. Two curves of this type are reproduced in Fig. 11, and show fairly close agreement between the two methods of measurement.

REFERENCES

1. J. H. Jeans, "Electricity and Magnetism," Cambridge University Press, 1927.

2. M. Mason and W. Weaver, "The Electromagnetic Field," University of Chicago Press, 1929.

3. O. D. Kellogg, "Foundations of Potential Theory," Julius Springer, Berlin , 1929.

4. D. Gaber, "Mechanical Tracer for Electron Trajectories," Nature, Vol. 139, p. 373; February 27, 1937.

5. H. Sallinger, "Tracing Electron Paths in Electric Fields," Electronics, Vol. 10, No. 10, pp. 50-54; October, 1937.

6. P. H. J. A. Kleynen, "The Motion of an Electron in a Two-Dimensional Electrostatic Field," Philips Tech. Rev., Vol. 2, No.11, pp. 321-352; 1937.

7. J. R. Pierce, "Electron Multiplier Design," Bell Tel. Record, Vol. 16, No.9, pp. 305-309; 1938.

8. V. K. Zworykin and J. A. Rajchman, "The Electrostatic Electron Multiplier," Proc. I.R.E., Vol. 27, pp. 558-566; September, 1939.

Lecture 16 *

ELECTRON OPTICS

Part II. ELECTRON-OPTICAL SYSTEMS WITH CYLINDRICALLY SYMMETRICAL FIELD-PRODUCING ELEMENTS

V. K. Zworykin and G. A. Morton

INTRODUCTION

The electron-optical systems to be considered in this lecture are those with cylindrically symmetrical field-producing elements. All electron lenses are based on configurations of this type. Conversely, it may be stated quite generally that any varying electrostatic or magnetic field which has cylindrical symmetry is capable of forming a first-order image, either real or virtual.

The practical importance of electron lenses is only now becoming apparent, although they have for some time attracted considerable attention in the realm of pure science.

Perhaps the first practical application of the electron lens was in the electron gun of cathode-ray oscilloscopes. This has been greatly improved in the past few years and is now used for television purposes in the Kinescope and the Iconoscope.

Another early use of the electron lens is found in the electron microscope. There are a great many forms of this microscope, all based essentially on the same principle. Fig. 1 illustrates a high-magnification instrument utilizing magnetic electron lenses. Not only is the electron microscope an important aid in the study of cathodes, secondary emitters, and metal surfaces, but, in addition, it has recently been adapted to biological work, and permits higher useful magnifications than can be obtained by optical means.

A third application which should be mentioned is its use in connection with the image tube. The image tube is of importance because it can be combined with the Iconoscope to make a television pickup tube which is many times more sensitive than the normal Iconoscope. Here again, there are a number of possible forms using electrostatic, magnetic, or combined electrostatic and magnetic lenses. An electrostatic image tube is shown in Fig. 2.

In order to make intelligent use of electron lenses, it is necessary to have a rather complete understanding of the process of image formation. The theory of the formation of images in electron optics may be regarded as an extension of that applying to light optics. It is, therefore, not out of place to review briefly the elements of ordinary optics.

OPTICAL PRINCIPLES

When a ray of light passes through a boundary between two media in which the velocity of light differs, the ray is bent by a process known as refraction. The law governing this refraction is the well-known Snell's law:

$$n \sin \beta = n' \sin \beta' \qquad (1)$$

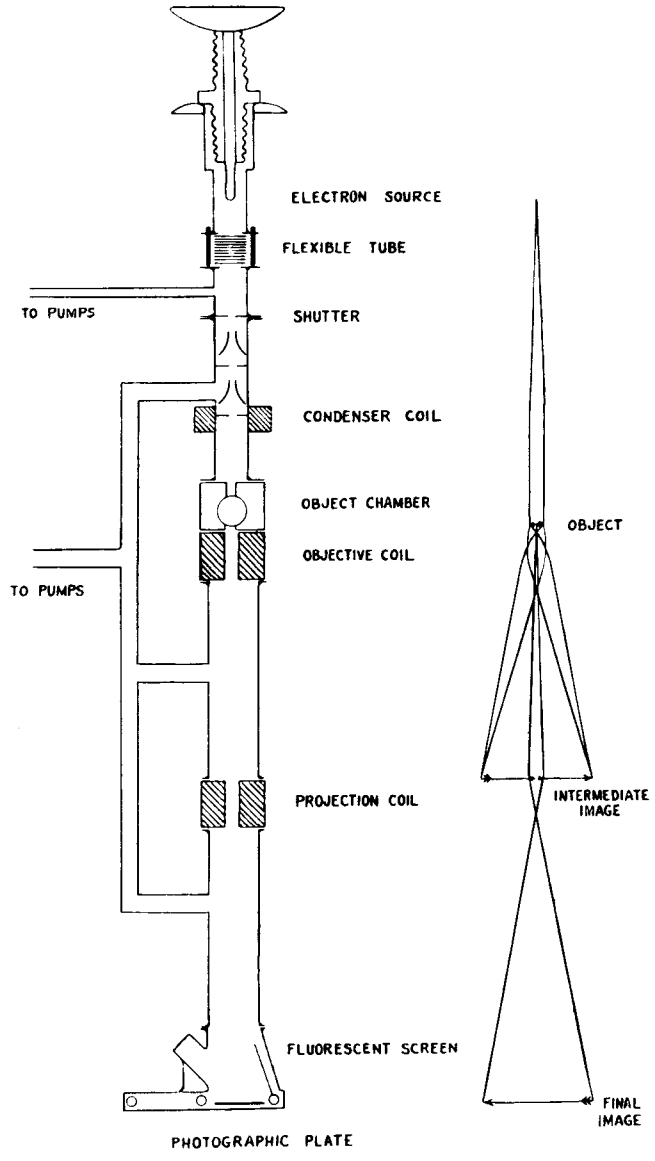


Fig. 1 - Electron microscope utilizing magnetic electron lenses.

* This lecture contains the same material given in pp. 91-120 of the book on TELEVISION by Zworykin and Morton. Used by courtesy of John Wiley and Sons.

where β and β' are the angles which the incident and refracted ray make with the normal to the boundary between the media having refractive indices of n and n', respectively.

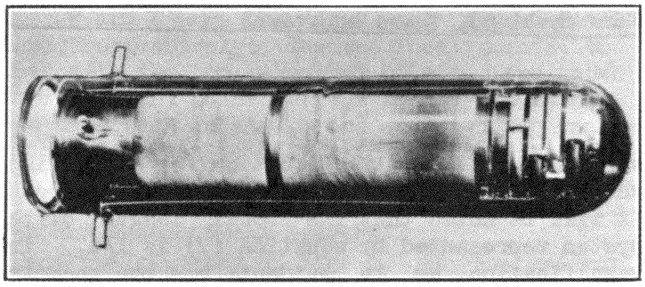


Fig. 2 - An electron image tube.

If the boundary separating the two media is a section of a spherical surface, a lens will be formed. Such a surface can be shown to have image-forming properties. By this is meant that, if the light rays from a small object enter the spherical refracting surface, the rays from any point will be bent in such a way that they converge on, or appear to diverge from, a second point known as its image point. Furthermore, the image points will be ordered in the same way as the emitting or object points, so that an image is formed of the original object. Where the rays travel so that they actually converge on the image points, the image formed is said to be real; if it is necessary to extend the rays backward to the point from which they appear to diverge, the image is virtual.

To prove the existence of the image-forming property of these spherical surfaces, it is necessary to show, first, that the rays from an object point converge on an image point and, second, that the ordering of object and image points is similar. The carrying out of this demonstration requires the imposing of two restrictions: namely, that the object be small, and that the rays make very small angles with the normal from the object point to the surface. Rays which meet these requirements are known as paraxial rays, and the image theory based on these restrictions is called the first-order theory, or Gaussian dioptrics.

In Fig. 3, P is the object point at a distance s from the spherical boundary of radius R. The ray PA emitted at an angle θ is refracted so that it reaches the axis at s', and thus makes an angle θ' with the axis. From the geometry of the figure, it is obvious that the angles of incidence and refraction are:

$$\beta = \theta + \frac{r}{R}$$

$$\beta' = -\theta' + \frac{r}{R}$$

Since, under the restrictions imposed, the angles of incidence and refraction are small, the sines appearing in Snell's law may be replaced by the angles themselves. Equation (1) becomes

$$n\beta = n'\beta' \qquad (2)$$

and, substituting in equation (2), it follows that

$$n\theta + \frac{rn}{R} = -n'\theta' + \frac{rn'}{R}$$

or

$$n\theta + n'\theta' = \frac{r}{R}(n' - n) \qquad (3)$$

However, $\theta = r/s$ and $\theta' = r/s'$, so that equation (3) can be written as follows:

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n' - n}{R} \qquad (4)$$

Since neither r nor θ appears in this expression, the equation proves that any ray leaving P must converge on the point P' at a distance s' from the surface. The point P' is, therefore, the image point of the object point P.

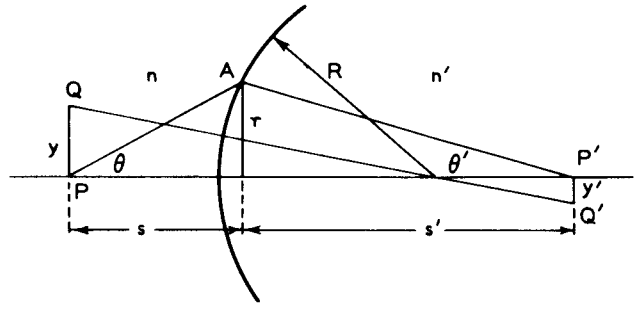


Fig. 3 - Refraction at a curved surface.

Considering next a particular ray QQ', which is normal to the refracting surface, it is evident from the similar triangles formed that

$$\frac{y'}{y} = \frac{+(s' - R)}{R + s}$$

Combined with equation (4), this becomes:

$$\frac{y'}{y} = \frac{+ns'}{n's} \qquad (5)$$

Since y'/y is the ratio of the separation of the two object points PQ, and the two image points P'Q', it is the lateral magnification m of the image. This magnification is independent of which point on the image is chosen; therefore, the ordering of the image and object points will

be similar, and the second condition for image formation is fulfilled.

Finally, substituting the relation $s'/s = \theta/\theta'$ into equation (5) gives the following equation:

$$ny\theta = n'y'\theta' \qquad (6)$$

This equation expresses the law of Lagrange and is valid for the first-order image formation by any number of refracting surfaces.

Referring back to Fig. 3, it is evident that it is possible to conceive of a plane surface which would bend the rays in exactly the same way as the spherical surface. As can be seen from the figure, the amount of bending required to produce an image would be

$$\alpha = kr \qquad (7)$$

where α is the angle between the incident and refracted ray, and r the radial distance from the axis at which the ray meets the surface.

THE THIN LENS

The formulas just derived are sufficient to determine the size and position of the first-order image formed by any number or type of curved refracting surfaces. However, it is often simpler to express the properties of the lens in another way. When dealing with the image-forming properties of a thin lens, such as is illustrated in Fig. 4, i.e., a refractive medium bounded by two spherical surfaces having radii

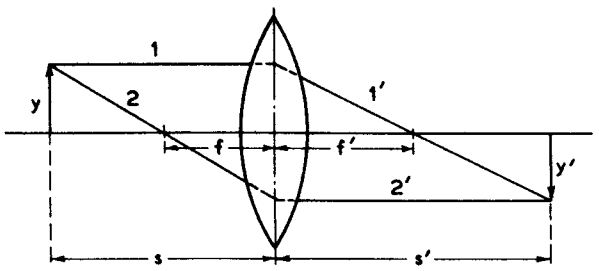


Fig. 4 - Properties of a thin lens.

of curvature large compared to the separation between them, it is most convenient to make use of the derived concept of focal length.

Since, for this lens, the index of refraction in the image and object space is the same, it follows from the successive application of equation (4) that

$$\frac{1}{s} + \frac{1}{s'} = \text{constant}$$

If, as for ray 1, s is made equal to infinity, this becomes:

$$\frac{1}{\infty} + \frac{1}{f'} = \text{constant}$$

or, similarly, for ray 2', where s' is infinite:

$$\frac{1}{f} + \frac{1}{\infty} = \text{constant}$$

When combined, these equations give

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} = \frac{1}{f'} \qquad (8)$$

which is the familiar equation for a thin lens where $f = f'$ is by definition the focal length. It will be noted that the focal length of the system represented by equation (7) is $1/k$. The magnification, as is evident from the geometry of the figure, will be

$$m = -\frac{s'}{s} \qquad (9)$$

THE THICK LENS

The relations developed above can be applied with a fair degree of accuracy for all lenses whose focal length is long compared to the thickness. However, if the thickness of the lens cannot be considered as negligible compared to the focal length, a more complicated set of relations is required. These apply to a single lens or to a system composed of several lens elements. Either system is spoken of as a "thick lens."

In order to determine the size and position of the Gaussian image formed by a thick lens, it is necessary to locate two reference planes known as "principal" planes, and the two focal points.

The two principal planes are the conjugate planes for which the optical system has a positive magnification of unity. In other words, an object at one principal plane produces at the other a virtual (erect) image which is the same size as the object. The intersections of these planes with the axis of the system are known as the principal points. Fig. 5 shows a thick-lens optical system, the planes MN and M'N' being the

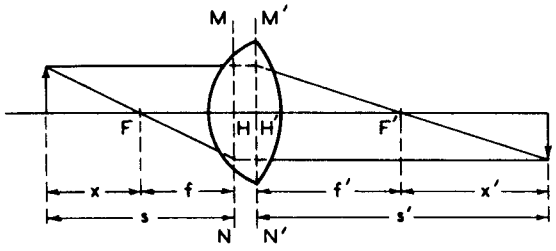


Fig. 5 - Image formation by a thick lens.

two principal planes, and H and H' the principal points. Any ray of light entering the lens from the object space parallel to the axis will be bent in such a way that it crosses the axis at the point F'. The parallel ray, and the ray through the focal point, will, if extended, meet

at the principal plane of the image space M'N'. Similarly, any ray through the point F in the object space will emerge from the lens system parallel to the axis. These two rays, when extended, meet in the first principal plane MN. The points F and F' are the first and second focal points, respectively. The two focal points, together with the two principal points, are known as the cardinal points of the lens system. Once these have been located, the first-order image of any object can readily be found. The distances f and f', between the focal points and their corresponding principal planes are known as the first and second focal lengths.

From the geometry of Fig. 5, it is evident that the lateral magnification m of the system is given by

$$m = -\frac{f}{x} = -\frac{x'}{f'} \qquad (10)$$

where x and x' are object and image distances from the focal points. From this it follows that

$$xx' = ff' \qquad (11)$$

The magnification and position can, of course, be referred to the principal points instead of the focal points. If s and s' are the distances of the object and image from the principal planes, their values are $s = x + f$ and $s' = f' + x'$. When these are substituted in equation (11), the latter becomes:

$$(s - f)(s' - f') = ff' \qquad (12)$$

$$\frac{f}{s} + \frac{f'}{s'} = 1$$

which corresponds to equation (8) for a thin lens. In the same terms, the magnification is

$$m = -\frac{f}{s - f} = -\frac{s' - f'}{f'} \qquad (13)$$

It can be shown that, if the medium in the object space has a refractive index n and that in the image space is n', the two focal lengths will be related as follows:

$$\frac{f}{f'} = \frac{n}{n'} \qquad (14)$$

As a consequence, if the media on the two sides of the system have the same refractive index, the first and second focal lengths will be equal. A derivation of these laws dealing with a thick lens can be found in any elementary textbook on optics.

As was mentioned above, any combination of thin lenses may be represented by the four cardinal points of a thick lens. Two thin lenses of focal lengths f_1 and f_2 separated by a distance d will serve as an illustrative example. Such a system is shown in Fig. 6. In the following, all quantities referring to the first lens will have the subscript 1; those to the

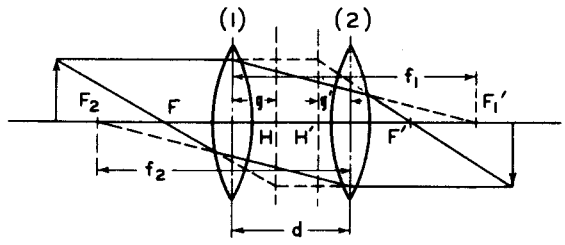


Fig. 6 - Cardinal points of a pair of thin lenses.

second lens, the subscript 2; those referring to the equivalent thick lens will have no subscript. An object at infinity, making $s_1 = s = \infty$, is imaged at F_1' by the first lens. This image is the object for the second lens, the object distance being

$$s_2 = d - f_1$$

The point at which the second lens images this virtual object will be the position of second focal point F' of the equivalent thick lens. The distance between F' and the second lens is, therefore, s_2'. Applying equation (12), it follows that

$$\left.\begin{aligned} \frac{1}{d - f_1} + \frac{1}{s_2'} &= \frac{1}{f_2} \\ s_2' &= \frac{f_1 f_2 - d f_2}{f_1 + f_2 - d} \end{aligned}\right\} \qquad (15)$$

which locates the second focal point. The first focal point can be located in a similar way by tracing a ray to an image point at infinity. The distance between the first lens and the first focal point will be found to be

$$\frac{f_1 f_2 - d f_1}{f_1 + f_2 - d} \qquad (15a)$$

Furthermore, from the geometry of the two sets of ray paths discussed, it can be shown that the distances g and g' between the first principal point and first lens, and between the second principal point and second lens, are:

$$g = \frac{d f_1}{f_1 + f_2 - d} \qquad (16)$$

$$g' = \frac{d f_2}{f_1 + f_2 - d} \qquad (16a)$$

In this way, the four cardinal points of the equivalent thick lens can be located.

INDEX OF REFRACTION IN ELECTRON OPTICS

From the foregoing it is evident that the concept of the index of refraction is important in optics. In electron optics the potential, or rather the square root of the potential, plays the role of the index of refraction. The following simple example illustrates the similarity. Consider an electron moving from a region at potential φ_1 through a narrow transition region into a region at potential φ_2. Referring to Fig. 7, it will be seen that the regions A and C are field-free and at potentials φ_1 and

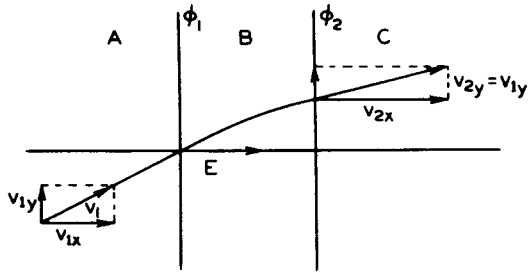


Fig. 7 - Refraction by potential double layer.

φ_2, respectively; while in the intervening transition space B, there exists a field of magnitude $(\varphi_2 - \varphi_1)/d$ normal to the boundary sheets. The boundaries separating the three regions are assumed to be conducting sheets which are transparent to electrons. Such sheets are, of course, purely fictitious and can only be approximated in practice; however, they are useful for illustrative purposes. The incident electron approaching the first sheet has a velocity

$$v_1 = \sqrt{2\frac{e}{m}\varphi_1}$$

If it makes an angle θ_1 with the surface normal, its velocity can be resolved into two components, which are:

$$v_{1x} = \sqrt{2\frac{e}{m}\varphi_1}\cos\theta_1 \qquad (17)$$

$$v_{1y} = \sqrt{2\frac{e}{m}\varphi_1}\sin\theta_1 \qquad (17a)$$

where x is normal to the boundary sheets and y is parallel to them. As the electron traverses the transition region, it is accelerated in the x direction by the field, while its y component of velocity remains unchanged. The total velocity of the emerging electron will, of course, be

$$v_2 = \sqrt{2\frac{e}{m}\varphi_2} \qquad (18)$$

From equation (17a) its transverse velocity is

$$v_{2y} = \sqrt{2\frac{e}{m}\varphi_1}\sin\theta_1 \qquad (19)$$

The angle of emergence θ_2 is determined by these two velocities and is given by

$$\frac{v_{2y}}{v_2} = \sin\theta_2$$

Finally, substituting equations (18) and (19) into this equation and transposing, it follows that

$$\sqrt{\varphi_1}\sin\theta_1 = \sqrt{\varphi_2}\sin\theta_2 \qquad (20)$$

Comparing this to Snell's law, it will be seen that $\sqrt{\varphi}$ is the exact counterpart of the index of refraction n.

This similarity is quite general, as can be proved by comparing Fermat's principle of optics:

$$\delta\int_A^B n\,ds = 0$$

and the principle of least action for an electron:

$$\delta\int_A^B v\,ds = \text{constant}\;\delta\int_A^B \sqrt{\varphi}\,ds = 0$$

SIMPLE DOUBLE-LAYER LENS

The simplest concept of an electron lens is probably that formed by two curved double layers of the type just described. The arrangement is illustrated in Fig. 8. The double layer on the object side is assumed to have a radius of curvature R_1 and that on the image side a radius R_2. The potential of the inner surfaces is φ_2; that of the outer, φ_1.

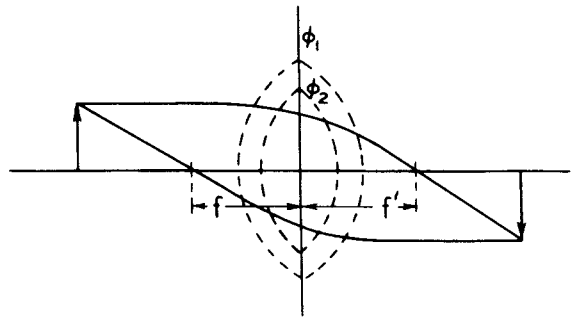


Fig. 8 - Doubler-layer lens.

This lens corresponds exactly to an optical thin lens with an index of refraction $n_2 = \sqrt{\varphi_2}$ immersed in a medium which has an index of re-

fraction $n_1 = \sqrt{\varphi_1}$. Its focal length can be calculated by elementary optics and is given by the equation:

$$\frac{1}{f} = \left(\sqrt{\frac{\varphi_2}{\varphi_1}} - 1\right)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

An electron source, located at a distance s from this lens, emitting electrons with a velocity $\sqrt{2(e/m)\varphi_1}$ will be imaged at a distance s'. These distances are related by the equation

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

CONTINUOUS LENSES

If the type of lens just described were the only kind that could be made, electron lenses would have little practical value. There exists a second, fundamentally different class, and upon this the importance of electron lenses rests. The operation of these lenses is based upon the following fact: Whenever there exists a region in which there is a varying electric field having cylindrical symmetry, this region will have properties analogous to those of an optical lens system; that is, it will be capable of forming a real or virtual image of an emitting source.

It is evident that a lens formed in this way is very different from the familiar glass lens treated in conventional optics. Instead of sharp boundaries between media of different index of refraction, the index of refraction varies continuously, both along the axis and radially. This being so, it has been found necessary to apply different mathematical methods in calculating the lens properties resulting from any specific field. These properties, once they have been found, can be represented by the same four cardinal points which describe an optical thick lens, namely, by two focal points and two principal points.

The simplest lens having a continuously variable index of refraction is that formed by an axially symmetric transition region between two constant fields of different magnitude. Physically, such a region is approximated by that in an aperture having different field strengths on the two sides. This is illustrated in Fig. 9a.

In order to obtain a physical picture of the action of this lens, it is convenient to make use of the concept of lines of electrostatic force, or field lines, whose direction at any point is that of the field and whose density is proportional to the field strength. A small volume element in the transition region of the lens is shown in Fig. 9b. In this region the Laplace equation is obeyed and, as was pointed out in the preceding lecture, this requires that the number of field lines which enter the volume element must equal that which leave. It is evident that, since the density of lines is different on the two ends, lines must enter or leave through the top of the volume. Referring to Fig. 9b, it will be seen that this condition can be fulfilled only if the field lines are curved. Accordingly, the field lines are not parallel to the axis, and

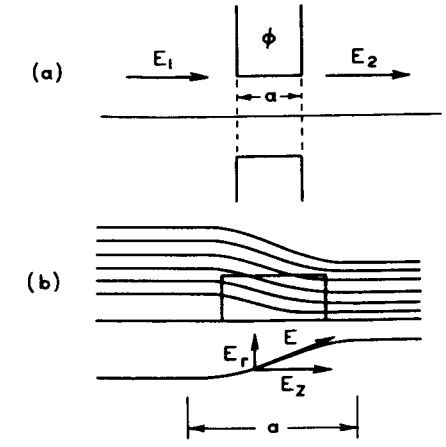


Fig. 9 - Idealized field distribution in simple aperture lens.

the field can be resolved into radial and axial components. The radial component of the field will deflect any electron passing through the transition region toward (or away from) the axis. This bending action increases with radial distance from the axis, as is required for image formation.

The quantitative properties of the lens can be determined as follows:[1] First, by applying the Laplace equation the radial component of the field is obtained. From the radial field component the change in radial momentum of an electron in passing through this region is calculated. This leads to the value of the change in angle of the electron trajectory and, hence, directly to the focal length of the system.

Since there are no charges, the divergence as well as the curl of the field are zero. In cylindrical coordinates these can be expressed as follows:

$$\frac{\partial E_z}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(rE_r) = 0 \qquad (21)$$

and

$$\frac{\partial}{\partial z}(rE_r) - r\frac{\partial E_z}{\partial r} = 0 \qquad (22)$$

The total differential of the radial field component is

$$d(rE_r) = \frac{\partial}{\partial z}(rE_r)\,dz + \frac{\partial}{\partial r}(rE_r)\,dr$$

[1] L.H. Bedford, "Electron Lens Formulas," Phys. Soc. Proc., Vol. 46, pp. 882-888; 1934.

Substituting from equations (21) and (22), it follows that

$$d\,(rE_r) = r\,\frac{\partial E_z}{\partial r}\,dz - r\,\frac{\partial E_z}{\partial z}\,dr \qquad (23)$$

which gives the radial component in terms of the axial field. To determine the first-order-image properties of the lens, only the field near the axis need be considered. In the vicinity of the axis both E_z and $\partial E_z/\partial z$ are to the desired approximation independent of r. Therefore, equation (23) may be written:

$$d\,(rE_r) = -\,r\,\frac{\partial E_z}{\partial z}\,dr$$

and integrated directly to give:

$$E_r = -\,\frac{r}{2}\,\frac{\partial E_z}{\partial z} \qquad (24)$$

The change in radial momentum of an electron subjected to this field for a time equal to that required to traverse the transition region, whose width is a, is:

$$\Delta\, mv_r = -\int eE_r\,dt = \int_0^a \frac{er}{2}\,\frac{\partial E_z}{\partial z}\,\frac{dz}{v_z} \qquad (25)$$

since $dt = dz/v_z$. As a is small, v_z can be assumed to be constant over the transition region, and this equation can be integrated, giving:

$$\Delta\, mv_r = \frac{er}{2v_z}(E_2 - E_1)$$

The change in angle is, therefore:

$$\alpha = \frac{\Delta\, mv_r}{mv_z} = \frac{er}{2v_z^2 m}(E_2 - E_1)$$

or replacing mv_z^2 by $2e\varphi$, where φ is the potential of the lens:

$$\alpha = r\,\frac{E_2 - E_1}{4\varphi} \qquad (26)$$

A comparison of equation (26) with equation (7) indicates that the lens will form a first-order image. Its focal length is

$$f = \frac{4\varphi}{E_1 - E_2} \qquad (27)$$

Although, in view of the restrictions assumed in the derivation, equation (27) is only a first approximation, it is nevertheless useful for estimating the behavior of apertures and other electron-optical systems. For example, a lens often encountered in practice consists of two apertures at different potentials as illustrated in Fig. 10. The focal lengths of the apertures considered separately are:

$$f_A = \frac{4\varphi_A d}{\varphi_B - \varphi_A}$$

$$f_B = \frac{-4\varphi_B d}{\varphi_B - \varphi_A}$$

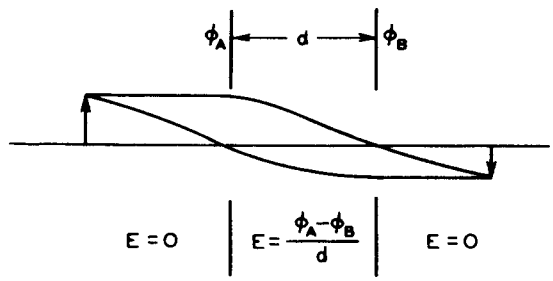


Fig. 10 - Ray paths in double-aperture lens (schematic).

When these are combined into a "thick lens", due account being taken of the converging action of the field between the apertures and the difference in potential on the two sides of the lens, the focal lengths f and f' on the object and image sides of the system are:

$$\left.\begin{aligned} f &= \frac{8d}{3\left(\frac{\varphi_B}{\varphi_A} - 1\right)\left(1 - \sqrt{\frac{\varphi_A}{\varphi_B}}\right)} \\ f' &= \frac{8d}{3\sqrt{\frac{\varphi_A}{\varphi_B}}\left(\frac{\varphi_B}{\varphi_A} - 1\right)\left(1 - \sqrt{\frac{\varphi_A}{\varphi_B}}\right)} \end{aligned}\right\} \qquad (28)$$

Again, warning should be given that these results are only approximate, and strictly applicable only when φ_A and φ_B are not too different, d is small compared with the focal length, and the aperture diameter small compared with d.

THE RAY EQUATION

In most practical cases, the lens region,

i.e., the region throughout which the field is varying, cannot be considered as small, as was assumed in the derivation given in the preceding section. To determine the imaging properties of a more general electron lens, it is necessary to make use of a differential equation of the ray path, which can be integrated over the region of

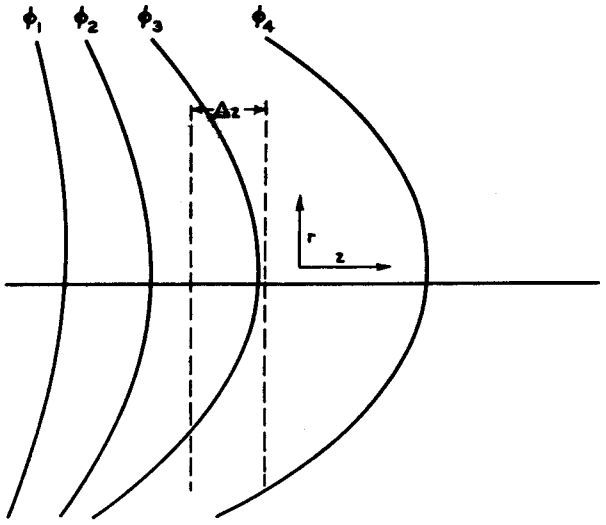


Fig. 11 - Elementary section of the potential field of an electron lens.

varying potential. Proceeding as in the previous example, but considering as the transition region an elementary strip between z and $z + \Delta z$, which is part of an extended region of varying potential, the change in radial momentum is

$$\Delta\, mv_r = \frac{re}{2} \frac{\partial E_z}{\partial z} \frac{\Delta z}{v_z}$$

Making the substitutions

$$v_r = \frac{dr}{dt} \quad \text{and} \quad \frac{\partial E_z}{\partial z} \cong - \frac{d^2\varphi}{dz^2}$$

it is found that

$$\Delta \frac{dr}{dt} = \frac{re}{2m} \frac{d^2\varphi}{dz^2} \frac{\Delta z}{v_z}$$

φ being the potential along the axis. Dividing through by Δz and letting the width of the Δz elementary strip approach zero, it follows that

$$\frac{d}{dz}\left(\frac{dr}{dt}\right) = - \frac{er}{2mv_z} \frac{d^2\varphi}{dz^2} \qquad (29)$$

By a suitable rearrangement of terms and the substitution

$$v_z \cong v = \sqrt{\frac{2e}{m}\varphi}$$

equation (29) can be written as

$$\frac{d^2r}{dz^2} = - \frac{1}{2\varphi} \frac{d\varphi}{dz} \frac{dr}{dz} - \frac{1}{4\varphi} \frac{d^2\varphi}{dz^2} r \qquad (30)$$

$$r'' = - \frac{r'\varphi'}{2\varphi} - \frac{r\varphi''}{4\varphi}$$

This is the fundamental ray equation used in all cylindrically symmetric electron-optical problems involving only electrostatic fields. When solved it gives the radial distance of the electron ray from the axis at every point along the axis. Furthermore, its form assures the fulfillment of Gaussian, or first-order, image requirements.

SOLUTION OF THE RAY EQUATION

When a solution of equation (30) is possible, the first-order-image properties of the electron lens system can be readily determined. The procedure is similar to that in optics. An electron ray parallel to the axis in object space is traced through the lens. Its intersection with the axis locates the second focal point. Extending a line tangent to the ray at the second focal point until it intersects the incident ray locates the position of the second principal plane. The first focal point and principal plane can be determined by similarly considering a ray parallel to the axis in image space. With the four cardinal points, the position and magnification of the image corresponding to an arbitrarily placed object can be readily calculated. However, it must be remembered that, in order to determine the cardinal points in this way, not only must both the image and object be outside of the lens (i.e., in field-free space) but the same must be true of the focal points.

An alternative method of determining the imaging properties of a lens is to trace through the system, first, a ray having a small radial initial velocity leaving an object point on the axis which is followed until it intersects the axis, thus locating the image plane, and second, a ray from the object leaving with zero radial initial velocity from a point off the axis whose intersection with the image plane will give the magnification. With the aid of these two ray paths the four cardinal points may also be determined. If the image and object are not in the lens, the cardinal points determined by this second procedure can be used irrespective of whether or not they fall within the lens. If the image or object lies within the lens, the calculation applies to that particular object or image position only.

The differential ray equation rarely permits an analytical solution in practical cases. Again, as was found to be necessary in the solution of the Laplace equation, numerical approximations or graphical methods must be resorted to.

A very simple and rapid approximate method has been proposed by Richard Gans.[2] With a little care, this method is capable of an accuracy adequate for most practical cases. The method consists of representing the potential along the axis by a series of straight-line segments, and applying the ray equation along the segments in turn. Over any straight segment the second derivative $d^2\varphi/dz^2$ is zero. The ray equation, therefore, becomes:

$$\frac{d^2r}{dz^2} = -\frac{1}{2\varphi}\frac{dr}{dz}\frac{d\varphi}{dz}$$

This can be integrated twice to give:

$$\frac{dr}{dz}\sqrt{\varphi} = C \qquad (31)$$

and

$$r = r_0 + \frac{2C\,(\sqrt{\varphi} - \sqrt{\varphi_0})}{\frac{d\varphi}{dz}} \qquad (31a)$$

where r_0 and φ_0 are the radial position and potential at the beginning of the segment.

At the point where two segments meet, $d^2\varphi/dz^2$ is infinite. Integrating equation (30) over the transition gives the equation:

$$\left(\frac{dr}{dz}\right)_2 - \left(\frac{dr}{dz}\right)_1 = -r\,\frac{\left(\frac{\partial\varphi}{\partial z}\right)_2 - \left(\frac{\partial\varphi}{\partial z}\right)_1}{4\varphi} \qquad (32)$$

The subscript 1 indicates values before the break-point, and 2, those after the break-point.

Finally, where the segment is parallel to the axis, the solution becomes

$$r = r_0 + \left(\frac{dr}{dz}\right)_0 (z - z_0) \qquad (33)$$

The method, because of its utility as a practical means for estimating the performance of any electron lens system for which the axial distribution is known, is worth illustrating by means of a simple numerical example.

The lens system to be considered consists of two equidiameter coaxial cylinders. The cylinders are assumed to have potentials equal to 2 and 12, respectively. Here the units of potential are quite arbitrary, the only thing of importance, as far as the lens properties are concerned, is the voltage ratio, which in this example is 6. The axial potential distribution of the system has been calculated by means of equation (35a) given in the next section, and is tabulated in Table I, in which the distance z is given in cylinder radii.

Table I

z	φ	z	φ
-2.0	2.05	0.1	7.66
-1.8	2.10	0.3	8.88
-1.6	2.16	0.5	9.88
-1.4	2.27	0.7	10.62
-1.2	2.44	0.9	11.12
-1.0	2.70	1.1	11.45
-0.8	3.12	1.3	11.66
-0.6	3.72	1.5	11.79
-0.4	4.58	1.7	11.87
-0.2	5.71	1.9	11.93
0.0	7.00		

It should be pointed out that the potential distribution for any desired voltage ratio can be obtained with the aid of the values given in Table I by adding an appropriate constant.

The potential distribution is shown by the dotted curve of Fig. 12. In addition, this distribution is approximated in the figure by a series of six straight-line segments intersecting at points A, B, C, D, and E. The intersections have the following values:

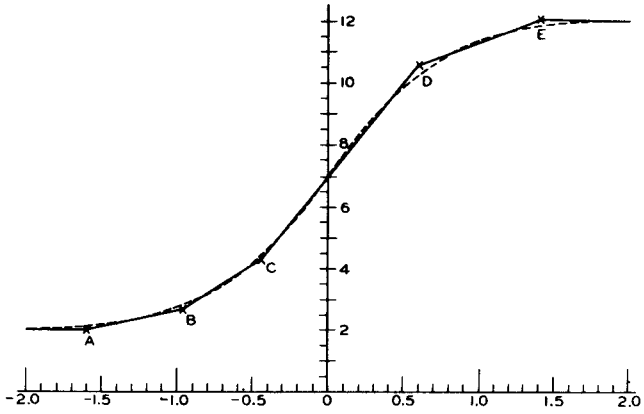


Fig. 12 - Axial potential distribution and its line segment approximation for a cylinder lens.

Point	z	φ
A	-1.60	2.0
B	-0.96	2.7
C	-0.46	4.3
D	0.60	10.6
E	1.40	12.0

The slopes of the segments, and consequently the potential gradients, φ', represented by them, are:

[2] R. Gans, "Electron Paths in Electron Optics," Z. tech. Physik, Vol. 18, pp. 41-48; 1937.

Segment	φ'
∞ A	0
A B	1.09
B C	3.20
C D	5.95
D E	1.75
E ∞	0

An electron leaves an object point on the axis at a distance -12 lens radii from the origin, with velocity corresponding to $\varphi = 2$ and a slope of 0.096. It is required to find the conjugate image point, that is, the point at which the electron again intersects the axis.

The electron will move in a straight line until it reaches the first break-point at A. At this point, its radial distance is $r_A = 1.00$ and its slope $r_{1A}' = 0.096$. Its slope at the other side of the break-point is given by equation (32):

$$r_{2_A}' = 0.096 - \frac{1.00 \times (1.09 - 0)}{4 \times 2.0}$$

$$= - 0.040$$

The value for C for the segment AB, given by equation (31), is

$$C = - 0.040 \times \sqrt{2.0} = - 0.057$$

Therefore, by equation (31):

$$r_{1_B}' = \frac{- 0.057}{\sqrt{2.7}} = - 0.034$$

and from equation (31a):

$$r_B = 1.00 + \frac{- 0.114 \ (\sqrt{2.7} - \sqrt{2.0})}{1.09}$$

$$= 0.97$$

The calculation continues in the same way for each succeeding break-point and segment. The results are given in Table II.

Table II

Point	r	r_1'	r_2'	C
A	1.00	0.096	-0.040	-0.057
B	0.97	-0.034	-0.022	-0.366
C	0.88	-0.18	-0.32	-0.657
D	0.62	-0.20	-0.14	-0.459
E	0.51	-0.13	-0.11	-

From point E the electron continues in a straight line, with a slope to the axis of -0.11. Therefore, it intersects the axis at a distance 0.51/0.11 = 4.6 radii from point E, or, in other words, the image point is 18.0 lens radii from the object point. A determination of the path by the more exact methods outlined in the next section gives the distance between the object and image point as 18.6 lens radii. Fig. 13 shows the exact path of the electron through the system, together with the points corresponding to the radial position at the break-point as calculated by the approximate method.

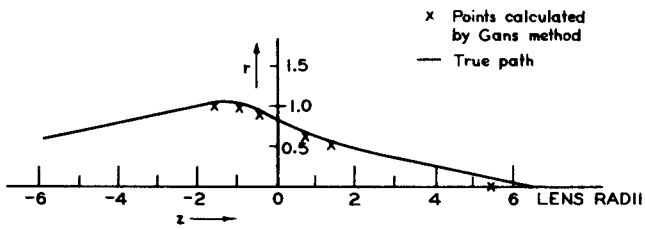


Fig. 13 - Electron trajectory through cylinder lens.

A closer approximation, of course, could have been obtained if a greater number of line segments had been used to represent the potential, particularly on the low-potential side of the lens.

SPECIAL LENS SYSTEMS

The electron-optical system consisting of two coaxial cylinders of equal diameters forms the basis of many practical lenses. The properties of this configuration can, of course, be determined by mapping the axial potential distribution with the aid of a plotting tank and then tracing suitable rays by the method just discussed. However, when high accuracy is required, a mathematical solution may be advantageous.

Fig. 14 illustrates, in cross-section, two semi-infinite coaxial cylinders spaced a negligible distance apart. The axis of the cylinders will be taken as the z-axis of the cylindrical coordinate system and the origin of the system will be located at the junction of the two cylinders.

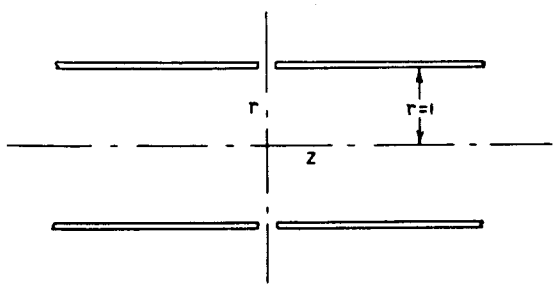


Fig. 14 - Coordinates for electrostatic cylinder lens.

The general solution of the Laplace equation for this type of configuration was discussed in the preceding lecture. By separating the vari-

ables, it was shown that the solution could be expressed in the form of the following integral:

$$\varphi\ (r,z) = \int A(k)\ G(r,k)\ F(z,k)\ dk$$

where

$$F = a\ e^{iks} + b\ e^{-iks}$$

$$G = c\ J_0(ikr) + d\ N_0(ikr)$$

The requirement that the potential remain finite as z increases eliminates all terms with complex k. Furthermore, the condition that it be finite along the axis requires that the coefficients of the Neumann function be zero. Finally, since

$$\varphi = \frac{(\varphi_1 + \varphi_2)}{2}$$

is an odd function of z, the only trigonometric functions to be considered are sines. Hence, the solution can be written as:

$$\varphi(r,z) = \int_0^\infty B(k)\ J_0(ikr)\ \sin kz\ dk + \frac{\varphi_1 + \varphi_2}{2} \qquad (34)$$

The coefficient B(k) can be found with the aid of boundary conditions for r = 1, the radius of the cylinders being taken as unit length. These are:

$$\varphi\ (1,z) = \varphi_1 \text{ for } z < 0$$

$$\varphi\ (1,z) = \varphi_2 \text{ for } z > 0$$

This evaluation leads to:

$$B(k)J_0\ (ik) = \frac{1}{\pi}\left[\int_{-\infty}^{0} -\frac{(\varphi_2 - \varphi_1)}{2}\sin kz\ dz + \int_0^\infty \frac{(\varphi_2 - \varphi_1)}{2}\sin kz\ dz\right]$$

$$B(k) = \frac{(\varphi_2 - \varphi_1)}{\pi k J_0(ik)}\left(1 - \lim_{z\to\infty} \cos kz\right)$$

When this value is placed in equation (34), the potential is found to be:

$$\varphi(r,z) = \frac{1}{\pi}\int_0^\infty \frac{\varphi_2 - \varphi_1}{kJ_0(ik)}\ J_0(ikr)\ \sin kz\ dk + \frac{\varphi_1 + \varphi_2}{2}$$

The axial potential and its first two derivatives are:

$$\varphi(0,z) = \frac{1}{\pi}\int_0^\infty (\varphi_2 - \varphi_1)\ \frac{\sin kz}{kJ_0(ik)}\,dk + \frac{\varphi_1 + \varphi_2}{2} \qquad (35a)$$

$$\varphi'(0,z) = \frac{1}{\pi}\int_0^\infty (\varphi_2 - \varphi_1)\ \frac{\cos kz}{J_0(ik)}\,dk \qquad (35b)$$

$$\varphi''(0,z) = -\frac{1}{\pi}\int_0^\infty (\varphi_2 - \varphi_1)\ \frac{k \sin kz}{J_0(ik)}\,dk \qquad (35c)$$

These integrals cannot be solved analytically, but must be evaluated by quadrature, which, though laborious, is a fairly straightforward procedure. Numerical values of potential thus obtained are used to calculate the electron paths from the ray equation.

The solution of the ray equation is expedited by the substitution:

$$c = -\frac{1}{r}\frac{dr}{dz}$$

The function c is the convergence of the ray. It is thus named because it is the reciprocal of the distance along the axis from the point at which c is determined to the intersection of a rectilinear extension of the ray with the axis.

When this substitution is made, the ray equation (30) becomes:

$$\frac{dc}{dz} = c^2 - \frac{c\varphi'}{2\varphi} + \frac{\varphi''}{4\varphi} \qquad (36)$$

The evaluation of the equation must be continued until the ray is substantially a straight line, or until it passes through the image plane.

The curves reproduced in Fig. 15 locate completely the four cardinal points for various values $\varphi 1/\varphi 2$ in the type of system just described. All distances are given in terms of lens radii.

Besides the coaxial cylinder lens using cylinders of the same diameters, an infinite number of other lenses can, of course, be formed by us-

ing cylinders of unequal diameters. The effect of this difference is to alter to some extent the position of the cardinal points. From the standpoint of Gaussian dioptrics, there is rarely any advantage in choosing cylinders of different diameters. Empirical data, however, indicate that certain of the image defects can be considerably reduced by a proper selection of diameter ratios.

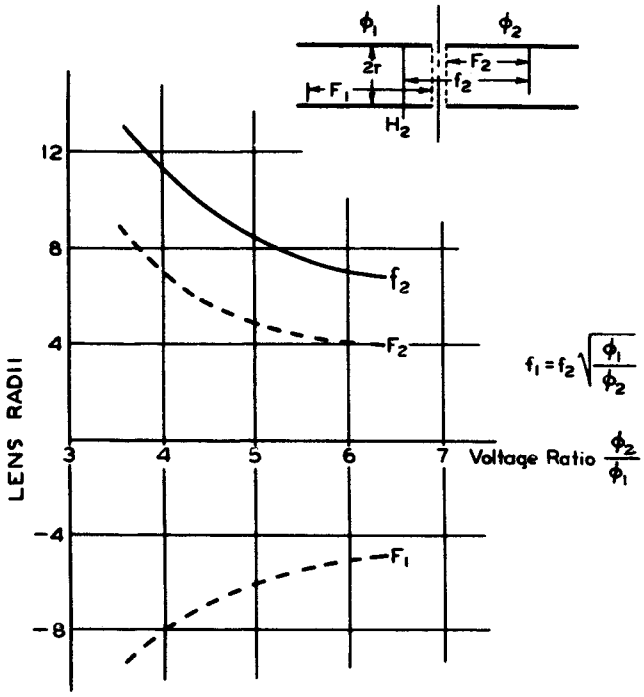


Fig. 15 - Focal length and position of focal points of cylinder as function of voltage ratio (after Epstein, reference 16).

APERTURE LENSES

Another very important class of lenses are those formed by a pair of apertured conductors at different potentials. Lenses of this class are used extensively both in the electron gun and in electron imaging devices such as the microscope.

The simplest lens of this type has already been discussed. The results obtained for this restricted case, however, have very limited application. They can be used only where the potential difference of the apertures is small so that the focal lengths are large compared to the spacing between the apertures and to their diameters.

A determination of the first-order-image properties of aperature lenses has been made by Polotovski by electrolytic and graphical methods. Fig. 16 reproduces some of his results. These curves give the second focal length for various voltages, and separation-to-diameter ratios. For comparison, equation (28) for the simplified case is included in this figure.

CATHODE LENS SYSTEMS

In many practical electron-optical lenses, the electrons enter the system with essentially zero velocity. Such a lens is to be found in the electrostatically focused image tube. This type of lens, shown in Fig. 17, differs quite radically in its properties from the lenses discussed so far.

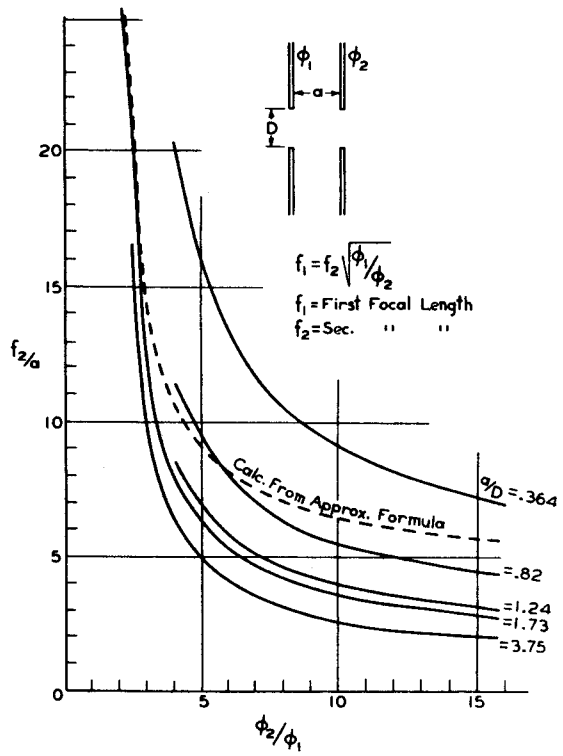


Fig. 16 - Focal length of double-aperture lenses (after Polotovski, reference 15).

The potential at the point where the electron enters the system is zero. In consequence, the electron ray originates in a region of zero index of refraction. Because of this condition, the first focal length is zero and the first principal plane is located at the object.

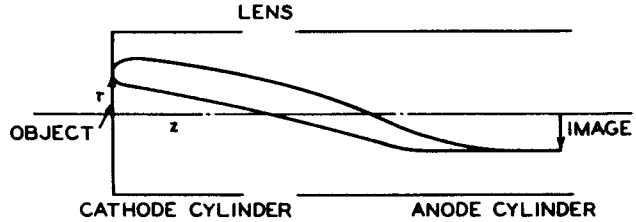


Fig. 17 - Cathode lens.

Under these conditions, if the electrons were actually emitted from the object, that is, the cathode, with zero velocity, the image plane

would be completely indeterminate. However, the assumption of a small radial initial velocity suffices to determine the position and, with it, the magnification of the image. In practice, both the radial and axial components of the initial velocity are far from negligible, and are responsible for certain aberrations to be discussed in the following lecture.

One of the simplest lenses of this type consists of two coaxial cylinders of equal diameter, one infinite in extent, the other terminated by a flat plane. The terminal plane is assumed to be the electron object. In a conventional image tube, for example, this plane would be a photoelectric cathode. The cathode surface is conducting and is electrically connected to the cathode cylinder, these elements being at zero potential. For convenience, the lens will be said to be located at the junction of the two cylinders, although actually it extends from the cathode to a distance of several cylinder radii beyond the junction. The geometry of the lens can be seen from Fig. 17.

The calculation of the potential distribution proceeds exactly as in the previous example of the two-cylinder lens. It is expedient here to use cylindrical coordinates with their origin at the cathode. The general solution of the Laplace equation, which can be evaluated from the boundary conditions of the lens, leads to the following integrals which give the potential along the axis together with the first and second derivatives:

$$\varphi(0,z) = \frac{2}{\pi}\int_0^\infty \varphi_1 \frac{\cos ku \ \sin kz}{k\, J_0\,(ik)}\, dk \qquad (37a)$$

$$\varphi'(0,z) = \frac{2}{\pi}\int_0^\infty \varphi_1 \frac{\cos ku \ \cos kz}{J_0\,(ik)}\, dk \qquad (37b)$$

$$\varphi''(0,z) = -\frac{2}{\pi}\int_0^\infty \varphi_1 \frac{k \cos ku \ \sin kz}{J_0\,(ik)}\, dk \qquad (37c)$$

As before, these integrals must be evaluated by quadrature.

The numerical values of the potential thus obtained can, with the aid of the ray equation, be used to obtain the lens properties. In this instance, the convergence function, c, becomes infinite at z = 0 and thus cannot be used. However, substitution of the function:

$$b = \frac{1}{2z} - \frac{1}{r}\frac{dr}{dz}$$

leads to:

$$\frac{db}{dz} = b^2 - b\left(\frac{1}{z} + \frac{\varphi'}{2\varphi}\right) + \frac{\varphi''}{4\varphi} + \frac{1}{2z}\left(\frac{\varphi'}{2\varphi} - \frac{1}{2z}\right)$$

and thus facilitates the solution by reducing the ray equation to a first-order differential equation.

Numerical determinations of two electron paths will locate the image and give its magnification. In Fig. 18 the potential and its first two derivatives are shown, together with two electron

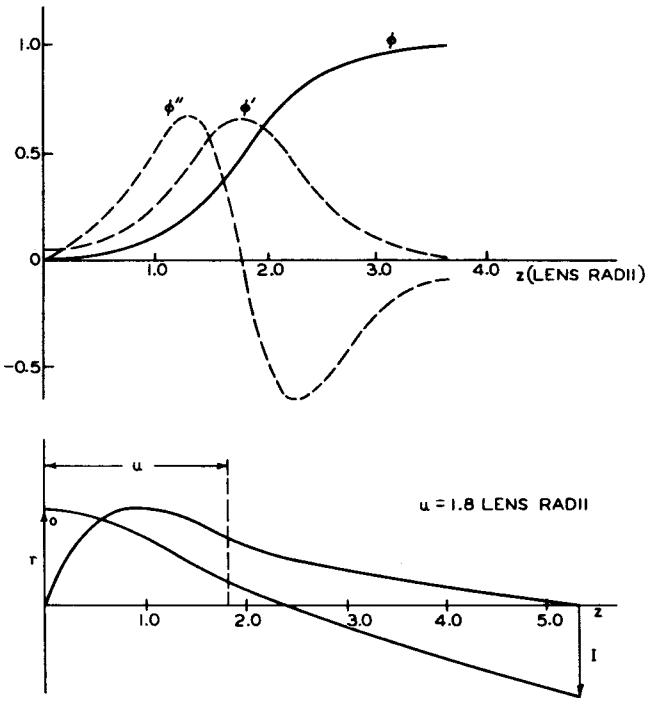


Fig. 18 - Axial potential distribution and electron trajectories in a cathode lens.

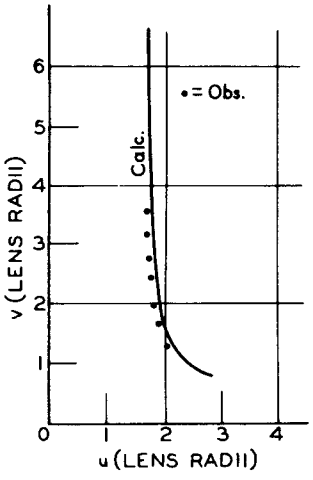
trajectories. It should be noted that the latter are independent of the applied potential, φ_1, and are a function of the cathode-to-lens distance, u, only. The relations between image distance, object distance, and magnification are given in Fig. 19a.

It is often desirable to arrange the lens in such a way that the focal length and image position can be varied electrically. This can be done by making the cathode cylinder of resistive material so that an axial potential gradient can be established on it between cathode and lens. In practice, it is convenient to divide the cathode cylinder into a number of rings having equal potential steps between them. Experimentally, this is found to simulate the resistive cathode cylinder qualitatively and quantatively.

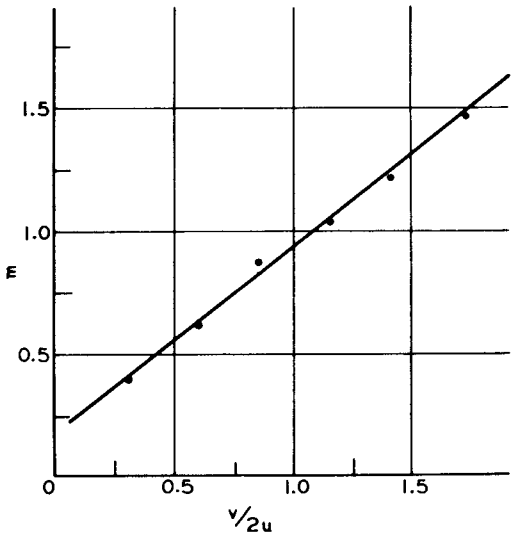
The performance of this system can be calculated by the methods already given. The results of such a determination are of value in many design problems and are shown in the two curves in Fig. 19b.

The magnification of the system just described is a function of the image and object distance alone and cannot be changed in a given tube configuration. It is possible, by incorporating a third element, to make a lens system which has a variable magnification. The additional electrode may be in the form of an aperture or a short cylinder placed between the anode and cathode cylinders. An electron optical system of this type is illustrated in Fig. 19c. For a given tube structure and overall voltage, the magnification is varied by changing the potential φ_3 on the extra electrode and refocusing the image by the potential φ_2 along the cathode cylinder.

Before leaving the subject of cathode lens systems, some mention should be made of the pro-

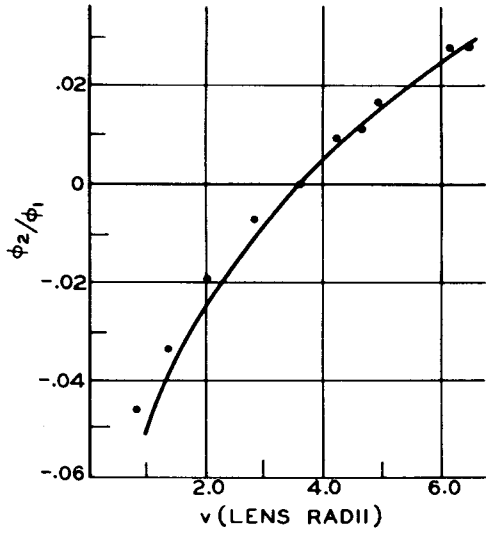


Focusing Relation
(Fixed Focus)

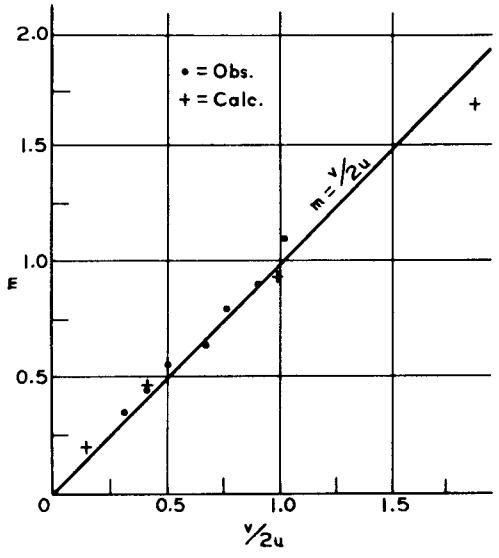


Magnification
(Fixed Focus)

Fig. 19a



Focusing Relation
(Variable Focus)



Magnification
(Variable Focus)

Fig. 19b

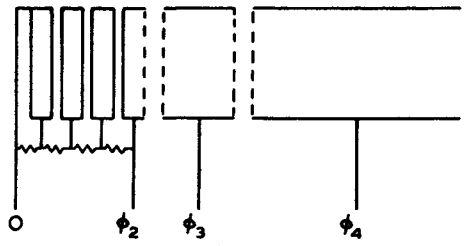


Fig. 19c - An electron lens system arranged to give variable magnification.

cedure required when they are to be investigated by the approximate method described in section on Solution of the Ray Equation. In order to determine the image position, it is necessary to assume an initial radial velocity. In Fig. 20, the axial potential distribution of a cathode

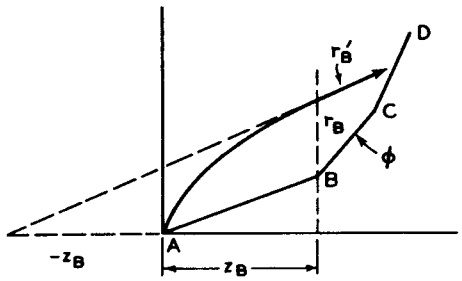


Fig. 20 - Determination of electron trajectory near cathode in Gans method of ray tracing.

lens system is represented by a series of straight-line segments. The electron is emitted from the object point, A, with a small radial initial velocity, into a region of uniform field as represented by the first segment, AB. The trajectory of the electron is a parabola. If at the end of the first segment its radial distance from the axis is r_B, then, from the properties of the parabola, the slope of its trajectory will be $r_B/2z_B$ where z_B is its axial position. Therefore, the electron will enter the first breakpoint on a path described as follows:

$$\left.\begin{aligned} \left(\frac{dr}{dz}\right)_1 &= \frac{r_B}{2z_B} \\ r &= r_B \\ \varphi &= \varphi_B \end{aligned}\right\} \qquad (38)$$

From this point on, the procedure is the same as for any lens system. The choice of the radial separation, r_B, is arbitrary, and does not affect the determination of the image position.

THE MAGNETIC LENS

Thus far the discussion has been limited to electrostatic lens systems. Magnetic lenses, however, rank as at least equal in importance. From a theoretical standpoint, they are very much more complicated because the principal electron paths are not confined to a plane.

The simplest magnetic lens consists of a uniform magnetic field parallel to the axis of the system and pervading all the region between anode and cathode.

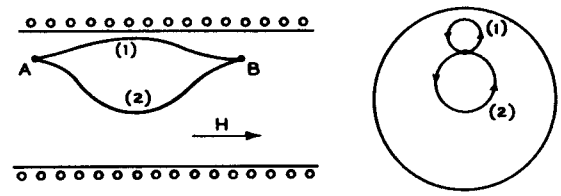


Fig. 21 - Electron paths in a uniform magnetic field.

An electron which leaves the emitter parallel to the axis is subject to no force from the magnetic field. The time required by this electron to traverse the distance between anode and cathode is given by the integral:

$$t_1 = \int_0^L \frac{dz}{v_z} \qquad (39)$$

where L is the total distance and v_z the axial velocity. When, however, the particle leaves with a radial initial velocity, v_r, it will experience a force

$$F = H\ e\ v_r$$

where H is the magnetic field strength. Since the direction of this force is at right angles to the transverse velocity and to the field, the electron path will be curved. The projection of this path on a plane normal to the axis is a circle, the radius ρ being given by the relation between the inward acceleration and the centripetal force, i.e.:

$$\frac{m\ v_r^2}{\rho} = H\ e\ v_r$$

$$\frac{v_r}{\rho} = \frac{e}{m} H$$

The circumference of this circle will be $2\pi\rho$, and, therefore, the time t_2 required to traverse the circle will be:

$$\left.\begin{aligned} t_2 &= \frac{2\pi p}{v_r} \\ &= \frac{2\pi m}{eH} \end{aligned}\right\} \qquad (40)$$

This time is independent of the initial radial velocity. If t_2 in equation (40), or any multiple thereof, is equal to t_1 in equation (39), all electrons leaving a point on the source will come together at a point in the image plane, regardless of their initial radial velocity. It should be noticed that the image, unlike that formed by an electrostatic lens, is erect.

The magnetic lens representing the other extreme is also amenable to simple calculation. This lens is the limiting case of the short lens. Here the magnetic field strength is negligible except over a distance which is small compared with the distance between object and image. The region over which the field is appreciable will be termed the lens. Furthermore, throughout this region it will be assumed that both the potential, φ_o, and the radial distance, r_o, of the ray from the axis are constant. The slope of the ray, dr/dz, of course, changes. On the basis of these assumptions it can be shown that:

$$\left(\frac{dr}{dz}\right)_A - \left(\frac{dr}{dz}\right)_B = \frac{er_o}{8m\varphi_o}\int_A^B H^2\,(z)\,dz \qquad (41)$$

where H(z) is the axial magnetic field, and the slopes with subscripts A and B are those of the ray as it enters and leaves the lens region. If $\underline{s}$ and $\underline{s}'$ are the object and image distances, respectively, for an object point on the axis, it follows that:

$$\left(\frac{dr}{dz}\right)_A - \left(\frac{dr}{dz}\right)_B = \frac{r_o}{s} + \frac{r_o}{s'}$$

Therefore, it follows that:

$$\frac{1}{f} = \frac{e}{8\ m\ \varphi_o}\int_A^B H^2\,(z)\,dz \qquad (42)$$

Its magnification is:

$$m = -\frac{s'}{s}$$

This type of system differs from that consisting of a uniform magnetic field in that the image is inverted.

In general, the practical magnetic lens lies between the two extremes just described. It is capable of forming a real image, and the image will be rotated through an angle whose magnitude depends upon the configuration and magnitude of the field.

Normally in magnetic lenses the variation in electrostatic potential throughout the system is such that its converging action is negligible compared with that due to the magnetic field, and the differential ray equation for a paraxial ray is:

$$\frac{d^2r}{dz^2} = -\frac{e\ H^2}{8\ m\ \varphi}\ r \qquad (43)$$

The image is rotated through an angle θ, where

$$\theta = \sqrt{\frac{e}{8\ m\ \varphi}}\int H\,dz \qquad (43a)$$

An exact solution of equation (43) is difficult and, usually, impossible. An approximate method of solution similar to that used for the electrostatic lens has been proposed by E. G. Ramberg.

To illustrate this method, it is assumed that the axial distribution of a field has been measured or estimated, leading to the smooth curve in Fig. 22. The actual distribution is approximated by a series of step segments as shown in the figure.

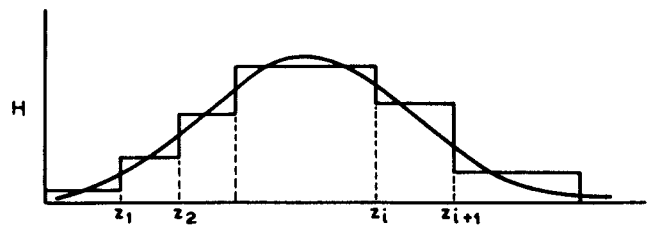


Fig. 22 - Approximate representation of axial magnetic field.

The solution of equation (43) over any segment, for example, that between z_i and z_{i+1}, is

$$r = r_i \cos\sqrt{\frac{e\ H^2}{8\ m\ \varphi}}\ (z - z_i) + \left(\frac{dr}{dz}\right)_i \frac{\sin\sqrt{\frac{e\ H^2}{8\ m\ \varphi}}\ (z - z_i)}{\sqrt{\frac{e\ H^2}{8\ m\ \varphi}}} \qquad (44)$$

where φ is the potential of the segment, H the magnetic field, $\underline{r}$ the radial distance to the ray, and $z_i < z < z_{i+1}$. Both $\underline{r}$ and dr/dz are contin-

uous at the break-points, so that the solutions for the individual segments given by equation (44) join smoothly.

With the aid of the initial conditions, the substituted field distribution, and equation (44), the path of the ray (or, more exactly, the radial displacement), can be traced through the lens just as is done for the electrostatic system. Two ray paths are sufficient to determine the first-order-image properties of the magnetic lens.

Finally, considering the most general case where the lens action is due both to magnetic and electrostatic fields, the differential ray equation becomes:

$$\frac{d^2r}{dz^2} = -\frac{1}{2\varphi}\frac{dr}{dz}\frac{d\varphi}{dz} - r\left(\frac{1}{4\varphi}\frac{d^2\varphi}{dz^2} + \frac{eH^2}{8m\varphi}\right) \tag{45}$$

where the symbols have the previously assigned meanings. The rotation of the image is given by

$$\theta = \sqrt{\frac{e}{8m}}\int\frac{H}{\sqrt{\varphi}}\,dz$$

REFERENCES

1. E. Brüche and O. Scherzer, "Elektronenoptik," Julius Springer, Berlin, 1934.

2. H. Busch and E. Brüche, "Beiträge zur Elektronenoptik," J. A. Barth, Leipzig, 1937.

3. I. G. Maloff and D. W. Epstein, "Electron Optics in Television," McGraw-Hill Book Company, Inc., New York, 1938.

4. H. Busch, "Cathode Ray Paths in Axially Symmetric Fields," Ann. Physik, Vol. 81, pp. 974-993; 1926.

5. M. Knoll and E. Ruska, "Geometric Electron Optics," Ann. Physik, Vol. 12, No. 5, pp. 607-640, 641-661; 1932.

6. J. Picht, "Theory of Geometric Optics for Electrons," Ann. Physik, Vol. 15, No. 8, pp. 926-964; 1932.

7. E. Brüche, "Fundamentals of Geometrical Electron Optics," Z. tech. Physik, Vol. 14, No. 2, pp. 49-58; 1933. See, also, Z. Physik, Vol. 78, No. 1-2, pp. 26-42; 1932.

8. O. Scherzer, "Theory of Electric Electron Condensing Lenses," Z. Physik, Vol. 80, No. 3-4, pp. 193-202; 1933.

9. W. Glaser, "Geometrical Optics of Electron Rays," Z. Physik, Vol. 80, No. 7-8, pp. 451-464; 1933.

10. V. K. Zworykin, "Electron Optics," J. Franklin Inst., Vol. 215, pp. 535-555; 1933.

11. O. Scherzer, "Errors of Electron Lenses," Z. Physik, Vol. 101, No. 9-10, pp. 593-603; 1936.

12. W. Rogowski, "Aberrations of Electron Images," Arch. Elektrotech., Vol. 31, No. 9, pp. 555-593; 1937.

13. L. H. Bedford, "Electron Lens Formulas," Phys. Soc. Proc., Vol. 46, pp. 882-888; 1934. See, also, C. J. Davisson and C. J. Calbick, Phys. Rev., Vol. 38, p. 585; 1931: and Vol. 42, p. 580; 1932.

14. R. Gans, "Electron Paths in Electron Optics," Z. Tech. Physik, Vol. 18, No. 2, pp. 41-48; 1937.

15. L. S. Polotovski, "Studies on Electron Optical Systems," Izveztia Electroprom. Slabovo Toka, No. 5, pp. 22-35, and No. 6, pp. 13-27; 1934.

16. D. W. Epstein, "Electron Optical Systems of Two Cylinders," Proc. I. R. E., Vol. 24, No. 8, pp. 1095-1139; 1936.

17. G. A. Morton and E. G. Ramberg, "Electron Optics of an Image Tube," Physics, Vol. 7, No. 12, pp. 451-459; 1936.

18. E. G. Ramberg, "Simplified Derivation of General Properties of an Electron Optical Image," J. Optical Soc. Am., Vol. 29, pp. 79-83; 1939.

19. W. Henneberg and A. Recknagel, "Chromatic Errors of Electron Optical Systems," Z. tech. Physik, Vol. 16, No. 8, pp. 230-235; 1935.

20. E. Gundert, "Demonstration of the Aberrations of Electron Lenses," Physik. Z., Vol. 38, No. 12, pp. 462-467; 1937.

21. B. von Borries and E. Ruska, "Electron Super Microscope," Wiss. Veröffentl. Siemenswerke, Vol. 17, No. 1; 1938.

Lecture 17

ELECTRON OPTICS

Part III . ABBERATIONS IN ELECTRON OPTICS

G. A. Morton and E. G. Ramberg

We have seen in the preceding lectures that an axially symmetric arrangement of electrodes at various potentials positive with respect to the source of the electrons, and of magnets, will form an image, real or virtual,[1] of any object emitting or irradiated by electrons. More precisely, we found that, if we limited the object point to the immediate neighborhood of the axis of symmetry and the angles of inclination of the electron paths to very small values and if we assumed the initial velocities leaving the object point to be identical in magnitude, there was a plane in which all of the rays from the object point considered intersected anew in a point whose distance from the axis was proportional to that of the object point from the axis. We called this new point the image point or the image of the object point. Thus, as at the same time the relative azimuthal position of the object points to each other, and image points to each other, were the same, a geometrically similar image of the object was produced in the image plane.

Suppose, now, that we drop the above-mentioned restrictions and turn our attention to real objects and images, finite in extent and produced by electron pencils of finite cross-section, made up of electrons of slightly varying initial velocities. We will now find that the image is no longer geometrically similar to the object and, furthermore, that an object point even on or near the axis is no longer imaged in a corresponding sharp point in the image plane, but rather in a disk — the rays of the finite pencil leaving the object point no longer intersect in a single point in the image plane. It is the major problem of electron optics to design electrode configurations and magnetic-field producing coils which minimize the deviations of the image obtained from a truly geometrically similar, sharp picture of the object. These deviations — more exactly the separations between the points of intersection of the real electron rays with the image plane and the points of intersection as predicted from an extrapolation of the faithful, Gaussian image produced by infinitely narrow pencils of uniform velocity in the neighborhood of the axis — are called the aberrations of the image.

In considering these aberrations it is convenient to separate them into components arising from different causes and to consider these individually. We thus shall treat first the aberrations due to the finite aperture of the electron pencils and the finite separations of the object points from the axis — the "Seidel aberrations" — and then those due to variations in the initial velocities of the electrons — the chromatic aberrations. Both terms are borrowed from optics: the latter is derived from the fact that, as for the different colors or wavelengths of light the refracting media have a different index of refraction, so for different initial velocities the effective index of refraction of the field governing the paths of the electrons differs — indeed, for purely electric fields, it has been brought out that this index is proportional to the velocity of the electron.

To begin with, we shall turn our attention to the Seidel aberrations, i.e., the aberrations for electrons leaving the object with identical velocities though, in general, in different directions. It is interesting to derive their general form as well as the general imaging properties of the axially symmetric electric and magnetic fields from conditions of symmetry. The magnitudes of the aberrations as well as the position and magnification of the image must in general be derived by evaluation of the ray equation as has been shown for the latter two quantities in the preceding sections.

Let us, then, consider an axially symmetric array of electrodes at various potentials, as well as of magnets, with the z-axis as their axis of symmetry (Fig. 1). Normal to this axis let there be an object plane O with the coordinate system x_o, y_o and, beyond the focusing fields, an image plane I (whose position is yet to be fixed) with coordinates x_i, y_i, the respective origins lying on the z-axis. We assume the space between

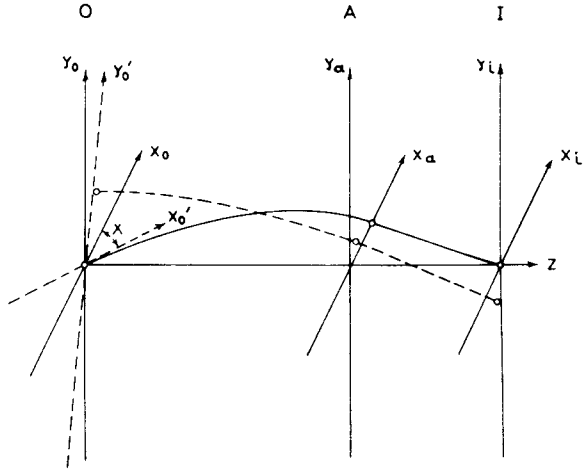


Fig. 1 - Object, aperture, and image planes.

[1] In this lecture we shall confine ourselves to the more important case of real images.

O and I to be free from matter(space charge negligible), so that Laplace's equation holds for both the electric and the magnetic field, and furthermore, that the space surrounding the plane I is field-free. We now introduce between O and I, also normal to the z-axis, an "aperture plane" A with coordinates x_a, y_a — close enough to I that the entire space between A and I may be regarded as field-free. We give the same orientation to the several coordinate systems, so that, e.g., the x_o-, x_a- and x_i-axes are coplanar and parallel.

Suppose now that electrons leave various points of the object plane, throughout with the same absolute velocity, though generally in different directions. If we fix our attention on one such electron leaving the point x_o, y_o in the object plane and aimed so that it passes through the point x_a, y_a of the aperture plane, its path will, in view of the fixed field conditions, be uniquely determined and accordingly also its intersection x_i, y_i with the image plane. Hence we may regard the coordinates x_i and y_i as functions of the coordinates of the ray in the object and aperture planes, x_o, y_o, x_a, and y_a; and, in view of the continuity of the force fields to which the electrons are subjected, express them in terms of a power series of these latter four coordinates, such as

$$x_i = a_0 + a_1x_o + a_2y_o + a_3x_a + a_4y_a + a_5x_o^2 + a_6x_oy_o + a_7y_o^2 + a_8x_ox_a + \ldots \quad (1)$$

There is a similar expression, with different coefficients, for y_i. We will see that we can simplify these expressions very considerably by making use of the symmetry conditions of the problem. We must here satisfy the obvious requirement that, if we rotate the radius vectors to the ray intersections in planes O and A through any given angle φ, the corresponding intersection in plane I is found by rotating the radius vector to x_i, y_i through the same angle φ. Thus equation (1) must remain satisfied if we carry out the simultaneous substitution:

$$\begin{aligned} x_i &\to x_i\cos\varphi - y_i\sin\varphi \\ x_o &\to x_o\cos\varphi - y_o\sin\varphi \\ x_a &\to x_a\cos\varphi - y_a\sin\varphi \\ y_o &\to x_o\sin\varphi + y_o\cos\varphi \\ y_a &\to x_a\sin\varphi + y_a\cos\varphi \end{aligned} \quad (2)$$

Putting $\varphi = 180^\circ$ simply changes the sign of all the coordinates and we obtain

$$-x_i = a_0 - a_1x_o - a_2y_o - a_3x_a - a_4y_a + a_5x_o^2 + a_6x_oy_o + a_7y_o^2 + a_8x_ox_a + \ldots$$

Adding equation (1), we find that the sum of all the terms of even order in the coordinates must vanish:

$$0 = 2a_0 + 2a_5x_o^2 + 2a_6x_oy_o + 2a_7y_o^2 + 2a_8x_ox_a + \ldots$$

As this must be true for all values in a continuous range of x_o, y_o, x_a, y_a, all the coefficients of the terms of even order $a_0, a_5, a_6, a_7, a_8 \ldots$ must vanish individually and we are left with terms in the first, third, fifth ... order in the coordinates. If we confine our attention to regions of the object very close to the axis and electron rays leaving these which, also, strike the aperture plane very close to the axis (i.e., imaging pencils of very small aperture), we gain sufficient information regarding the pattern in the image plane if we consider the terms of the first order only. Let us, then, substitute equation (2) in equation (1) with φ left arbitrary. As the resulting equation must be valid for all values of φ, it must be fulfilled by the coefficients on the two sides of the equation of cos φ and sin φ individually. Hence, we obtain

$$\begin{aligned} x_i^{(1)} &= a_1x_o + a_2y_o + a_3x_a + a_4y_a \\ y_i^{(1)} &= -a_2x_o + a_1y_o - a_4x_a + a_3y_a \end{aligned} \quad (3)$$

the superscript (1) indicating that the expressions constitute an approximation taking account of the first-order terms only.

Consider now a ray leaving the point $x_o = 0$, $y_o = 0$ and passing through $x_a = c$, $y_a = 0$. For it we have

$$\begin{aligned} x_i^{(1)} &= a_3c \\ y_i^{(1)} &= -a_4c \end{aligned} \quad (4)$$

Now the only tangential forces which the electron undergoes arise from the action of the magnetic field. According to Larmor's theorem its angular velocity about the axis is, however, equal to $eH_z/(2m)$, where H_z is the longitudinal component of the magnetic field at the point of observa-

tion.[2] As, however, $H_z = 0$ in the aperture plane and beyond, the angular velocity of the electron must here be zero throughout. The path of the electron between the aperture and image planes is thus a straight line proceeding in a meridional plane — in our example in the zx-plane. Hence, in (4),

$$y_i^{(1)} = 0 \quad \text{and, as} \quad c \neq 0\,, \quad a_4 = 0$$

We now take one more step: the ray in the zx-plane which we are considering will, in general be inclined to the axis of symmetry. Let us, hence, follow it to its point of intersection with the axis and fix the image plane I as the plane normal to the axis passing through this point.[3] With this choice of I, we have $x_i = 0$ and hence also $a_3 = 0$. Thus we obtain finally:

$$\begin{aligned} x_i^{(1)} &= a_1 x_0 + a_2 y_0 = \beta x_0' \\ y_i^{(1)} &= -\,a_2 x_0 + a_1 y_0 = \beta y_0' \end{aligned} \tag{5}$$

where

$$\begin{aligned} x_0' &= x_0 \cos\chi - y_0 \sin\chi \\ y_0' &= x_0 \sin\chi + y_0 \cos\chi \\ \chi &= \arctan\left(-\frac{a_2}{a_1}\right) \\ \beta &= (a_1^2 + a_2^2)^{\frac{1}{2}} \end{aligned} \tag{6}$$

x_0', y_0' are the coordinates of the object point referred to a coordinate system rotated with respect to the original one (and the systems in planes A and I) through the angle χ. We thus have the final result that, to a first approximation in the coordinates of the object and aperture planes, all the rays leaving any point in the object plane O are reunited in a point in the image plane I. Furthermore, to any array of object points there corresponds a geometrically similar array of image points with separations magnified, or reduced, by the factor β and rotated, one relative to the other, through the angle χ. In short, to a first approximation, our axially symmetrical field combination gives a sharp and faithful "Gaussian" image of the object.

For purely electric fields (as for an ordinary optical lens system), the angle χ vanishes, as rays starting in a given meridional plane do not leave this plane. Owing to the fact that in this case all of the forces acting on the electron are radial or axial, the relation (1) must remain true if we reflect all of the coordinates at any meridional plane, e.g., change the signs of the x-coordinates and leave those of the y-coordinates unchanged (reflection at the yz-plane). This requirement leads to $a_2 = 0$, and hence $\chi = 0$.

If we drop the assumption of an image field of infinitesimal extent and imaging pencils of infinitesimal aperture, we must take into consideration the terms of higher order in the expressions for x_i and y_i — they inform us of the deviations of the real image from the Gaussian sharp and geometrically faithful image — the aberrations. For many purposes, if the image field and apertures are not too large, it is sufficient to restrict our attention to the terms of the third order. We obtain, then, the so-called third-order (Seidel) aberrations. Thus, we may write, giving a new meaning to the coefficients a_1, a_2, a_3, a_4,

---

[2] If we introduce the polar coordinates r, φ, the exact integration of Newton's equation for the φ-direction

$$\frac{d}{dt}(mr^2\dot{\varphi}) = re\,(\dot{r}H_z - \dot{z}H_r)$$

taking account of Laplace's equation

$$\frac{\partial H_z}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(rH_r) = 0$$

yields

$$r^2\dot{\varphi} = [r^2\dot{\varphi}]_{z=0} + \frac{e}{m}\left[\frac{H_z}{2}r^2 - \frac{1}{8}\left(\frac{\partial^2 H_z}{\partial r^2}\right)_{r=0} r^4 - \frac{1}{72}\left(\frac{\partial^4 H_z}{\partial r^4}\right)_{r=0} r^6 \ldots\right]_0^z$$

This vanishes, if $r = 0$ at $z = 0$, in any plane in which $H_z \equiv 0$.

[3] Our original assumptions regarding the image plane I are equivalent to postulating that this point fall into the field-free space beyond the focusing fields.

$$\begin{aligned}
\Delta x_i = {} & a_1 x_a^3 + a_2 x_a^2 y_a + a_3 x_a y_a^2 + a_4 y_a^3 && \textit{Aperture Defect (Spherical Aberration)} \\
& + b_1 x_0' x_a^2 + b_2 x_0' x_a y_a + b_3 x_0' y_a^2 + b_4 y_0' x_a^2 + b_5 y_0' x_a y_a + b_6 y_0' y_a^2 && \textit{Coma} \\
& + c_1 x_0'^2 x_a + c_2 x_0'^2 y_a + c_3 x_0' y_0' x_a + c_4 x_0' y_0' y_a + c_5 y_0'^2 x_a + c_6 y_0'^2 y_a && \textit{Curvature of Field and Astigmatism} \\
& + d_1 x_0'^3 + d_2 x_0'^2 y_0' + d_3 x_0' y_0'^2 + d_4 y_0'^3 && \textit{Distortion}
\end{aligned} \tag{7}$$

and a similar expression for Δy_i. Δx_i and Δy_i are the third-order deviation of the ray intersection with the image plane from the corresponding first-order, Gaussian, intersection or image point:

$$x_i^{(3)} = x_i^{(1)} + \Delta x_i ; \qquad y_i^{(3)} = y_i^{(1)} + \Delta y_i$$

For convenience we now refer our object plane coordinates to the rotated (primed) coordinate system. To the right of every group of terms involving the same power of the coordinates of the object plane and of the coordinates of the image plane, we have written down the customary designation of the corresponding aberration. Their origin will become evident later.

An application of our symmetry condition will show that the twenty coefficients in equation (7) and the further twenty coefficients in the corresponding expression Δy_i are by no means independent. Let us, then, substitute equation (2) in equation (7) and satisfy the resulting equation individually for the coefficients of sin φ, cos φ, $\sin^3\varphi$, and $\cos^3\varphi$. Expressing the coefficients in terms of each other as far as the resulting relations permit, we find:

$$\begin{aligned}
\Delta x_i = {} & a_1 x_a (x_a^2 + y_a^2) + a_2 y_a (x_a^2 + y_a^2) \\
& + b_1 x_0' (x_a^2 + y_a^2) + b_5 y_a (y_0' x_a - x_0' y_a) + b_2 y_a (x_0' x_a + y_0' y_a) + b_4 y_0' (x_a^2 + y_a^2) \\
& + c_1 x_0' (x_0' x_a + y_0' y_a) + c_5 y_0' (y_0' x_a - x_0' y_a) + c_3 y_0' (x_0' x_a + y_0' y_a) + c_2 y_a (x_0'^2 + y_0'^2) \\
& + d_1 x_0' (x_0'^2 + y_0'^2) + d_2 y_0' (x_0'^2 + y_0'^2)
\end{aligned} \tag{8}$$

$$\begin{aligned}
\Delta y_i = {} & - a_2 x_a (x_a^2 + y_a^2) + a_1 y_a (x_a^2 + y_a^2) \\
& - b_4 x_0' (x_a^2 + y_a^2) - b_5 x_a (y_0' x_a - x_0' y_a) - b_2 x_a (x_0' x_a + y_0' y_a) + b_1 y_0' (x_a^2 + y_a^2) \\
& + c_1 y_0' (x_0' x_a + y_0' y_a) - c_5 x_0' (y_0' x_a - x_0' y_a) - c_3 x_0' (x_0' x_a + y_0' y_a) - c_2 x_a (x_0'^2 + y_0'^2) \\
& - d_2 x_0' (x_0'^2 + y_0'^2) + d_1 y_0' (x_0'^2 + y_0'^2)
\end{aligned}$$

If we again consider a ray leaving the point on the axis in the object plane ($x_0' = y_0' = 0$) and passing through the point ($x_a = c$, $y_a = 0$), Larmor's theorem tells us that the ray must proceed in the zx-plane between A and I and that thus $\Delta y_i = 0$ and hence $a_2 = 0$. We have thus reduced the number of independent third-order coefficients to 11. For purely electric fields, in which case, as we have mentioned, the relations must hold if we reflect the coordinates at any meridional plane, we have in addition $b_2 = b_4 = c_3 = c_2 = d_2 = 0$, so that the number of independent coefficients reduces to 6. To reduce them to their true number, i.e., 8 for mixed and 5 for purely electric fields, we must introduce a new factor into our considerations. We may derive this final simplification if we postulate the existence of a wave-surface — i.e., a continuous surface normal to all the rays of a pencil originating in one particular object point — which the central ray of the pencil intersects in the aperture plane. Let the equation of this wave-surface be $f(x,y,z) = 0$. The direction cosines of the normal at any point of the surface, coinciding with an electron ray intersecting the image plane in the point (x_i, y_i, z_i), will be given by

$$\begin{aligned}
\frac{\partial f}{\partial x} \Big/ \frac{\partial f}{\partial s} &= \frac{(x_i - x)}{\left[(x_i-x)^2 + (y_i-y)^2 + (z_i-z)^2\right]^{\frac{1}{2}}} \\
\frac{\partial f}{\partial y} \Big/ \frac{\partial f}{\partial s} &= \frac{(y_i - y)}{\left[(x_i-x)^2 + (y_i-y)^2 + (z_i-z)^2\right]^{\frac{1}{2}}}
\end{aligned} \tag{9}$$

As we shall see further on, the gradient $\partial f/\partial s$ of the wave-function $\underline{f}$ is a constant in a field-free space; furthermore, as the inclinations of the rays to the axis are assumed to be small, we may replace the root in the denominator on the right by the constant $Z = z_i - z_a$. Differentiating, we then find

$$\frac{\partial^2 f}{\partial x\,\partial y} = c\,\frac{\partial x_i}{\partial y} \approx c\,\frac{\partial x_i}{\partial y_a} \qquad \frac{\partial^2 f}{\partial y\,\partial x} = c\,\frac{\partial y_i}{\partial x} \approx c\,\frac{\partial y_i}{\partial x_a} \tag{10}$$

as x_a and y_a differ only in the second order from the corresponding coordinates $\underline{x}$ and $\underline{y}$ of the wave-surface. If the wave-surface, or the function $\underline{f}$, exists, the two second derivatives appearing in equations (10) must be equal and we obtain as final condition

$$\frac{\partial x_i}{\partial y_a} = \frac{\partial y_i}{\partial x_a} \tag{11}$$

In view of the approximations made in its derivation, this condition applies only as long as we restrict the terms in x_i and y_i to those of the first and third order in the object and aperture coordinates.

Applying equation (11) to equation (8) we obtain, finally:

$$\begin{aligned}
\Delta x_i &= a_1 x_a (x_a^2+y_a^2) \\
&+ \frac{b_1}{3}\left[x_0'(x_a^2+y_a^2) + 2x_a(x_0'x_a+y_0'y_a)\right] + \\
&\frac{b_4}{3}\left[y_0'(x_a^2+y_a^2) - 2x_a(x_0'y_a-y_0'x_a)\right] \\
&+\left[\frac{(c_1+c_5)}{2}\right]x_a(x_0'^2+y_0'^2) + \left[\frac{(c_1-c_5)}{2}\right] \\
&\left[x_a(x_0'^2-y_0'^2) + 2x_0'y_0'y_a\right] + \\
&c_2\left[y_a(x_0'^2-y_0'^2) - 2x_0'y_0'x_a\right] \\
&+ d_1 x_0'(x_0'^2+y_0'^2) + d_2 y_0'(x_0'^2+y_0'^2)
\end{aligned} \tag{12}$$

$$\begin{aligned}
\Delta y_i &= a_1 y_a (x_a^2+y_a^2) \\
&+ \frac{b_1}{3}\left[y_0'(x_a^2+y_a^2) + 2y_a(x_0'x_a+y_0'y_a)\right] - \\
&\frac{b_4}{3}\left[x_0'(x_a^2+y_a^2) + 2y_a(x_0'y_a-y_0'x_a)\right] \\
&+\left[\frac{(c_1+c_5)}{2}\right]y_a(x_0'^2+y_0'^2) - \left[\frac{(c_1-c_5)}{2}\right] \\
&\left[y_a(x_0'^2-y_0'^2) - 2x_0'y_0'x_a\right] + \\
&c_2\left[x_a(x_0'^2-y_0'^2) + 2x_0'y_0'y_a\right] \\
&+ d_1 y_0'(x_0'^2+y_0'^2) - d_2 x_0'(x_0'^2+y_0'^2)
\end{aligned}$$

We shall now consider the component aberrations individually, and introduce polar coordinates

$$x_0' = r_0 \cos\varphi_0' \qquad y_0' = r_0 \sin\varphi_0' \qquad x_a = r_a \cos\varphi_a \qquad y_a = r_a \sin\varphi_a$$

to facilitate the discussion.

If we consider an object point on the axis, the only aberration arising is the so-called <u>aperture defect</u> or <u>spherical aberration</u>:

$$\Delta x_i = a_1 r_a^3 \cos\varphi_a \qquad \Delta y_i = a_1 r_a^3 \sin\varphi_a \tag{13}$$

The rays passing through a circular zone of the aperture plane with a radius equal to r_a pass through a second circular zone in the image plane

$$\Delta x_i^2 + \Delta y_i^2 = (a_1 r_a^3)^2 \tag{13'}$$

whose radius is proportional to the cube of r_a. Fig. 2 is taken from a paper by Diels and Wendt.[4] At the top it shows the origin of the aperture defect in a lens, and below this, beam cross-section patterns obtained experimentally at the positions 1, 2, and 3 relative to a magnetic electron lens imaging a point source. The upper row was obtained with a pencil of small aperture; the lower

[4] K. Diels and G. Wendt, "Die 8 Bildfehler dritter Ordnung Magnetischer Elektronenlinsen," Zeits. f. techn. Physik," Vol. 18, No.3, p. 65; 1937.

one with one of large aperture. The aperture defect arises from the fact that the marginal rays are bent more strongly toward the axis and cut the axis at points between the lens and the image plane. In the presence of this aberration, we see from the figure that the sharpest image is obtained at a point between the lens and the Gaussian image plane (position 3).

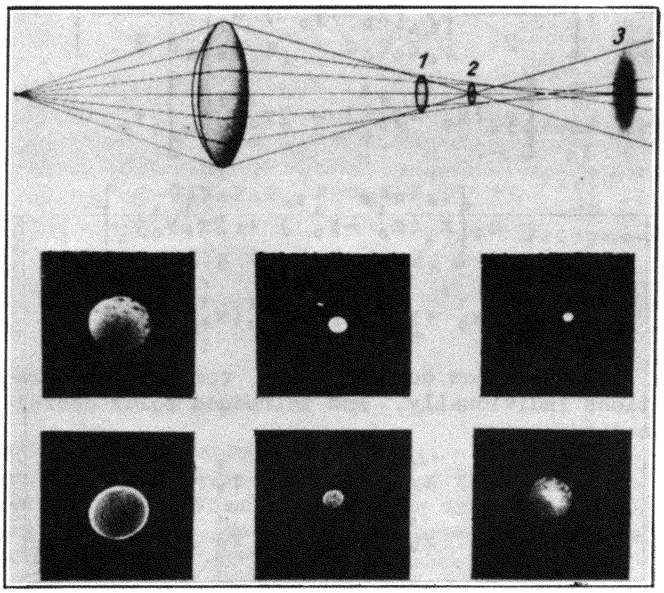


Fig. 2 - Aperture defect.
(K. Diels and G. Wendt, reference 4)

The terms containing the aperture coordinates in the second and those of the object plane in the first power lead to another aberration, known as coma. It is the aberration which, together with the spherical aberration already described, affects primarily points lying only little off the axis, but imaged with rather large aperture pencils. For purely electric fields we find for points lying on the y_0'-axis of the object plane,

$$\Delta x_i = \frac{b_1}{3} r_0 r_a^2 \sin 2\varphi_a$$
$$\Delta y_i = \frac{b_1}{3} r_0 r_a^2 (2 - \cos 2\varphi_a) \qquad (14)$$

or, for an aperture zone of radius r_a:

$$\Delta x_i^2 + \left(\Delta y_i - \frac{2 b_1 r_0 r_a^2}{3}\right)^2 = \left(\frac{b_1 r_0 r_a^2}{3}\right)^2 \qquad (14')$$

For aperture zones of increasing radius, this corresponds to a series of circles enclosed by two lines at 30° to the y-axis intersecting at

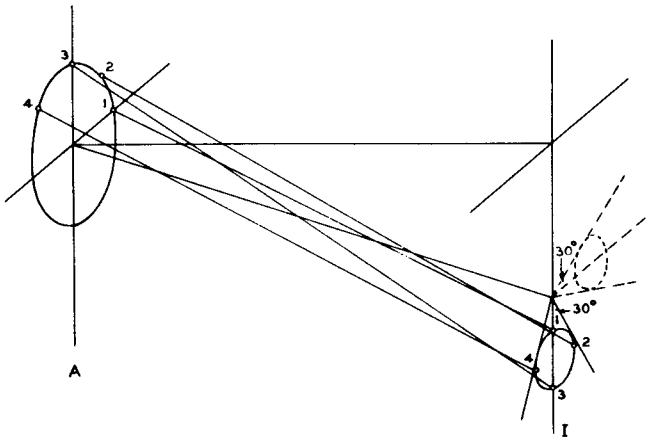


Fig. 3a - Coma.

the Gaussian image point (Fig. 3a). The superposition of these circles together with the spherical aberration giving the vertex of the 60° sector a diffuse appearance results in the comet-like structure shown in the upper middle photo-

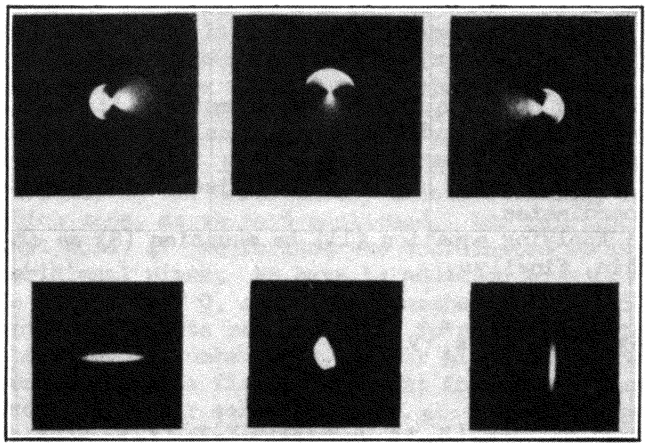


Fig. 3b - Coma and astigmatism.
(K. Diels and G. Wendt, reference 4)

graph in Fig. 3b.[4] In the presence of a magnetic field ($b_4 \neq 0$), the aberration figure (equation 14') is merely changed in scale and rotated about the Gaussian image point through the angle arc tan (b_4/b_1):

$$\Delta x_i = r_0 r_a^2 \frac{\left[b_1 \sin 2\varphi_a + b_4 (2 + \cos 2\varphi_a)\right]}{3}$$
$$\Delta y_i = r_0 r_a^2 \frac{\left[b_1 (2 - \cos 2\varphi_a) + b_4 \sin 2\varphi_a\right]}{3} \qquad (15)$$

[4] Loc. cit.

$$\left(\Delta x_i - \frac{2b_4 r_o r_a^2}{3}\right)^2 + \left(\Delta y_i - \frac{2b_1 r_o r_a^2}{3}\right)^2 = (b_1^2 + b_4^2)\left(\frac{r_o r_a^2}{3}\right)^2 \qquad (15')$$

The figures on the left and right of the upper row in Fig. 3b represent the extreme case $b_1 = 0$, $b_4 \neq 0$, indicated in Fig. 3a by dotted lines. The two figures were obtained by reversing the current in the magnetic focusing coils.

If we deal with extended objects and beams of relatively small aperture, such as is the case, for instance, in the image tube, the sharpness of the image is primarily influenced by the aberrations proportional to the square of the distance of the object point from the axis and to the first power of the aperture distance. These are, for a purely electric field: (1) curvature of field, and (2) astigmatism. For $x_o' = 0$, we have

1.
$$\Delta x_i = \left[\frac{c_1+c_5}{2}\right] r_o^2 r_a \cos\varphi_a$$
$$\Delta y_i = \left[\frac{c_1+c_5}{2}\right] r_o^2 r_a \sin\varphi_a \qquad (16)$$

2.
$$\Delta x_i = -\left[\frac{c_1-c_5}{2}\right] r_o^2 r_a \cos\varphi_a$$
$$\Delta y_i = \left[\frac{c_1-c_5}{2}\right] r_o^2 r_a \sin\varphi_a \qquad (17)$$

Together, for an aperture zone of radius r_a, these give rise to an aberration figure in the shape of an ellipse:

$$\frac{\Delta y_i^2}{(c_1 r_o^2 r_a)^2} + \frac{\Delta x_i^2}{(c_5 r_o^2 r_a)^2} = 1 \qquad (16')$$

If $c_1 = c_5$ (astigmatism = 0), the figure becomes a circle. Fig. 4 shows the origin of the ellipse about the Gaussian image point. We also see here that the imaging pencil converges into a line segment normal to the meridional plane of the principal ray at one point and into a line segment in this plane somewhat further on. The first we call the tangential, the second the sagittal image of the object point. The surfaces of revolution, tangent to the Gaussian image plane I on the axis, which contain the tangential and the

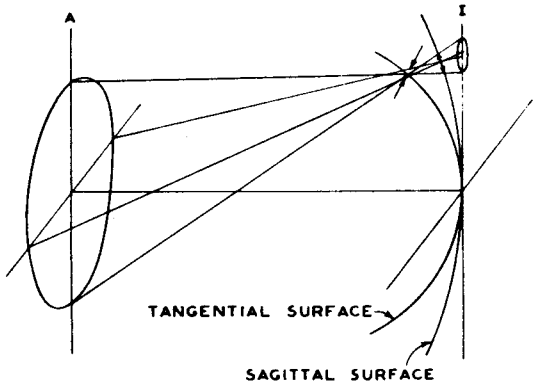


Fig. 4 – Curvature of field and astigmatism.

sagittal images of all the points in the object plane are called the tangential and sagittal image surfaces, respectively. From equation (16) and equation (17), we may show that their radii of curvature on the axis are given by

$$R_t = \frac{m^2}{2c_1 Z} \quad \text{and} \quad R_s = \frac{m^2}{2c_5 Z} \qquad (18)$$

respectively, where m is the magnification of the image and Z is the separation of the aperture and image planes. For $c_1 = c_5$, the two image surfaces coincide and we have on this surface a sharp, stigmatic image of the object. For $c_1 = -c_5$ (curvature of field = 0), on the other hand, the radii of curvature of the two surfaces are equal and opposite in sign, so that the image is sharpest in the Gaussian image plane. If both conditions are fulfilled ($c_1 = c_5 = 0$), the image is, of course, in the absence of other aberrations, both stigmatic and situated in the Gaussian image plane.

Fig. 5 shows the origin of astigmatism and curvature of field in an electrostatic image tube with a flat cathode. Two flat imaging pencils, in the meridional plane and normal thereto, have been drawn in and, in addition, the tangential (T) and sagittal (S) image surfaces. It is seen that in this case the curvatures are extremely large. A further example of astigmatism and curvature of field is given by the lower row of pictures in Fig. 3b. These pictures show the image

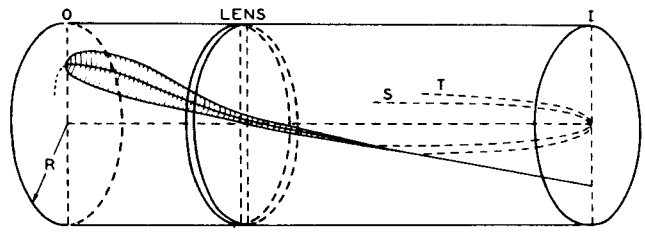


Fig. 5 – The image formation in a flat-cathode image tube.

of a point in the tangential and the sagittal image surface and on a surface halfway between these.[4]

In the presence of a magnetic field the aberration figure for combined curvature of field and astigmatism is given by

$$\Delta x_i = c_5 r_0^2 r_a \cos \varphi_a - c_2 r_0^2 r_a \sin \varphi_a$$
$$\Delta y_i = c_1 r_0^2 r_a \sin \varphi_a - c_2 r_0^2 r_a \cos \varphi_a \qquad (19)$$

This, again, may be shown to be an ellipse, with its axes rotated through an angle equal to

$$\frac{1}{2} \arctan \frac{2c_2}{c_1 - c_5}$$

For $c_1 = -c_5$, the ellipse degenerates to a circle; even if $c_1 = -c_5 = 0$, however, there will be no stigmatic image surface as long as $c_2 \neq 0$.

We now come to the last two of the Seidel aberrations, ordinary and magnetic distortion. These depend only on the coordinates of the object plane. We have

1.
$$\Delta x_i = d_1 r_0'^3 \cos \varphi_0'$$
$$\Delta y_i = d_1 r_0'^3 \sin \varphi_0' \qquad (20)$$

2.
$$\Delta x_i = d_2 r_0'^3 \sin \varphi_0'$$
$$\Delta y_i = -d_2 r_0'^3 \cos \varphi_0' \qquad (21)$$

These aberrations do not affect the sharpness of the image (as they do not depend on the aperture coordinates), but only the similarity to the object. The effect of the first distortion is to compress the outer parts of the image if the sign of d_1 is negative and to expand them if it is positive; in the first case we have barrel-shaped, in the second cushion-shaped distortion (see Fig. 6a). The latter is the type usually encountered in electron optics.

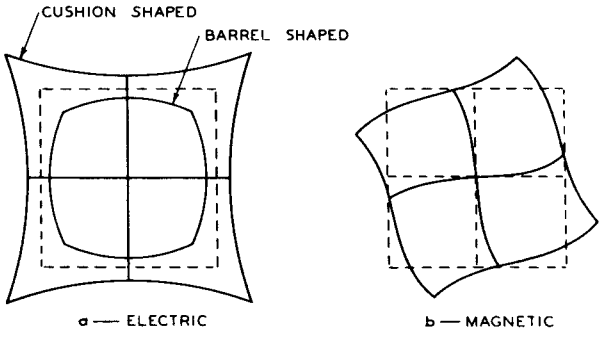


Fig. 6 - Distortion.

In the second case the image is twisted with respect to the Gaussian image, the direction of the twist being determined by the sign of d_2 or the direction of the currents in the focusing coils (see Fig. 6b). The axes themselves go over into cubic parabolas. Fig. 7 shows experimentally obtained electron images of a wire mesh[4] obtained in the presence of ordinary distortion and with both types of distortion present at the same time. The second and third pictures differ owing to the reversal of the currents in the focusing coils.

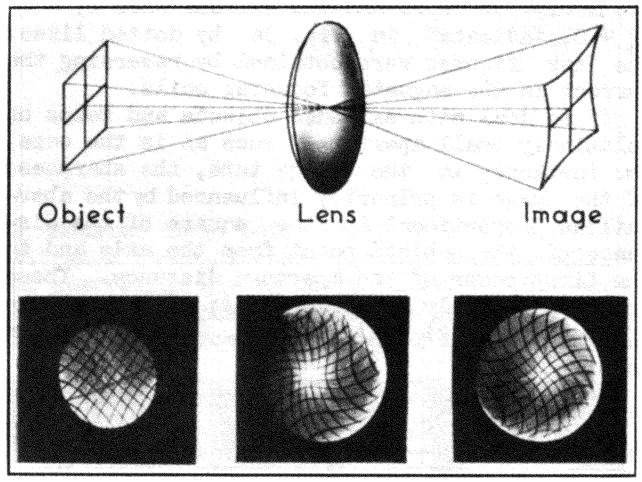


Fig. 7 - Distortion.
(K. Diels and G. Wendt, reference 4)

A remark should be added regarding the choice of the aperture plane. So far we have spoken of it as a rather arbitrary, physical plane. It is in many ways more advantageous to choose it, in agreement with the custom in optics, as the plane of the exit pupil, i.e., the normal plane whose z-coordinate is given by the intersection with the axis of the central ray of a pencil imaging a point slightly off the axis, continued backwards as a straight line from the image point. Defined in this manner, it will in general be a virtual plane, but the ray paths in the space between it and the image plane will in any case, by definition, be straight and all the theory given so far as well as that still to be given will apply. Only with this choice will the effective aperture normally be circular and concentric with the axis as assumed in deriving the aberration figures and will furthermore be substantially independent of the position of the object point. In this case only do the aberration figures become simple and does the separation of the aberration into the eight separate components discussed above become significant.

In obtaining the final expression (equation 12) for the aberration, we made use, without further justification, of the existence of a "wave-surface" of the electrons. This may be

[4] Loc. cit.

proved readily with the aid of the Hamiltonian theory of mechanics which forms the basis of an alternative derivation of the aberration[5] which will be outlined below.

To begin with, this derivation introduces the ray function, otherwise called the point iconal or Hamiltonian characteristic function:

$$S = \int_{P_1}^{P_2} \left(\frac{v}{c} - \frac{e}{mc} A \cos\chi \right) ds = \int_{P_1}^{P_2} n ds \qquad (22)$$

where

$v = \left(\frac{2e\varphi}{m} \right)^{\frac{1}{2}}$ is the velocity

φ is the electric potential

A is the magnetic vector potential of the field (H = curl A)

χ is the angle between the direction of the magnetic vector potential A and the path of the electron s. The integration is to be carried out over that path connecting the points P_1 and P_2 which minimizes the integral S — this, we shall see, is the actual path traveled by an electron leaving P_1 and arriving at P_2. For, if we express the integral in the form

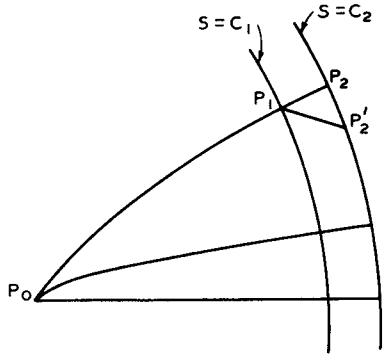
$$S = \int_{z_1}^{z_2} F \, dz$$
$$F = n \left[1 + \left(\frac{dx}{dz} \right)^2 + \left(\frac{dy}{dz} \right)^2 \right]^{\frac{1}{2}} \qquad (23)$$

the path minimizing it is a solution of the pertinent Euler equations

$$\frac{d}{dz} \frac{\partial F}{\partial x'} - \frac{\partial F}{\partial x} = 0 \quad ; \quad x' = \frac{dx}{dz} \quad ;$$
$$\frac{d}{dz} \frac{\partial F}{\partial y'} - \frac{\partial F}{\partial y} = 0 \quad ; \quad y' = \frac{dy}{dz} \quad ; \qquad (24)$$

which may be shown to be simply the equations of motion of the electron in our combined electric and magnetic field. For a given field, the ray function is a function of the coordinates of the initial point P_1 and of the final point P_2 of the path. If we ascribe to the field the (variable and anisotropic) index of refraction n, it may be described simply as the "electron-optical distance" between these two points.

Suppose now that we keep the point P_1, a particular object point acting as source of electrons, fixed. Then S = constant describes a wave-surface for the electrons leaving P_1: if we consider two such surfaces, $S = C_1$ and $S = C_2$, where C_1 and C_2 differ very little (Fig. 8), we readily see that, in regions where the magnetic



$$\Delta S = C_2 - C_1 = \int_{P_1}^{P_2} n \, ds \leq \int_{P_1}^{P_2'} n \, ds \quad \text{FOR ALL } P_2' \text{ ON } S = C_2$$

Fig. 8 – Electron wave-surfaces.

field vanishes (A = 0), such as at the aperture plane, all the electron rays must proceed normal to the surfaces. Only then will the increase in S in passing from the first to the second surface be a minimum, as required by the definition of S. Thus $S(P_2) - C$ is, for the proper choice of C, the function $f(x,y,z)$ which we introduced in equation (9), and the existence of the former function establishes the existence of the wave-surface there postulated.

Returning to the derivation of the aberrations with the aid of the ray function, we set up first the ray function for a pencil converging in the point

$$x_i^{(1)}, \; y_i^{(1)}, \; z_i^{(1)}$$

of the image plane, the coordinates x_a, y_a, z_a in the aperture plane of the constituent rays entering in as parameters:

[5] W. Rogowski, "Über Fehler von Elektronenbildern," Archiv für Elektrotechnik," Vol. 31, No.9, p. 555; 1937.

$$S_p = S_i - n_a\left[(x_i^{(1)} - x_a)^2 + (y_i^{(1)} - y_a)^2 + Z^2\right]^{\frac{1}{2}}$$

$$= S_i - n_a Z - \frac{n_a}{2Z}\left[(x_i^{(1)\,2} + y_i^{(1)\,2}) - 2(x_i^{(1)} x_a + y_i^{(1)} y_a) + (x_a^2 + y_a^2)\right] +$$

$$\frac{n_a}{8Z^3}\left[(x_a^2+y_a^2)^2 - 4(x_a^2+y_a^2)(x_i^{(1)} x_a + y_i^{(1)} y_a) + 2(x_a^2+y_a^2)(x_i^{(1)\,2} + y_i^{(1)\,2}) + 4(x_i^{(1)} x_a + y_i^{(1)} y_a)^2 - 4(x_i^{(1)} x_a + y_i^{(1)} y_a)(x_i^{(1)\,2} + y_i^{(1)\,2}) + (x_i^{(1)\,2} + y_i^{(1)\,2})^2\right] + \cdots \qquad (25)$$

Here n_a is the constant index of refraction for the field-free space between the aperture and image planes and Z is their separation, which is assumed to be large compared to the lateral coordinates entering into the expressions. S_i is an adjustable constant.

We compare this expression with the expression for a pencil diverging from the object point x_o, y_o, z_o, expanded in terms of the coordinates in the object and aperture planes:

$$S = S_o + a_1 x_o + a_2 y_o + a_3 x_a + a_4 y_a + a_5 x_o^2 + \cdots \qquad (26)$$

If we make use of the condition that the "electron-optical path distance" be invariant under rotation about the axis of the system, we can simplify this expansion as before and may write

$$S = S_o + A_1(x_o^2+y_o^2) + A_2(x_o x_a + y_o y_a) + A_3(x_a^2+y_a^2) + A_4(x_o y_a - y_o x_a) + B_1(x_a^2+y_a^2)^2 + B_2(x_a^2+y_a^2)(x_o x_a + y_o y_a) + B_3(x_a^2+y_a^2)(x_o^2+y_o^2) + B_4(x_o x_a + y_o y_a)^2 + B_5(x_o x_a + y_o y_a)(x_o^2+y_o^2) + B_6(x_o^2+y_o^2)^2 + B_7(x_o^2+y_o^2)(x_o y_a - y_o x_a) + B_8(x_o x_a + y_o y_a)(x_o y_a - y_o x_a) + B_9(x_a^2+y_a^2)(x_o y_a - y_o x_a) \qquad (27)$$

If we now fix the position of the image plane by requiring

$$Z = -\frac{n_a}{2A_3}$$

we find that, excluding expressions of the fourth and higher orders in the coordinates, the function S indeed represents a point pencil converging to the image point

$$x_i^{(1)} = -\frac{(A_2^2 + A_4^2)^{\frac{1}{2}}}{2A_3} \cdot x_o'$$

$$y_i^{(1)} = -\frac{(A_2^2 + A_4^2)^{\frac{1}{2}}}{2A_3} \cdot y_o' \qquad (28)$$

where

$$x_o' = x_o \cos\chi - y_o \sin\chi$$

$$y_o' = x_o \sin\chi + y_o \cos\chi$$

$$\chi = \arctan\frac{-A_4}{A_2}$$

as it then becomes, within these limits, if we set

$$S_i - n_a Z + (A_2^2 + A_4^2)\frac{(x_o^2 + y_o^2)}{4A_3} = S_o + A_1(x_o^2 + y_o^2)$$

identical with S_p. We may then substitute equation (28) in the expression for S_p (equation 25) and find, retaining the terms of the fourth order:

$$S - S_p = S_1(x_a^2+y_a^2)^2 + S_2(x_a^2+y_a^2)(x_o' x_a + y_o' y_a) + S_3(x_a^2+y_a^2)(x_o'^2 + y_o'^2) + S_4(x_o' x_a + y_o' y_a)^2 + S_5(x_o' x_a + y_o' y_a)(x_o'^2 + y_o'^2) + S_6(x_o'^2 + y_o'^2)^2 + S_7(x_o'^2 + y_o'^2)(x_o' y_a - y_o' x_a) + S_8(x_o' x_a + y_o' y_a)(x_o' y_a - y_o' x_a) + S_9(x_a^2+y_a^2)(x_o' y_a - y_o' x_a) \qquad (29)$$

where the $S_1 \ldots S_9$ are new coefficients.

Now, as S is of the form

$$\int_{P_1}^{P_2} n \, ds$$

the normal to the wave-surface through x_a, y_a has the direction cosines

$$\frac{1}{n_a}\frac{\partial S}{\partial x_a} \qquad \frac{1}{n_a}\frac{\partial S}{\partial y_a}$$

If, now, the ray leaving P_1 (x_0', y_0', z_0') and passing through x_a, y_a in the aperture plane intersects the image plane in x_i, y_i, its direction cosines between the aperture and image planes are given by

$$\frac{x_i - x_a}{\left[(x_i-x_a)^2 + (y_i-y_a)^2 + Z^2\right]^{\frac{1}{2}}} \approx \frac{x_i - x_a}{Z}$$

and

$$\frac{y_i - y_a}{\left[(x_i-x_a)^2 + (y_i-y_a)^2 + Z^2\right]^{\frac{1}{2}}} \approx \frac{y_i - y_a}{Z} ,$$

respectively. As this ray must coincide with the above-mentioned normal, the two pairs of direction cosines must be identical. Similarly, the direction cosines for the ray of the corresponding Gaussian point pencil passing through the same point of the aperture plane are

$$\frac{1}{n_a}\frac{\partial S_p}{\partial x_a} \approx \frac{x_i^{(1)} - x_a}{Z} , \quad \frac{1}{n_a}\frac{\partial S_p}{\partial y_a} \approx \frac{y_i^{(1)} - y_a}{Z}$$

Hence, we have for the third-order aberrations:

$$\Delta x_i = \frac{Z}{n_a}\frac{\partial (S - S_p)}{\partial x_a} ; \quad \Delta y_i = \frac{Z}{n_a}\frac{\partial (S - S_p)}{\partial y_a} \tag{30}$$

Carrying out the differentiations, we arrive at an expression for the aberration identical with that in equations (12) except for the notation employed for the 8 independent coefficients involved (the term with S_6 in equation (29) drops out in the differentiation).

Having now derived and considered the character of the individual Seidel or monochromatic aberrations, we still have to determine their magnitude, i.e., the magnitude of the coefficients $a_1 \ldots d_2$ or $S_1 \ldots S_9$, respectively. Scherzer[6] and Glaser[6] have derived expressions for these in terms of the axial electric and magnetic fields. The procedure used by Scherzer is, very briefly, the following. In a preceding section it has been shown that the path of an electron very close to the axis is given, in the case of a purely electric field, by the differential equation

$$r'' = - r' \frac{\Phi'}{2\Phi} - r \frac{\Phi''}{4\Phi} \tag{31}$$

where Φ is the electric potential along the axis. With magnetic field, this becomes

$$r'' = - r' \frac{\Phi'}{2\Phi} - r\left(\frac{\Phi''}{4\Phi} + \frac{eH^2}{8m\Phi}\right) \tag{32}$$

This equation may be obtained from Newton's second law

$$\begin{aligned} m\ddot{z} &= er\dot{\theta}H_r + e\frac{\partial \varphi}{\partial z} \\ m(\ddot{r} - r\dot{\theta}^2) &= - er\dot{\theta}H_z + e\frac{\partial \varphi}{\partial r} \\ \frac{d}{dt}(mr^2\dot{\theta}) &= - e\dot{z}H_r + e\dot{r}H_z \end{aligned} \tag{33}$$

by eliminating the time, making use of

$$\varphi(r,z) = \Phi(z) - r^2\frac{\Phi''(z)}{4} + r^4\frac{\Phi^{IV}(z)}{64} - \cdots \tag{34}$$

and dropping all terms involving higher powers of r and r' than the first. If we retain terms of the third order in r and r', and introduce for convenience the variable

$$u = re^{i\left(\theta - \int \frac{eH_z}{2m} dt\right)} = re^{i\left(\theta - \int \sqrt{\frac{e}{8m\Phi}} H_z \, dz\right)} \tag{35}$$

the equation becomes

$$u'' + u'\frac{\Phi'}{2\Phi} + u\left(\frac{\Phi''}{4\Phi} + \frac{eH^2}{8m\Phi}\right) = \frac{B_0}{\Phi} \tag{36}$$

where B_0 is a function of u, u^2, u', u'^2, their conjugate functions, and the axial fields together with their derivatives down to the third. Substituting in B_0 solutions of the homogeneous differential equation, it is possible, by familiar theorems governing differential equations of this type, to solve the same in terms of solutions of the homogeneous equation. From the

[6] In H. Busch and E. Brüche, "Beiträge zur Elektronenoptik." J. A. Barth, Leipzig, 1937.

solutions can be derived the complete expressions for the aberrations. Scherzer has carried this through and has given the expressions in a form involving derivatives of the electric potential only down to the second and of the magnetic field to the first.[6]

The practical utilization of these expressions and the general conclusions reached concerning their magnitude has been so far rather limited. In particular, it is of interest to know which of these aberrations can be made to disappear entirely. An examination of the form of the several coefficients has shown here that:

1) Distortion and coma can be caused to vanish completely, e.g., by making the magnification 1:1, the electric potential symmetrical about the midway plane between object and image, and the magnetic field antisymmetrical about this plane. In cases where the electrons start from rest, as for instance in the image tube, this method is, however, of no value, as they would have to arrive at the screen with zero energy as well.

2) Spherical aberration can be made to vanish under no circumstances.[7] The minimum spherical aberration for weak electric and magnetic lenses has been calculated by Scherzer[8] and by Rebsch and Schneider.[9] As weak lenses have little practical interest, we shall merely indicate their general conclusions:

a. The spherical aberration for weak electric and magnetic lenses is of the same order. In the optimum case, it is 20 per cent less for electric than for magnetic lenses of equal aperture, focal length, and length.

b. The limit for immersion objectives and magnetic lenses with image rotation is 1/9 as large as that for electric unit lenses (with the same potential on object and image side) and magnetic lenses without image rotation (compensating coils).

c. No improvement with respect to spherical aberration over either the electric or magnetic lens can be obtained by combining these two.

The spherical aberration of an electron lens consisting of two concentric cylinders has been studied experimentally, in view of its great importance in connection with the production of fine spots with electron guns, by Epstein[10] and Gundert.[11] Epstein placed a diaphragm with a series of apertures in the second anode cylinder and observed the image on a movable screen. He measured out the separations of the spots due to the outer apertures from the central spot. Using two diameter ratios, 1.5 and 3.6, the larger one being at the higher potential, Epstein obtained slightly better results with the ratio 3.6.

Gundert used the very different technique of introducing traces of gas into the tube — he could show that this did not affect his results appreciably — and photographing thread or gas-focused rays which he directed by means of deflecting coils. He found that the spherical aberration could be reduced considerably by making the second anode (at the higher potential) diameter smaller than that of the first anode. In this case, by bringing the marginal rays close to the edge of the electrode in the negative portion of the lens, he found that they are here refracted away from the axis so much more than the paraxial rays, that the excessive refraction toward the axis in the positive portion of the lens is partly compensated. Fig. 9 illustrates this condition.

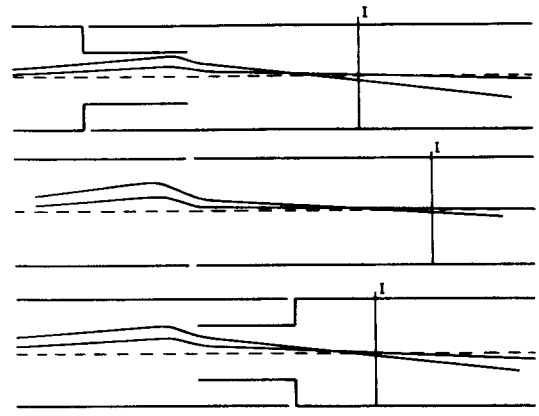


Fig. 9 - Aperture defect for cylinder lenses.

3) Curvature of field and astigmatism present special problems. Riedl[12] has been able to show that in the case of the electric weak short unit lens the curvature and astigmatism cannot be made to vanish for a plane object — the tangential image is invariably concave toward the lens. While this proof refers to a very limited case, all electron-optical systems considered so far (except the uniform electric and magnetic field types) appear to suffer to some extent from

[6] Loc. cit.

[7] O. Scherzer, "Über einige Fehler von Elektronenlinsen," Zeits. f. Physik, Vol. 101, No.9-10, p. 593; 1936.

[8] O. Scherzer, "Die Schwache Elektrische Einzellinse Geringster Sphärischer Aberration," Zeits. f. Physik, Vol. 101, No.1-2, p. 23; 1936.

[9] R. Rebsch and W. Schneider, "Der Öffnungsfehler Schwacher Elektronenlinsen," Zeits. f. Physik, Vol. 107, No.1-2; p. 138; 1937.

[10] D. W. Epstein, "Electron Optical System of Two Cylinders as Applied to Cathode-Ray Tubes," Proc. I.R.E., Vol. 24, No.8, p. 1095; August, 1936.

[11] E. Gundert, "Nachweis der Abbildungsfehler von Elektronenlinsen nach der Fadenstrahlmethode," Phys. Zeits., Vol. 38, No.12, p. 462; June, 1937. (Later papers contradict some of his conclusions.)

[12] H. Riedl, "Die Bildfehler 3 Ordnung der Kurzen Schwachen Rein Elektrischen Elektronen-Einzellinse," Zeits. f. Physik, Vol. 107, p. 210; 1937.

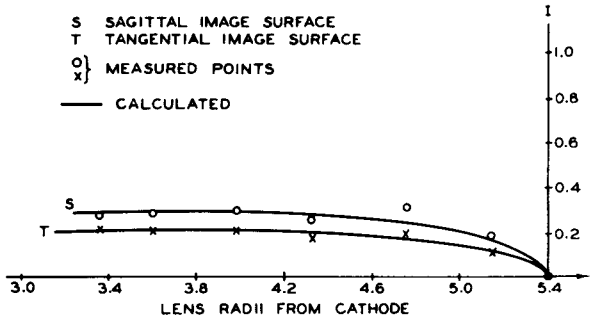


Fig. 10 - Curvature of field of image tube.

the same defect. Fig. 10 shows the shape of the tangential and sagittal surfaces as measured and calculated for an electrostatic image tube with a flat cathode.[13] The image of a simple square pattern obtained under these circumstances is

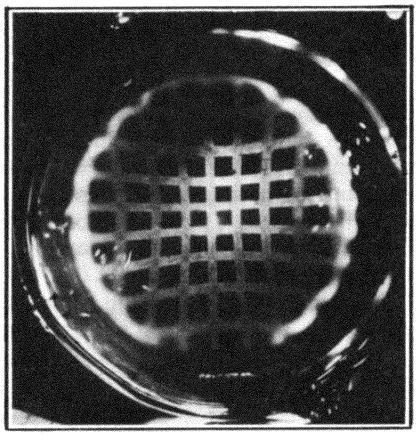


Fig. 11a - Image field of flat-cathode image tube.

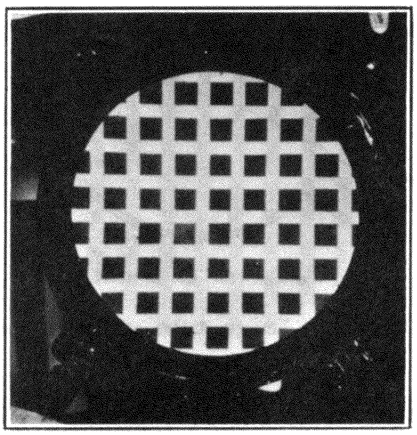


Fig. 11b - Image field of curved-cathode image tube.

[13] G. A. Morton and E. G. Ramberg, "Electron Optics of an Image Tube," Physics, Vol. 7, p. 451; December, 1936.

shown in Fig. 11a. The only practical way found so far for obtaining a completely flat image field is to curve the cathode or object; the improvement obtained by giving the cathode a radius of curvature approximately equal to its distance from the lens is shown in Fig. 11b. A rough idea of why this improvement is obtained may be gained from Fig. 12, in which a principal ray is traced through the field of both types of tubes. In the flat-cathode tube the marginal ray is

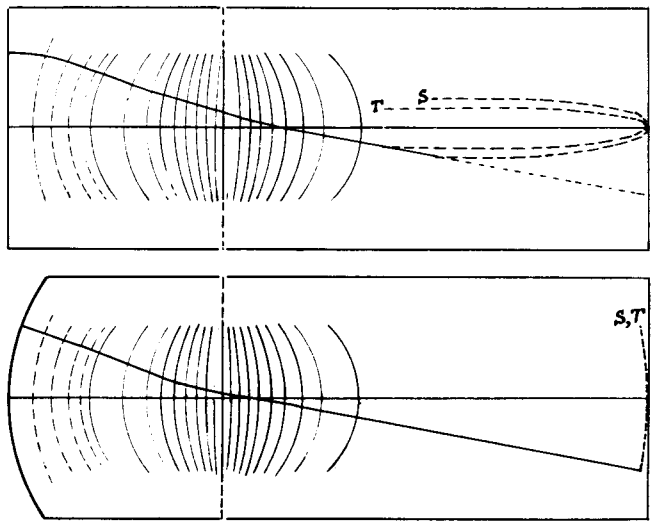


Fig. 12 - Potential distribution and image curvature in tubes with flat and curved cathode (Schematic).

strongly refracted in the first part of its passage. As a result, there are great differences in the focusing distances along it and the central ray as well as in the focusing distances of the sagittal and tangential pencils. This condition gives rise to strong curvature and astigmatism. In the curved-cathode tube all of the principal rays proceed practically normal to the first few equipotentials and are thus refracted similarly. Consequently, there are approximately equal focusing distances along them.

We now come to the chromatic aberrations of electron lenses, i.e., the image defects due to variations in the initial velocities of the electrons. Scherzer[7] was able to show quite generally that for any electric and magnetic lens combination, electron rays leaving the object were refracted less and came to a focus at a point further removed from the lens in proportion as their initial velocities are greater. Thus, achromatic lenses are impossible in electron optics. This statement does not, however, apply to systems involving electron mirrors, though even with their aid we cannot achieve true achromatism if the electrons leave the object with

[7] Loc. cit.

initial velocities ranging from zero upward. This very important case is realized wherever the object is in the form of a photoelectric, thermionic, or secondary-emission cathode. If, here, we denote the initial radial velocity in volts by V_r, and the axial velocity by V_z, the diameter of the corresponding circle of aberration is given by

$$\Delta = \left(\frac{V_r}{U}\right)^{\frac{1}{2}} \left[a\left(\frac{V_z}{U}\right)^{\frac{1}{2}} + b\,\frac{V_z}{U} + \cdots \right]$$

where a and b are constants and U is the overall applied potential. If the electron-optical system contains a mirror, b may take on negative values of such magnitude that Δ will vanish for any one particular value of V_z — not for all values of V_z, provided they are sufficiently small, as true achromatism would require. However, the improvement that can be gained in this respect by adding a mirror is very considerable.

Before considering this improvement from a quantitative point of view, however, it is well to introduce some simple formulas derived for simple types of image tubes by Henneberg and Recknagel.[14] If the initial volt velocity of the electrons is given by V, the distance between the object and image by L, and the field at the cathode by F, we have for the diameter of the circle of diffusion to a first approximation (see Fig. 13):

1) For a simple accelerating field

$$\Delta = 4\,L\left(\frac{V}{U}\right)^{\frac{1}{2}} \qquad (37)$$

2) For a uniform magnetic field superposed on a uniform electric field

$$\Delta = 2\,L\,\frac{V}{U} = \frac{2V}{F} \qquad (38)$$

3) For the image tube with electric or magnetic lens

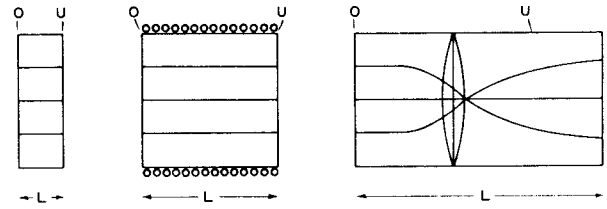
$$\Delta = 2\,m\,\frac{V}{F} \qquad (m = \text{magnification}) \qquad (39)$$

For the short magnetic lens with the whole field applied between the cathode and the lens, equation (39) becomes

$$\Delta = \frac{2m}{2m+1}\,L\,\frac{V}{U} \qquad (40)$$

[14] W. Henneberg and A. Recknagel, "Der Chromatische Fehler bei Elektronenoptischen Anordnungen Insbesonder beim Bildwandler," Zeits. f. techn. Phys., Vol. 16, No.8, p. 230; 1935.

To show the relative magnitude of the chromatic aberration as calculated for an image tube with mirror, an uncorrected electrostatic image tube, and an image tube with short magnetic lens with the same object distance, we may write down a



(a) SIMPLE ACCELERATING FIELD
(b) UNIFORM ELECTRIC AND MAGNETIC FIELD
(c) IMAGE TUBE

Fig. 13

few numerical values of the diameter of the circle of diffusion as shown in Table I.

The image tube with mirror is represented in Fig. 14. A transverse magnetic field directs the electron rays from the photocathode neck into the mirror and from here to the screen. The potential of the mirror electrode U_m is adjusted for every focusing voltage U_1 so that the photocathode remains focused on the screen. It is seen that, in general, very small changes in U_m are sufficient for this.

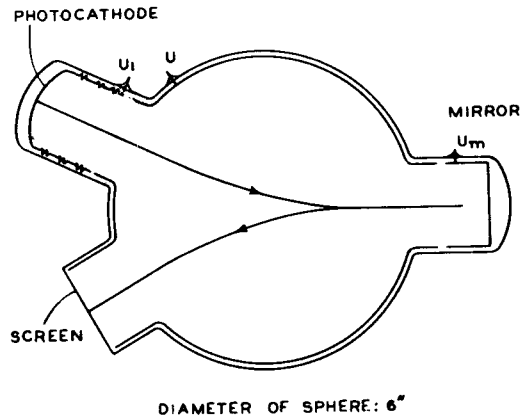


Fig. 14 – Image tube with mirror.

The table shows that, for favorable adjustment of the potentials (U_1 = 1000 volts), the chromatic aberration for electrostatic focusing together with an electron mirror may be as low or even lower in the desired range than that of the most favorable magnetic focusing arrangement. Without a mirror the magnetic image tube gives very much less chromatic aberration than the electrostatic tube operated in the usual manner

Table I

Initial Electron Velocity V Volts	DIAMETER OF THE CIRCLE OF DIFFUSION (IN INCHES) FOR					
	Image Tube With Mirror				Uncorrected Electrostatic Image Tube	Image Tube With Short Magnetic Lens
	Focusing Volts U_1 =	4000	2000	1000	600	
	Mirror Electrode Volts U_m =	-1475	-1475	-1495	-1525	
0.5		.0006	.0002	.0008	.0015	.0028	.0005
1.0		.0022	.0010	.0013	.0027	.0054	.0009
2.0		.0068	.0039	.0021	.0050	.0109	.0018
3.0		.0130	.0081	.0026	.0070	.0164	.0027
5.0		.0292	.0195	.0035	.0103	.0274	.0045

with direct focusing, as in the latter the initial fields are much weaker.

Henneberg and Recknagel also considered the effect of the initial velocities off the axis. For the short magnetic lens, in particular, they found the simple expression

$$\Delta' = r_o(m + 1)\frac{V}{U} \qquad (41)$$

for the deviation of the intersection of the principal ray in the image plane from that corresponding to zero initial velocities. As $r_o m$, the radius vector to the image point, is generally much less than L, the distance between the object and the image, $\Delta' \ll \Delta$ and can usually be neglected.

It may be remarked that in the above considerations on the chromatic aberrations of image tubes we have, in effect, treated chromatic and spherical aberration together, as the spread in initial velocities here determines, at the same time, the aperture of the imaging pencils.

Lecture 18

RADIO RECEIVING TUBE COMPONENTS AND THEIR MANUFACTURE

N. R. Smith

The design of radio tubes (particularly receiving types) is subject to the same restrictions that govern the design of any appliance. The external size and shape of the tube are more often determined by the equipment in which the tube is to be used than by requirements which would facilitate manufacturing conditions. Of course, when a tube has been placed on the market these dimensions must be maintained. Just how accurately these dimensions must be maintained was demonstrated during the conversion of the "S" to the "ST" bulbs in 1932 when a difference of as little as 1° in slope of the side of the bulb and a 1/8" increase in bulb-neck diameter caused difficulty in replacements in certain installations.

ENCLOSURES

Prior to the development of the metal tubes, the types of enclosures for production-type receiving tubes could be classified into five separate groups varying in bulb size and shape, The metal tubes produced five new enclosures which were shortly increased by revisions made to the glass tubes so as to make them interchangeable with the metal types. Following these additions, new tubes were designed in glass to approximate the size of the metal tube which further increased the types of enclosures until at the present time there are between fifteen and twenty separate types of enclosures required for receiving tubes. Duplication of tubes of a given type in various enclosures divides the production on the type and increases the burden on a factory which has to adjust its equipment to accommodate each type of enclosure. Such a condition, naturally, produces an undesirable manufacturing situation which must be relieved from one of the following sources:

1) A reduction in cost of a line of tubes in one of the existing types of enclosures sufficient to eliminate competition from all other types.

2) The design of mounts which can be used interchangeably with the various types of enclosures.

3) A new line of tubes so improved in design and reduced in cost as to eliminate competition from existing types.

Regardless of what the solution to this problem will finally be, it is apparent that structural design will be the most essential factor in obtaining the answer.

BASES

The majority of bases used on radio tubes today consist of a plastic material molded under heat and pressure to the desired shapes. The material used must have a low power factor at high frequencies, must stand the operating temperature of the tube, and must not be affected adversely by the presence of moisture. The material used in radio-tube bases will stand approximately 170°C before softening or blistering, and is non-hygroscopic.

The contact pins are of two varieties; one type is drawn from flat brass stock into the long tubular section by means of an eyelet machine, and is designated as a solid or drawn pin; the second type is made from flat brass stock formed into a tubular shape and swaged to size and is known as a seamed pin.

Other bases which are used for special applications have a metal shell with a ceramic or mica composition insert. These bases are considerably more expensive and are used only when an application demands.

In the early manufacture of plastic composition bases, the contact pins were molded into the shell. Now it is almost universal practice to mold the shell and stake the pins into the base in a subsequent operation.

It is interesting to note in the history of the development of the radio-tube base, that the initial bases were made of a plastic material which was replaced during the early production of tubes in 1916 by the brass-shell, ceramic-insert base and did not come into general use again until the fall of 1924.

The early bases had four short pins, the ends of which made contact with the socket prongs. These bases also had a pin on the side of the base shell by means of which the tube was held in the socket. The short pins were replaced by longer ones which fitted into sleeve contacts in the socket. These side contacts provided more positive connections, and eliminated the need for the side or bayonet pin. These long-pin bases were known as "UX" bases, and were used universally for receiving tubes until 1935 when the present octal base was developed for metal tubes.

Recently, tubes have been designed without bases. The external contacts are provided by 0.050"-diameter leads sealed into a glass-button stem.

BULBS

Glass bulbs can be obtained in a great variety of shapes; however, for any given shape, the bulbs are usually designed in accordance with empirical formulas which assure proper glass distribution and correct proportions.

Bulbs are designated by a descriptive nomenclature consisting of one or more letters indicating the bulb shape followed by one or more

digits expressing the maximum diameter of the bulb in eighths of an inch followed by a letter and numeral designating the shape of the neck and cullet, respectively. Thus, the designation ST12D1 indicates a combination "S" and "T" bulb 12/8" or 1-1/2" in diameter. All bulb dimensions are outside or mold dimensions.

When a bulb is designed, the diameters are made concentric with a vertical axis $\overline{AB}$ and all vertical dimensions are referred to a plane $\overline{CD}$ normal to the $\overline{AB}$ axis and cutting the bulb at the point where the side wall meets the neck.

The choice as to size of bulb to use for a given application is governed by the heat dissipation expected from the tube. Although no fixed value has been established for soft glass, experience indicates that approximately 3 watts per square inch of bulb surface above the $\overline{CD}$ line is satisfactory. However, a value less than this is generally used in order to keep the operating temperature of the stem within safer values. For metal tubes, a value of 2.5 watts per square inch of shell surface has been used as a maximum value in order to keep the shell temperature below 200°C and protect the painted surface.

The method used for blowing a glass bulb depends on the production required, the size, and the shape. Small bulbs may be molded from tubing. Bulbs made in this manner usually are fairly uniform in cross-section, since the glass distribution is controlled in the manufacture of the tubing. A small quantity of special bulbs is usually hand-blown especially in those sizes which cannot be obtained from tubing. Bulbs which are required in large quantities are blown either on the Westlake machine or the Ribbon machine. In the operation of the Westlake machine, a mass of molten glass is transferred to the mold revolving with the turret and the bulb is blown to shape. In the operation of the Ribbon machine, molten glass flows from a furnace continuously under constant temperature and pressure between rolls which form the molten glass into a ribbon. Blanks are then stamped from this ribbon and blown into molds. As many as 500 bulbs per minute can be obtained from this machine.

STEMS

Four types of stems have been used in the production of radio tubes. The glass-flare stem (Fig. 1) consists of a piece of glass tubing with one end formed into a flare, and the opposite end pinched to form the seal containing the leads. Until a short time ago, the use of gauge 41 or larger tubing to produce a stem of this character shorter than 29 millimeters was considered impractical; however, stems as short as 14 millimeters have been used successfully in production. The factors which combine to fix the length of a flare are the spread of the leads which determine the diameter of tubing required and the length of the press. The length of the press has been reduced on the 14-millimeter stems from 10 to 6 millimeters. The flare diameter is governed by the size of the base used. The total clearance usually allowed between the flare and the inside walls of the base is 5 millimeters. This permits sufficient room for cement. An exception to this rule is found in the GT tubes, in which the bulb is seated in the bottom of the base. In this case not so much cement is necessary, and since it is desirable to use as large a flare as possible, the clearance has been reduced to 2 millimeters. The smallest flare diameter should always exceed the tubing diameter by at least 3 millimeters, and if the flare is short, this ratio should be increased in order to permit satisfactory sealing and eliminate danger of stuck seals.

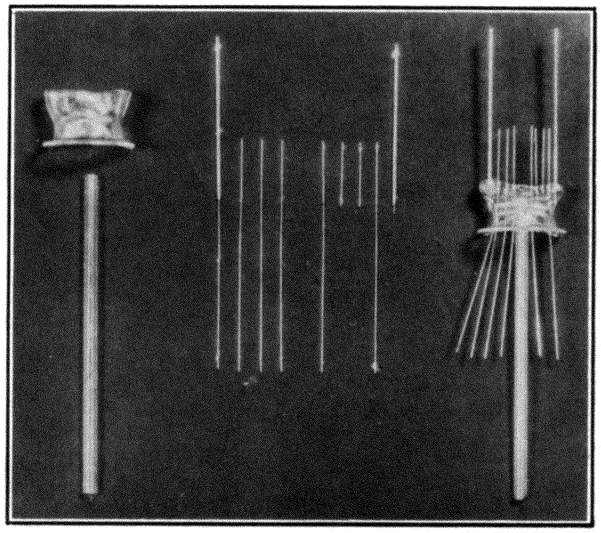


Fig. 1 - Glass-flare stem.

The clearance between leads in a flat-press stem should permit a minimum of 0.030" of glass. This spacing is determined by the diameter of the inner lead or weld knot, whichever is the larger. In many instances, of course, this spacing must be increased to eliminate possible difficulty from electrolysis.

In the metal-tube stem shown in Fig. 2a, each lead is brought into the tube through an individual metal eyelet, and made vacuum tight by means of a glass seal. The amount of assembly work necessary on a stem of this type with the consequent cost has caused it to be abandoned in favor of the metal-tube button stem shown in Fig. 2b.

The metal-tube button stem consists of a glass button molded into a metal sleeve. In the fabrication of the stem, a chrome-iron sleeve having approximately the same coefficient of expansion as the glass is welded to a cold-rolled steel flange. This assembly is loaded onto a mold in which the lead wires have been located; and one short length of glass tubing is placed inside the lead wires and one outside. The whole assembly is heated and the glass pressed into the metal sleeve. The cooling of the button is carefully

controlled in order to obtain the correct compression strain in the finished part.

The fourth stem (shown in Fig. 3) is the all-glass button stem, and is fabricated in much the same manner as the metal-tube button stem, except that the metal sleeve is omitted and the glass is shaped by the mold.

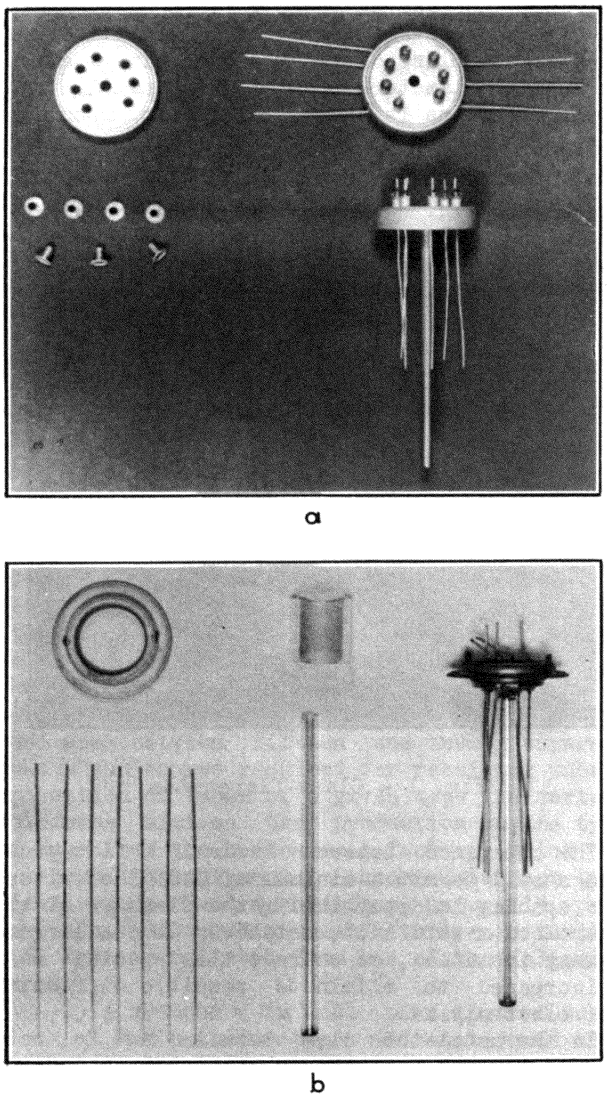


Fig. 2 - (a) Metal-tube stem with metal eyelets.
(b) Metal-tube button stem.

The principal advantage of the glass button stem over the flared stem is that it offers an opportunity to reduce the overall length of the tube slightly, and improves the shielding conditions which are rather important in single-ended designs.

FILAMENT COATING

There are three methods available for coating filaments. The most economical procedure to date is known as drag coating. Under this method the filament base wire or ribbon is continuously coated by passing the wire alternately over grooved wheels running in baths of coating material, then through electric furnaces to bake the coating on, and continuing this process until the proper amount of coating has been applied. After coating, the wire is respooled and stored until required for forming and use in tubes.

Formed filaments which are sufficiently rugged may be sprayed. This method is slower than the drag-coating process, but is advantageous

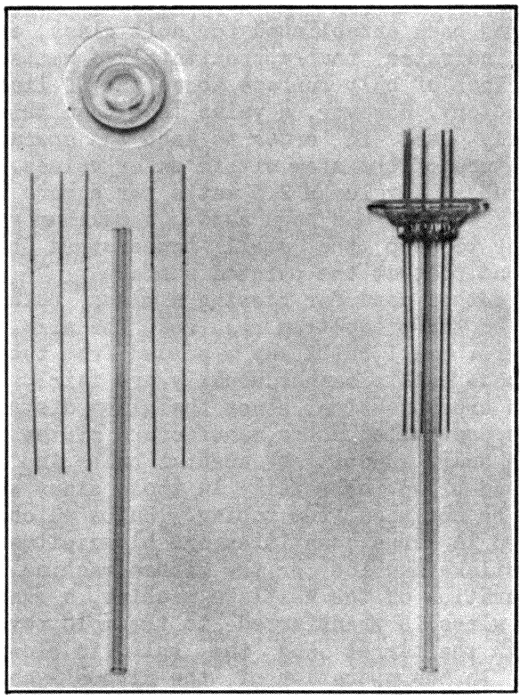


Fig. 3 - All-glass button stem.

inasmuch as it permits forming the base wire before the coating is applied.

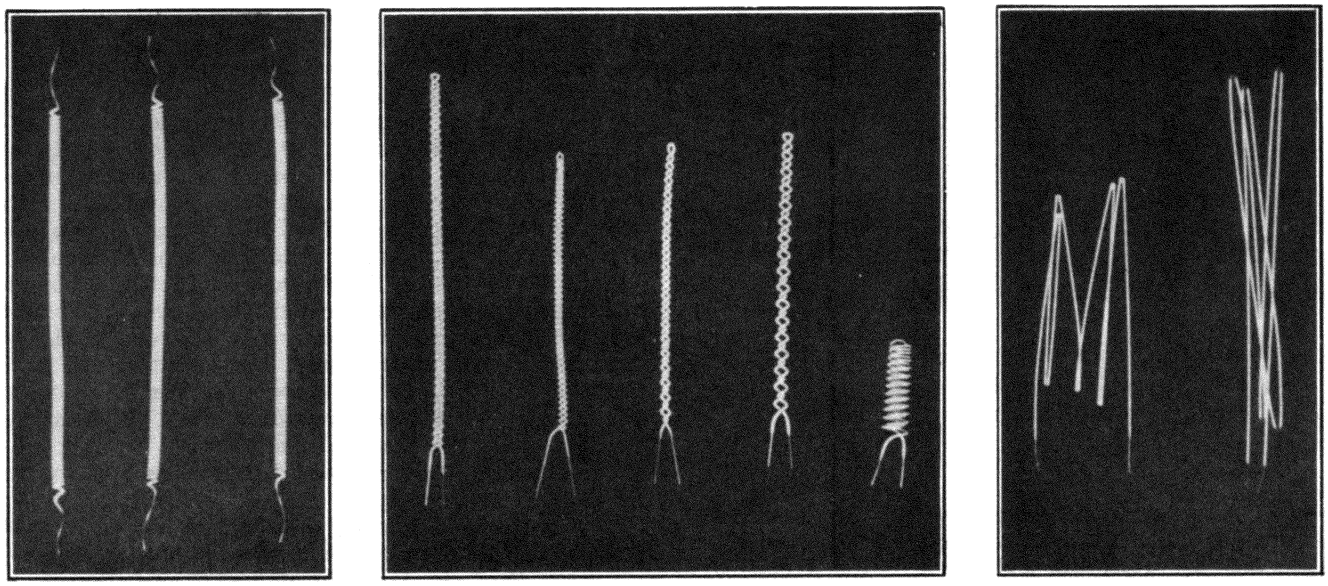
In a third method, which has not proved of sufficient advantage to have more than a limited application, the coating is applied by depositing it electrically on the wire. This can be done as a continuous process or applied after forming the filament.

HEATERS

Heaters used in receiving tubes can be grouped into two general classes, coil heaters and straight-wire folded heaters, as shown in Fig. 4. Of the various coil heaters, the double helix is the most popular and is used in the majority of the applications because of its superior electrical characteristics, its ease of fabrication, and its convenience in tube assembly. Other coil heaters which are used are the straight-wound

coil and the folded coil heaters. All coil heaters are sprayed with an insulating material. This operation involves hand-racking, either hand or automatic spraying, and baking at high temperature to cure the insulation material.

The straight-wire folded heaters have the advantage that the wire can be continuously coated and baked. This procedure is much faster than the spraying method required for coil-type heaters. However, forming and assembly operations for the straight-wire folded heater at the present time, are sufficiently more difficult so that the gain made through continuous coating is for the most part lost, and the net advantage of this heater when compared with the double helix is doubtful, except for special applications. Where wide flat cathodes are used, the folded heater may offer better heat distribution and some gain in the strength of the assembly.



Straight-wound coil heaters | Double-helix coil heaters | Straight-wire folded heaters

Fig. 4 - Types of heaters.

CATHODES

Of the two forms of cathodes used in the manufacture of receiving tubes, the lock-seam cathode serves for the majority of the applications, due to the fact that it can be conveniently formed from flat stock to the various shapes desired at a more reasonable cost than can be obtained with the use of seamless tubing. Seamless tubing is obtained by either an electrolytic or an extrusion method, and is received in random lengths which must be cut to finished size in a lathe. Embossing or forming the cathode to other than its original cylindrical shape is a difficult and comparatively slow operation when compared with the forming operation for the lock-seam cathode.

Although nickel is the basic material used for all cathodes, various percentages of other elements are alloyed with it to either increase the strength of the cathode or to improve the emissive qualities.

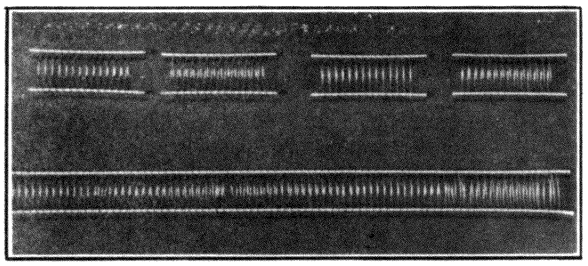
The emissive coating is applied in all cases by racking the cathodes and spraying. The density, weight, and character of the coating is controlled by the number of passes, the distance of the rack from the gun, and the density of the spray mixture.

GRIDS

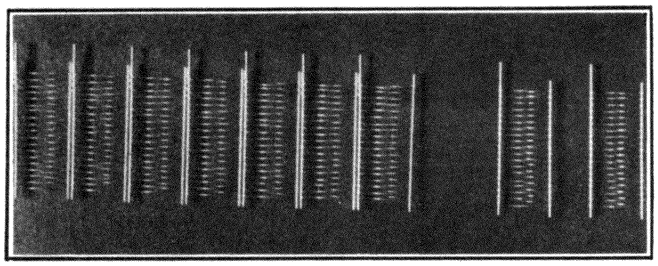
The two types of grids used in production are shown in Fig. 5a and b. The wound grid, shown in Fig. 5a, has been standard since the earliest manufacture of tubes, and is formed in a long strip on a lathe by winding the grid wire around a mandrel through notches cut in the side-rods. The notches are then closed or peened over the grid wire to hold the turns in place. Depending on the length of the finished grid, 6 to 10 grids can be obtained from a strip. The chief items of criticism of a wound grid are: notching of the side-rods reduces the strength of the grid, and creates a rough, uneven surface difficult to insert into the mica spacers; and the displacement of the metal during notching causes the side-rods to bow, especially on grids having small diameter side-rods.

The ladder type of grid shown in Fig. 5b is made on a machine which has recently been developed. In this method, the side-rod wires and the grid wires are fed into a welding machine at right

angles to each other. Each grid wire is unreeled from a separate spool and for proper spacing is threaded through the teeth of a comb. Two sets of wires and two sets of combs are used to make a complete grid with the grid wires attached to opposite sides of the side-rods. As all the grid wires are pulled simultaneously over the side-



a



b

Fig. 5 - (a) Wound grid.
(b) Ladder-type grid.

rods, two welding electrodes draw the wires over a pair of shaping mandrels and press the wires into contact with the side-rods. The wires are then simultaneously welded to the side-rods by the welding electrodes. The completed grid is cut off by shears. Grids can be made very rapidly in this manner. At present this method is most successful when applied to grids with fairly wide side-rod spacing and few turns per inch.

PLATES

Until recently, the plate in a tube was designed to serve but one purpose and that was as an anode. However, this electrode has now been pressed into service, not only to serve as an anode, but also as a means for locating and holding the whole mount assembly together. To meet this demand, plates in various shapes and forms have been produced. Ears or lugs have been provided, which protrude through the insulating spacers for location and retention of the mount parts. Although the shape and character of a lug may vary with the design of the plate, it is essential that the lug be shaped so that it automatically provides the correct location of the parts independently of the operator. It is also essential, of course, that the plates be so designed as to insure the correct dimensions being held during processing. This is most easily accomplished by embossing the active surface of the plate and so hinging this rigid section to the rest of the plate as to permit the necessary expansion during heating without seriously altering the initial plate dimensions. At present all of the plates used in receiving tubes are made from nickel or nickel-plated steel.

MICAS

Mica is the principal material used for electrode spacers in receiving tubes. It has the advantage over other available materials, in that it can be stamped to any shape and design desired. Mica can be split to any thickness, but since splitting is a hand operation, sufficient tolerance has to be allowed so as not to make the operation unduly expensive. The range used for most spacers is 0.008" to 0.015".

Since mica is purchased already graded to certain standard trimmed sizes, it is desirable to design spacers so as to permit the maximum number of pieces to be stamped from a standard size blank. Most spacers for receiving tubes are stamped from No.6 mica which is approximately a square blank having 1-1/2 square inches. Consequently, if a mica spacer were designed 1/2" wide by 3/4" long, probably only one spacer could be obtained from a No.6 blank, but if the width were reduced to 3/8", it would be possible to obtain two spacers from the same blank and so obtain a reduction in the cost of the mica.

The design factors and tolerances applicable to mica stampings are:

1) Hole alignment will be commensurate with any other die job, and can be held to ± 0.0005".

2) Minimum diameter or size of any hole should not be less than the thickness specified for the mica. Variation in hole diameter can be as little as ± 0.0002", but due to the feathery edge left after punching, measurements to this accuracy are impractical.

3) Minimum distance between adjacent openings in a mica should not be less than 0.010" - 0.012".

4) All corners should have sufficient radius, if possible, to prevent splitting of mica during punching and transportation.

SUMMARY

This short survey of component parts and manufacturing methods used in radio tubes to-day in comparison with the parts and methods used in the first production of 1920, indicates few radical changes or improvements in the mechanical design of tubes. This fact leads to one of two conclusions: either the initial design of a vertical, concentric mount is fundamentally the cheapest

and best design which can be conceived, or the possibilities to be derived from mechanical design have not been thoroughly investigated.

Recently, considerable effort has been directed by all manufacturers toward the improvement, reduction in size, and reduction in cost of the vertical, concentric mount. Much can be accomplished in this direction; however, the results of this effort have in many instances only served to complicate the manufacturing problems without attaining real advantages.

Consequently, I believe it is a safe prediction to say that the ultimate answer to the demand for smaller, lower-priced tubes will not be found in improvements which can be made on the vertical-concentric mount, but will be the result of an entirely new structural design adaptable to more economical manufacturing methods.

Lecture 19

ANALYSIS OF RECTIFIER OPERATION

O. H. Schade

INTRODUCTION

The design of diodes for rectification of a-c power requires unusual care even though their design is simple from the standpoint of the number of electrodes involved. The cathode must be capable of supplying the average current and peak current which will be encountered in service. The plate must operate at a safe temperature, which in turn depends on plate dissipation, the physical dimensions of the plate, and the thermionic and thermal emissivity of the plate material. Another factor to be considered is the perveance of the diode, i.e., the steepness of its voltage-current characteristic. Perveance affects both cathode- and plate-design because it controls current density and voltage gradient at the effective surfaces of both for a given cathode area. Then, too, there is the requirement for adequate insulation. Perveance and insulation requirements depend on the type of service in which the diode is to be used.

It is apparent that an analysis of diode-operating conditions in rectifier circuits is essential in order to arrive at optimum design specifications in each particular case, because the required data bear no simple relationship to the values of rectified current and voltage.

PRINCIPLES OF RECTIFICATION

1. General

Rectification is a process of synchronized switching. The basic rectifier circuit consists of one synchronized switch in series with a single-phase source of single frequency and a resistance load. The switch connection between load terminals and source is closed when source and load terminals have the same polarity, and is open during the time of opposite polarity. The load current consists of half-wave pulses. This simple circuit is unsuitable for most practical purposes, because it does not furnish a smooth load current.

The current may be smoothed by two methods: (a) by increasing the number of phases, and (b) by inserting reactive elements into the circuit. The phase number is limited to two for radio receivers. The circuit analysis which follows later on will treat single- and double-phase rectifier circuits with reactive circuit elements.

Switching in reactive circuits gives rise to "transients." Current and voltage can, therefore, not be computed according to steady-state methods.

The diode functions as a self-timing electronic switch. It closes the circuit when the plate becomes positive with respect to cathode and opens the circuit at the instant when the plate current becomes zero.

The diode has an internal resistance which is a function of current. When analyzing rectifier circuits, it is convenient to treat the internal resistance of the diode rectifier as an element, separated from the "switch action" of the diode. Fig. 1 illustrates the three circuit elements so obtained. The diode characteristic is the geometric sum of these characteristics. The resistance r_d is effective only when the switch is closed, i.e., during the conduction period of the diode. The effective diode resistance must, therefore, be measured or evaluated within conduction-time limits. Consider a switch in series with a fixed resistance and any number of other circuit elements connected to a battery of fixed voltage. The d-c current and rms current which flow in this circuit will depend on the time intervals during which the switch is closed and open; the resistance value is not obtainable from these current values and the battery voltage. The correct value is obtained only when the current and voltage drop in the resistance are measured during the time angle φ (Fig. 2) when the switch is closed.

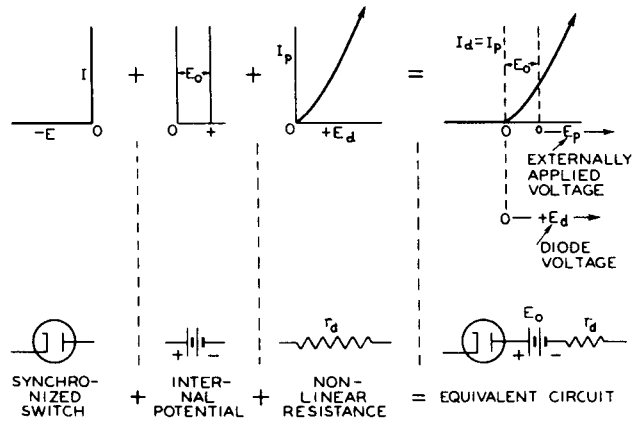


Fig. 1

The method of analysis of rectifier circuits to be discussed in this lecture is based on the principle that the non-linear effective resistance of the diode may be replaced analytically by an equivalent fixed resistance which will give a diode current equal to that obtained with the actual non-linear diode resistance. The correct value to be used for the equivalent fixed resistance depends upon whether we are analyzing for peak diode current, average diode current, or rms diode current. As will be seen later, the relations among these three equivalent resist-

ances depend upon the circuit under consideration.

2. Definitions of Resistance Values

The instantaneous resistance (r_d) of a diode is the ratio of the instantaneous plate voltage, e_d, to the instantaneous plate current, i_p, at any point on the characteristic measured from the operating point (see Fig. 1). It is expressed by the equation

$$r_d = \frac{e_d}{i_p} \qquad (1)$$

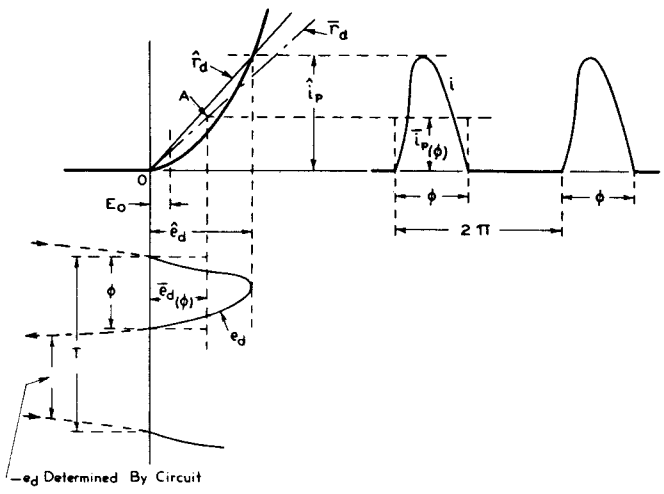


Fig. 2

The operating point (0) of a diode is a fixed point on the characteristic, marked by beginning and end of the conduction time. It is, therefore, the cut-off point $I_d = 0$ and $E_d = 0$, as shown in Fig. 1. The operating point is independent of the waveform and of the conduction time φ (see Fig. 2).

The peak resistance ($\hat{r}_d$) is a specific value of the instantaneous resistance and is defined as

$$\hat{r}_d = \frac{\hat{e}_d}{\hat{i}_p} \quad \text{(See Fig. 2)} \qquad (2)$$

Peak voltage $\hat{e}_d$ and peak current $\hat{i}_p$ are measured from the operating point 0.

The equivalent average resistance ($\bar{r}_d$) is defined on the basis of circuit performance as a resistance value determining the magnitude of the average current in the circuit. The value $\bar{r}_d$ is, therefore, the ratio of the average voltage drop $\bar{e}_{d(\varphi)}$ in the diode during conduction time to the average current $\bar{i}_{p(\varphi)}$ during conduction time, or

$$\bar{r}_d = \frac{\bar{e}_{d(\varphi)}}{\bar{i}_{p(\varphi)}} \qquad (3)$$

The curved diode characteristic is thus replaced by an equivalent linear characteristic having the slope $\bar{r}_d$ and intersecting the average point A, as shown in Fig. 2. The coordinates $\bar{e}_{d(\varphi)}$ and $\bar{i}_{p(\varphi)}$ of the average point depend on the shape of voltage and current within the time angle φ. The analysis of rectifier circuits shows that the shape of the current pulse in actual circuits varies considerably between different circuit types.

The equivalent rms resistance ($|r_d|$) is defined as the resistance in which the power loss P_d is equal to the plate dissipation of the diode when the same value of rms current $|I_d|$ flows in the resistance as in the diode circuit. It is expressed by the equation

$$|r_d| = \frac{P_d}{|I_d|^2} \qquad (4)$$

3. Measurement of Equivalent Diode Resistances

The equivalent resistance values of diodes can be measured by direct substitution under actual operating conditions. The circuit arrangement is shown in Fig. 3. Because the diode under test must be replaced as a whole by an adjustable resistance of known value, a second switch (a mercury-vapor diode identified in the figure as the ideal diode) with negligible resistance must be inserted in order to preserve the switch-action in the circuit.

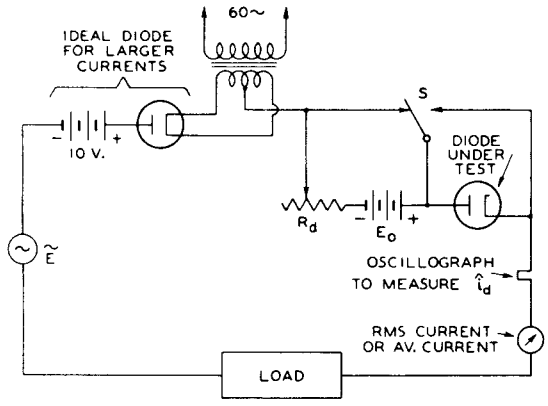


Fig. 3 - Half-wave circuit for measuring $\hat{r}_d$, $\bar{r}_d$, and $|r_d|$.

When a measurement is being made, the resistor R_d is varied until the particular voltage or current under observation remains unchanged for both positions of the switch S. We observe (1) that it is impossible to find one single value

of R_d which will duplicate conditions of the actual tube circuit, i.e., give the same values of peak, average, and rms current in the circuit; (2) that the ratio of these three "equivalent" resistance values of the diode varies for different combinations of circuit elements; and (3) that equivalent average or rms substitution resistances may have different values when the average or rms current in the diode circuit is adjusted to the same value in different types of rectifier circuits. The "peak resistance" value obtained for a given peak current is found to be substantially independent of the type of circuit.

4. Waveforms and Equivalent Resistance Ratios for Practical Circuit Calculations

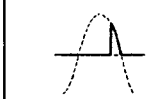
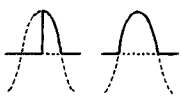
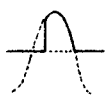
The form of the current pulse in practical rectifier circuits is determined by the power factor of the load circuit and the phase number. Practical circuits may be divided into two main groups: (a) circuits with choke-input filter; and (b) circuits with condenser-input filter.

The current pulse in choke-input circuits has a rectangular form on which is superimposed one cycle of the lowest ripple frequency. In most practical circuits, this fluctuation is small as compared with the average amplitude of the wave and may be neglected when determining the equivalent diode resistances. It is apparent then that the equivalent diode resistance values are all equal and independent of the type of diode characteristics for square waveforms. Hence, for choke-input circuits, we have

$$\hat{r}_d = \bar{r}_d = |r_d| \qquad (5)$$

The current pulse in condenser-input circuits is the geometric summation of a sine-wave section and a current having an exponential decay. It varies from a triangular form for $\varphi < 20^\circ$ to a full half-cycle ($\varphi = 180^\circ$) as the other extreme. In Table I are given the ratios of voltages, currents, and resistance values during conduction time for two principal types of rectifier characteristics: the 3/2-power-law characteristic of high-vacuum diodes, and the idealized rectangular characteristic of hot-cathode, mercury-

Table I

CONDUCTION TIME ANGLE φ	WAVE SHAPE	$\frac{\bar{i}_{p_\varphi}}{\hat{i}_p}$	$\frac{\lvert i_p \rvert_\varphi}{\hat{i}_p}$	3/2-POWER RECTIFIER CHARACTERISTIC			RECTANGULAR CHARACTERISTIC		
				$\frac{\bar{e}_{d_\varphi}}{\hat{e}_d}$	$\frac{\bar{r}_d}{\hat{r}_d}$	$\frac{\lvert \bar{r}_d \rvert}{\hat{r}_d}$	$\frac{\bar{e}_{d_\varphi}}{\hat{e}_d}$	$\frac{\bar{r}_d}{\hat{r}_d}$	$\frac{\lvert \bar{r}_d \rvert}{\hat{r}_d}$
CONDENSER-INPUT CIRCUITS									
$\leq 20^\circ$		0.500	0.577	0.593	1.185	1.120	1.0	2.00	1.500
90° & 180°		0.637	0.707	0.715	1.120	1.057	1.0	1.57	1.272
130°		0.725	0.780	0.787	1.085	1.030	1.0	1.38	1.190
CHOKE-INPUT CIRCUITS									
180°		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

vapor diodes. In this table, the designation $|i_p|_\varphi$ represents the rms value of the current during the conduction time.

It follows that the relation

$$\hat{r}_d = 0.88\ \bar{r}_d = 0.93\ |r_d| \qquad (6)$$

is representative for the group of condenser-input circuits containing high-vacuum diodes, and holds within ± 5 per cent over the entire range of variation in wave shape. The actual error in circuit calculations is smaller, as the diode resistance is only part of the total series resistance in the circuit.

EMISSION AND SATURATION OF OXIDE-COATED CATHODES

1. Oxide-Coating Considerations

The normal operating range of diodes (including instantaneous peak values) is below the saturation potential because the plate dissipation rises rapidly to dangerous values if this potential is exceeded. Saturation is definitely recognized in diodes with tungsten or thoriated-tungsten cathodes as it does not depend on the time of measurement, provided the plate dissipation is not excessive. The characteristic of such diodes is single-valued even in the saturated range, i.e., the range in which the same value of current is obtained at a given voltage whether the voltage has been increased or decreased to the particular value.

Oxide-coated cathodes have single-valued characteristics up to saturation values only under certain conditions. The cathode coating has resistance and capacitance, both of which are a function of temperature, current, and the degree of "activation," and vary during the life of the tube. The conception of a "coating impedance" is essential in explaining many peculiar effects observed when testing and operating oxide-coated diodes in rectifier circuits and will be treated briefly.

The cathode coating is usually a mixture of carbonates (barium, strontium, calcium) which are "broken down" into oxides, metal atoms, and gases during the exhaust process. A highly emitting monatomic layer of barium on oxygen is formed on the surface of the coating, which, when heated, supplies the electron cloud forming the space charge above the coating surface (see Fig. 4). The coating itself consists of non-conducting oxide clumps or crystals (shown as shaded areas) interposed with metal atoms and ions (circles), which are produced by high cathode temperature and electrolysis during the activation and aging process of the cathode.

The distances between these metal atoms or atom groups are very small. The current flow from the base material through the coating is effected by relay-emission from atom to atom or particle to particle under the influence of electrostatic potentials. The coating is not conducting when cold. Electronic conduction through the coating is high when many metal relay chains, not broken by insulating oxides, have been formed,

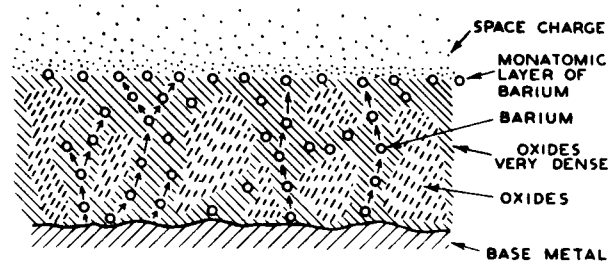


Fig. 4 - Structure of cathode coating.

and also when the electron emission is raised as a result of loosening the atomic structure by increased temperatures. Hence, the coating conductance may be represented by a large number of extremely close-spaced diodes in series-parallel arrangement, as shown in Fig. 4a.

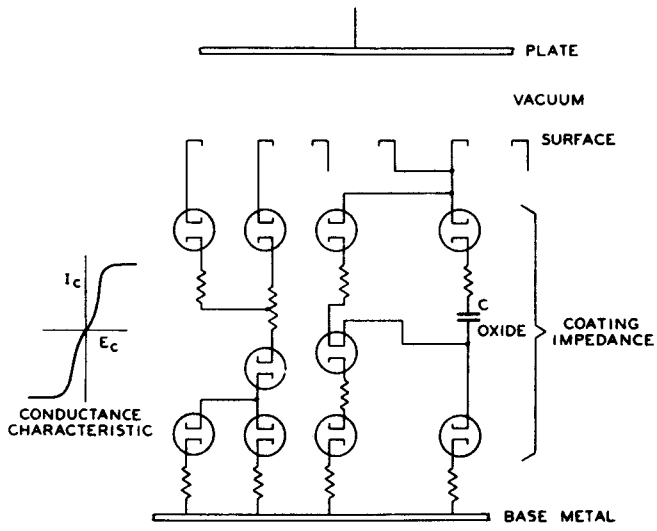


Fig. 4a - Representative diode network.

The coating activation is good when a fully emitting surface layer can be saturated without limitation of electron flow in the coating. This condition is indicated by a single-valued characteristic up to and beyond the saturation potential as shown in Fig. 5. In general, single-valued characteristics are obtained if the surface saturates before the coating conductance becomes limited. (See temperature-limited conditions in Fig. 5).

The conductance of the coating may, however, limit the electron flow before the surface emission is saturated, and cause a peculiar voltage-current characteristic. Consider a high plate

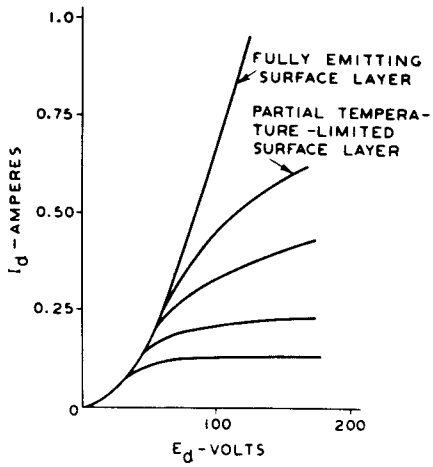


Fig. 5 - Single-valued saturation characteristic of diode with high coating conductance.

voltage suddenly applied over a periodic switch to a diode as in the circuit of Fig. 6. If the coating is not limiting, the current obtained is that at a point P on the corresponding diode characteristic. Hence, the current waveform in the circuit is as shown in Fig. 6a. If the surface emission is assumed to be unchanged, but

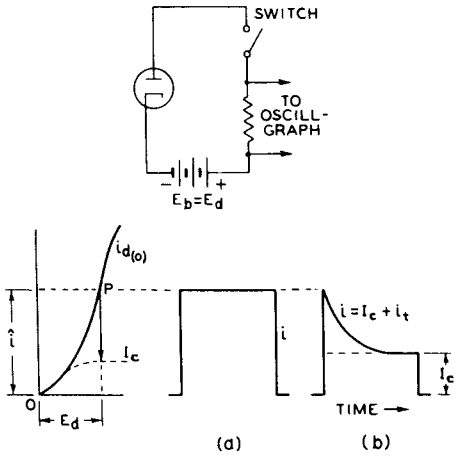


Fig. 6

the coating conductance is limited, due to an insufficient number of "coating-diodes" and too many non-conducting oxide groups, the waveform of Fig. 6b is obtained. At the instant after closing the switch, the current value $\hat{I}$ is demanded by E_d from the surface layer; the conduction current in the coating, however, is limited to the value I_c by saturation of the "coating-diodes." There are, however, many free electrons in the coating which can be moved under the influence of the electric field to the surfaces of the insulating oxides, but cannot move throughout the entire coating. Hence, a displacement current can flow, and this charges up the oxides until all free electrons have been moved. The oxide groups thus act like condensers in parallel to the coating resistance but with the peculiarity that their charge may be limited by hypothetical series diodes.

The coating resistance is extremely low* below saturation, due to the small spacing and high gradient in the "coating-diodes" but becomes infinite when the conduction current is saturated; hence, the charging current must flow in the plate circuit (external) of the diode. The total plate current is, therefore, the sum of the conduction current I_c and the "transient" electron displacement current. The "coating transient" decays to zero like normal transients at a rate depending on the actual shunt conductance value and the total series resistance in the circuit (Fig. 6b). The decay can be changed by adding external resistance in the plate circuit. When the surface emission is good, i.e., as long as the total vacuum-space plate current is space-charge limited, the current will rise initially to the value (point P) determined by the applied potential, but will then decay to the saturation value determined by the coating conductance.

The condition of oxide-coated cathodes can, therefore, not be judged alone by their capability of furnishing high peak currents, but the rate of change in current flow, and hence the current waveform, must also be carefully considered, because the diode characteristic may not be single-valued. Fig. 7 shows characteristics which are not single-valued. These were taken within 1/120 of a second (1/2 of a 60-cycle sine wave) with a cathode-ray curve tracer.

2. Current Overload and Sputter

The degree of activation is not stable during the life of the cathode. Coating conductance and surface emission change. Factors affecting the change are the coating substances, the evaporation of barium depending on the base material, and the operating conditions to which the cathode is subjected. This life history of the cathode is the basis on which current ratings are established. Rectifier tubes especially are subject to severe operating conditions. If a diode is operated with too high a current in a rectifier circuit and its surface emission is decreased to the saturation value, then the tube voltage drop will increase rapidly, and cause excessive plate dissipation and destruction of the tube. Should the coating conductance in this diode decrease to a value which limits the demanded current, power is dissipated in the now-saturated "coating-diodes," with the result that the coating voltage drop and coating temperature are raised. The temperature rise may cause reactivation but also may become cumulative and melt the coating mate-

* Its magnitude depends on the number of series diodes and, hence, on the barium content and thickness of the coating.

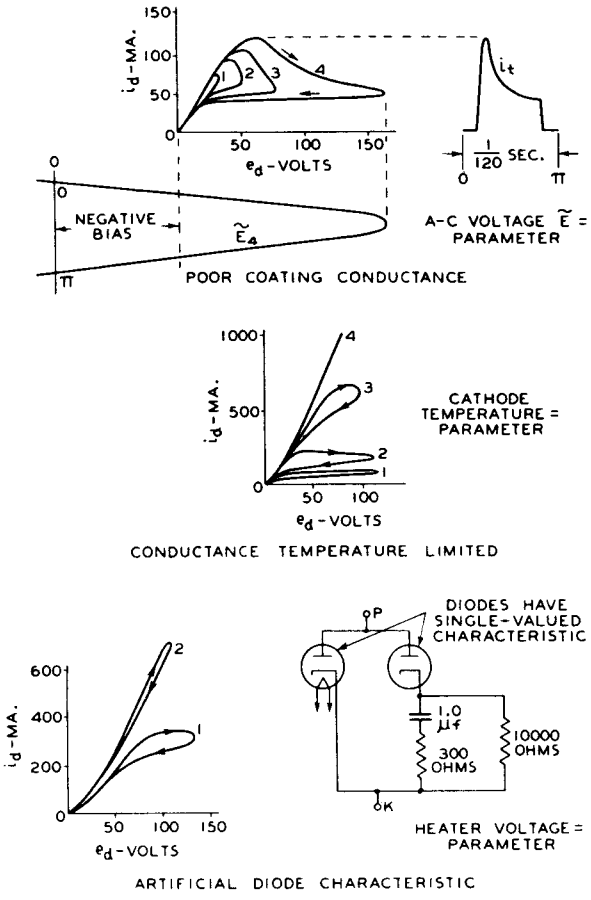


Fig. 7

rial. Vapor or gas discharges may result from saturated coatings and cause breakdowns during the inverse voltage cycle. These breakdowns are known as "sputter," and destroy the cathode. Sputter of this type will occur more readily in diodes with long cylindrical electrodes and close spacing than in diodes with flat cathodes of narrow width and parallel-plane plates, because the activation gas can diffuse faster in the latter.

A second type of sputter is caused by the intense electrostatic field to which projecting "high spots" on the plate or cathode are subjected. These high spots are formed by loose carbon on the plate or large cathode-coating particles. The resulting current concentration causes these spots to vaporize with the result that an arc may be started.

3. Hot-Cathode Mercury-Vapor Diodes

The ionization potential E_i of mercury vapor is a function of the gas pressure and temperature. It is approximately 10 volts in the RCA-83 and similar tubes. A small electron current begins to flow at $E_p = 0$ (see Fig. 8), and causes ionization of the mercury vapor at $E_p = E_i$. This action decreases the variational diode resistance r_p to a very low value. The ionization becomes cumulative at a certain current value ($r_p = 0$ at 40 milliamperes in Fig. 8a), and causes a discontinuity in the characteristic. Hence, it is not single-valued within a certain voltage range. Beyond this range (see Fig. 8b), the slope (r_p) of the characteristic becomes again positive until saturation of the emitter is reached.

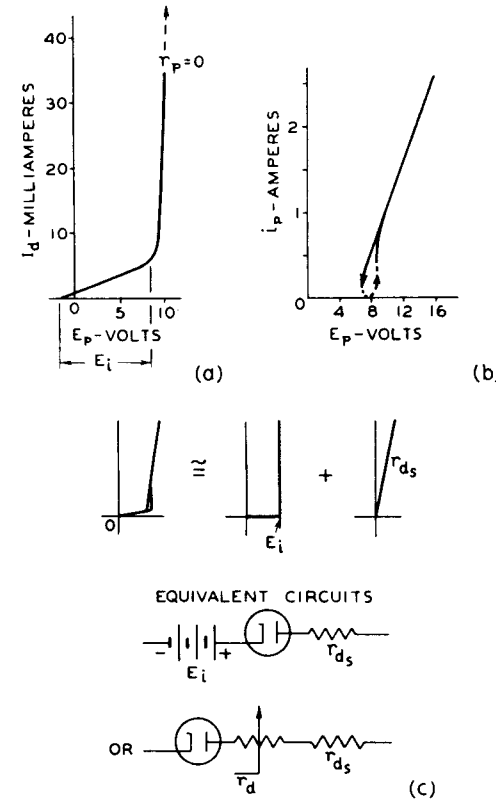


Fig. 8

For circuit analysis, the mercury-vapor diode may be replaced by a bucking battery having the voltage E_i and a fixed resistance,

$$r_{d_s} \cong \frac{\hat{e}_d - E_i}{\hat{i}}$$

as shown in Fig. 8c; or the diode characteristic may be replaced by an ideal rectangular characteristic and its equivalent resistance values and the series resistance r_{d_s} as shown.

The first representation is advantageous when making approximate calculations, as the equivalent diode resistance is then the substantially fixed value r_{d_s}. The value r_{d_s} is in the order of 4 ohms. The low series resistance and the small constant voltage drop E_i are distinct advantages for choke-input filters, as they cause very good regulation; the low resistance, however, will give rise to enormously high starting transients in condenser-input circuits, in case

all other series resistances are also small. Mercury-vapor diodes as well as high-perveance (close-spaced), high-vacuum diodes having oxide cathodes should, therefore, be protected against transient-current overloads when they are started in low-resistance circuits to prevent destruction of the cathode coating. The destruction of the coating in mercury-vapor diodes is apparently caused by concentration of current to small sections of the coating surface and not by heat dissipation in the coating.

CIRCUIT ANALYSIS

The rectifier diode is a switch operated in synchronism with the applied a-c frequency. The diode alternately opens and closes the circuit. This action causes a series of transients in reactive circuits. According to the decay time of the transients, fundamental rectifier circuits may be classified into two principal groups: (1) circuits with repeating transients in which the energy stored in reactive elements decreases to zero between conduction periods of the diode; and (2) circuits with chain transients in which (a) the magnetic energy stored in the inductance of the circuit remains above zero value, and (b) the electric energy stored in the capacitance of the circuit remains above zero value. The much used "choke-input" and "condenser-input" circuits fall under the second group.

The complete analysis of rectifier-circuit operation requires more time than this lecture provides. It is based on the fact that the total current in a circuit is the sum of all steady-state currents and transient currents within the time between two switching operations. We will analyze briefly the operation in two important circuits, i.e., the full-wave, choke-input circuit, and the condenser-input circuit.

1. The Full-Wave Choke-Input Circuit

Circuit and operation are shown in Fig. 9. The construction is made by considering first one of the diodes shorted to obtain the phase relation of the a-c voltage $\tilde{e}$, and the steady-state current $\tilde{i}_s$, as shown. If we assume that the diode closes the circuit at the time $\tilde{e} = 0$, a transient i_t with the initial value $i_{t_0} = -\tilde{i}_{s(0)}$ will flow in the circuit. The total current, i, is the sum of the currents $\tilde{i}_s + i_t$. It starts, therefore, at zero and rises as shown until the second switching operation occurs at the commutation time $t = \pi$ when the second diode receives a positive plate voltage. The total current, i, after

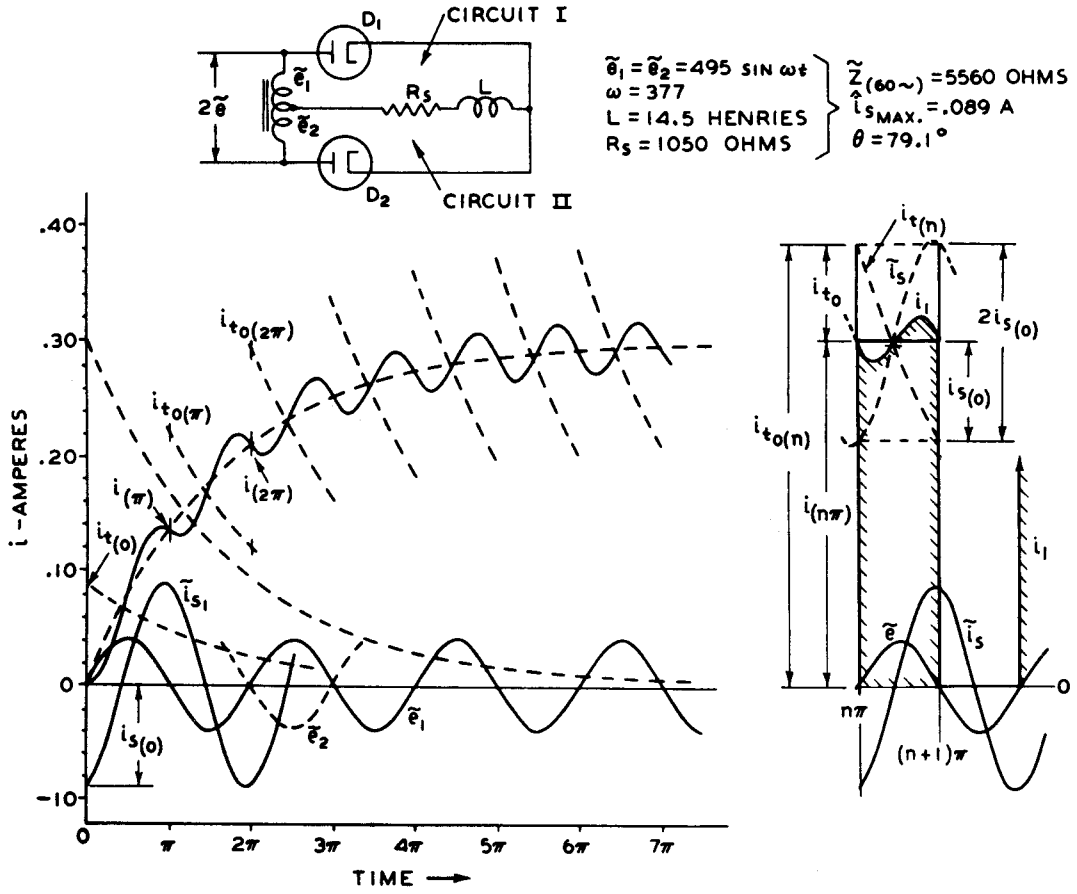


Fig. 9 - Full-wave choke-input circuit and its operation.

$t = \pi$ is again the sum of currents $\tilde{i}_s + i_t$, but the initial value i_{t_o} is increased by the value $i_{(\pi)}$ still flowing in the circuit.

The current i_{t_o} increases, therefore, at every new switching time until the decay of the transient $i_{t(n)}$, during the time $t = \pi$, is numerically equal to the steady-state current rise $2\tilde{i}_{s(o)}$, as shown in Fig. 9. This leads to the equation for the final operating current at the n^{th} commutation time:

$$i_{n(\pi)} = -\tilde{i}_{s(o)} \frac{1 + \varepsilon^{-\frac{R}{2fL}}}{1 - \varepsilon^{-\frac{R}{2fL}}} \qquad (7)$$

A broken line is shown connecting all commutation-current values. This line represents closely the average current $\bar{I}$ in the common circuit branch. The final average current $\bar{I}_\varphi$ during the conduction time, and the average current $\bar{I}$ in the load resistance $\bar{R}$, have the same value and are given by equation (7). The average plate current per diode is $\bar{I}_p = 0.5\ \bar{I}$, since each diode conducts alternately, and passes a current pulse shown by the shaded area in Fig. 9. With the numerical values of the construction substituted in equation (7) and setting

$$\varepsilon^{-\frac{R}{2fL}} = 0.547$$

we obtain

$$i_{(n\pi)} = \bar{I} = 0.298 \text{ ampere.}$$

The described solution is illustrative in understanding starting and operating conditions of the choke-input circuit, but its accuracy is limited in cases of low power factor (large L or small R), since the decay terms in equation (7) become very small. The construction, however, shows the exact phase relations of current and voltage. The oscillogram in Fig. 10 was taken with the circuit values indicated in Fig. 9.

a. Critical Inductance — The analysis has shown that the transient currents in the particular case never decay to zero and are "chain transients." In practical circuits, the d-c load resistance $\bar{R}$ is shunted by a capacitance C. This does not affect the circuit performance, provided $\bar{R}$ does not increase beyond a critical value $\bar{R}_{cr}$. Expressing this statement symbolically, we have

$$\bar{R} \leqq \bar{R}_{cr} = 1.5\ Z_{(2f)} \qquad (8)$$

in which $Z_{(2f)}$ is the impedance of one branch circuit at double line-frequency. For large values of C,

$$Z_{(2f)} \cong 4\pi fL$$

and hence

$$\bar{R} \leqq 6\pi fL \quad \text{or} \quad L \geqq \frac{\bar{R}}{6\pi f}$$

which can be converted into the form

$$L \geqq \frac{0.91\ |\tilde{E}|}{6\pi f\ \bar{I}} \qquad (9)$$

where

$|\tilde{E}|$ = rms voltage of one-half the secondary winding

and

$\bar{I}$ = d-c load current.

The relation of equation (9) is characterized by the fact that the current i in the final operating condition (Fig. 9) has an a-c component which is so large that i becomes zero at one instant. Any further increase of this component caused by

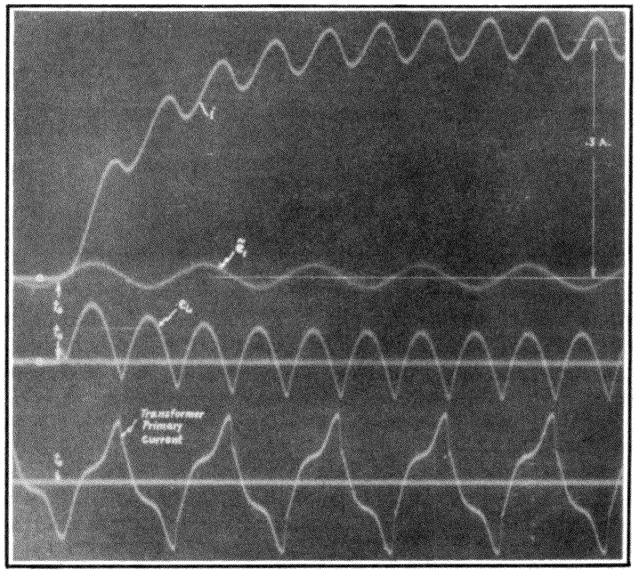


Fig. 10 - Oscillogram taken with circuit values in Fig. 9.

decreasing L still further will, therefore, open the diode circuit and hence interrupt chain-current operation. The circuit then becomes a type of condenser-input circuit and the output voltage $\bar{E}$ rises beyond the value given by the equivalent circuit for the chain-current operating range.

b. Equivalent Circuit for the Chain-Current Operating Range — Commutation of the cur-

rent in the chain-current range occurs at the instant when the supply voltage passes through zero. The voltage energizing the common branch of choke-input rectifier circuits has the form of a commutated sine wave for any load value $\bar{R} \leqq \bar{R}_{cr}$. Since the circuit is never interrupted, currents and voltages may be considered steady-state components in a circuit energized by the equivalent commutated sine voltage as shown in Fig. 11a. The single generator in Fig. 11a may

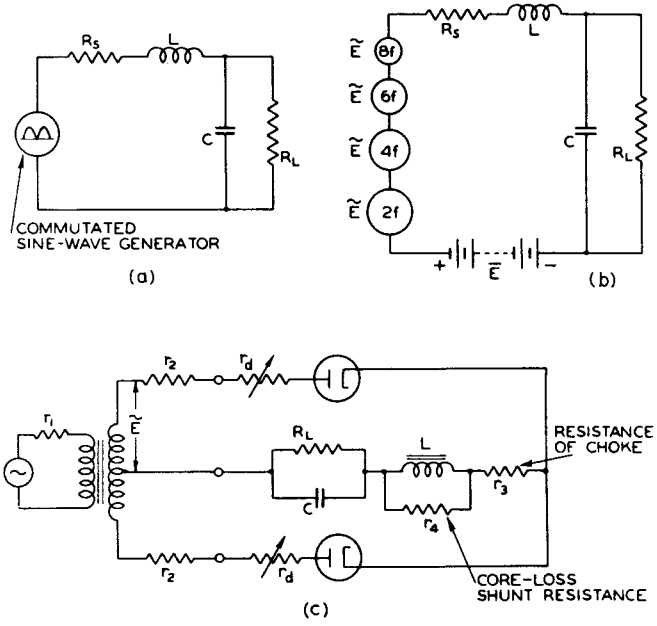


Fig. 11

be replaced as shown in Fig. 11b by a battery and a series of sine-wave generators having frequencies and amplitudes, as given by the following equation of the commutated sine wave.

$$\tilde{e} = \frac{2\,\tilde{e}_{max}}{\pi}\left[1 - \frac{2\cos 2f}{1\times 3} - \frac{2\cos 4f}{3\times 5} - \frac{2\cos 6f}{5\times 7} - \cdots\right] \qquad (10)$$

The d-c component or battery voltage is the total average voltage $\bar{E}$ discussed in the preceding sections.

Some useful relations of voltage components are:

Line voltage induced in one-half of the secondary winding (rms)	$\lvert\tilde{E}\rvert = 1.1\ \bar{E}$	(11)
Total average voltage	$\bar{E} = \begin{cases} 0.90\ \lvert\tilde{E}\rvert \\ 0.637\ \tilde{e}_{max} \end{cases}$	
Voltage of frequency 2f (rms)	$\lvert\tilde{E}\rvert_{2f} = \begin{cases} 0.424\ \lvert\tilde{E}\rvert \\ 0.471\ \bar{E} \end{cases}$	
Voltage of frequency 4f (rms)	$\lvert\tilde{E}\rvert_{4f} = \begin{cases} 0.085\ \lvert\tilde{E}\rvert \\ 0.0945\ \bar{E} \end{cases}$	
Total choke voltage (rms)	$\lvert E\rvert_L = \begin{cases} \sqrt{\lvert\tilde{E}\rvert^2 - \bar{E}^2} \\ 0.482\ \bar{E} \end{cases}$	

The current components in the common circuit branch are calculated from the above voltages divided by the impedance of one branch-circuit at the particular frequency. Because the current is commutated every half-cycle of the line frequency from one to the other branch-circuit, the average current in each diode circuit is one-half of the total average current; and rms values of currents or current components in each branch-circuit are obtained by multiplying the rms current values in the common circuit branch by $1/\sqrt{2}$. The peak current in each diode circuit has the same value as in the common circuit branch.

Average load current	$\bar{I} = \dfrac{\bar{E}}{R_s + R_L}$	(12a)
Average plate current (per diode)	$\bar{I}_p = 0.5\ \bar{I}$	
Double-frequency current (rms) in common circuit branch	$\lvert\tilde{I}\rvert_{2f} = \dfrac{\lvert\tilde{E}\rvert_{2f}}{Z_{(2f)}}$	
Total current (rms) in common circuit branch	$\lvert I\rvert_L = \sqrt{\bar{I}^2 + \lvert I\rvert_{2f}^2}$	(12b)
Rms diode current or rms current per transformer winding	$\lvert I\rvert_d = \dfrac{\lvert I\rvert_L}{\sqrt{2}}$	
Peak diode current	$\hat{i}_d = \bar{I} + (\lvert\tilde{I}\rvert_{2f} \times \sqrt{2})$	

The total power dissipated in diode and load circuits of the practical secondary circuit shown in Fig. 11c is the sum of the power losses in the circuit resistances. In equation form, it is

Total Power = Series-Resistance Loss
+ Choke-Core Loss
+ D-C Power in Load

The plate dissipation per diode is given by:

$$P_d = 0.5\ |I|_L^2 \times |r_d| \qquad (13)$$

With reference to equation (5), we have

$$P_d = 0.5\ |I|_L^2 \times \frac{\bar{e}_d}{\bar{I}} \qquad (14)$$

where $\bar{e}_d$ is the diode voltage taken from the static diode characteristic at the output-current value $\bar{I}$.

The regulation of choke-input circuits is determined by the total series resistance $\bar{R}_S$, since the voltage $\bar{E}$ in the circuit is constant in the useful working range for an energizing a-c voltage of constant value. Thus, the regulation curve has the slope $\bar{R}_S$ (see Fig. 12), which includes the diode resistance. The regulation curve for a circuit with high-vacuum diodes is the geometric sum of the 3/2-power-law diode characteristic and the ohmic series resistance r_2 of one branch-circuit as shown in Fig. 12. The curve is correct for constant voltage $\tilde{e}$ and beyond the critical-current value. In practical circuits, the voltage source $\tilde{e}$ has a certain equivalent resistance, which must be added to r_2.

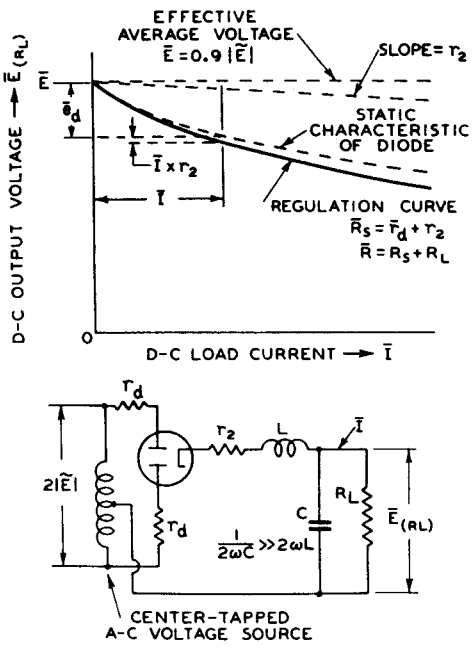


Fig. 12

The equivalent internal resistance of the rectifier circuit as a d-c supply source is the slope of the regulation curve at the current value under consideration. This value should be used for steady-output conditions only, since the reactances in the load circuit cause transients at the instant of sudden load changes.

2. The Condenser-Input Circuit

In rectifier circuits with shunt-condenser-input loads, the condenser is alternately charged and discharged. In the final state of operation, charge and discharge are balanced. The graphic analysis of such circuits is comparatively simple and readily followed. Formulas for the calculation of specific circuit conditions are easily derived from the constructions.

a. Circuits Without Series Resistance — The graphic analysis of a half-wave rectifier circuit without series resistance ($\bar{R}_S$) is illustrated in Fig. 13. Steady-state voltage $\tilde{e}$ and current $\tilde{i}_S$ are constructed on the assumption that the diode is shorted. The steady-state condenser voltage $\tilde{e}_c$ coincides with $\tilde{e}$ because $\bar{R}_S = 0$.

The diode timing is as follows: The diode opens the circuit at point 0 when the diode current becomes zero. The condenser voltage at this instant has the value

$$e_{c(o)} = \tilde{e}_{max} \sin \Theta \qquad (15)$$

Since the condenser-discharge circuit consists of C and R_L, the condenser voltage decays exponentially as shown. At point C it has become equal to the energizing voltage $\tilde{e}$. The diode becomes conducting and closes the circuit. Because there is no potential difference between the steady-state voltages $\tilde{e}$ and $\tilde{e}_c$ at this or any instant, the condenser does not receive a transient charge. The current, therefore, rises instantly to the steady-state value as shown on the $\tilde{i}_S$ curve and then decreases until it is zero at point 0.

The timing of the full-wave circuit in Fig. 14 is quite similar. The time for the condenser discharge through R_L is reduced since $\tilde{e}_c$ meets the positive half-cycle $\tilde{e}_2$ and thus closes the circuit through D_2. Point C in Fig. 14 is located at a higher value of $\tilde{e}$ than in Fig. 13. The conduction angle φ is consequently reduced although C, R_L, and Θ have the same values in both circuits. The average current in the full-wave circuit is, therefore, smaller than twice that of the half-wave circuit.

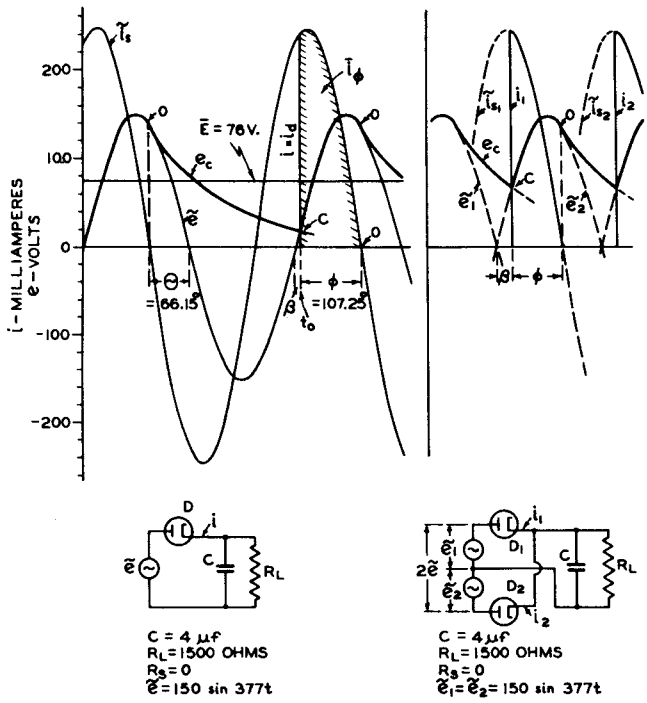


Fig. 13 (left) - Half-wave, condenser-input circuit without series resistance.
Fig. 14 (right) - Full-wave, condenser-input circuit without series resistance.

b. Circuits With Series Resistance — In circuits with series resistance, the steady-state condenser voltage $\tilde{e}_c$ does not coincide with the supply voltage $\tilde{e}$, as illustrated in Figs. 15a and 15b. Phase displacement and magnitudes of current and voltage under steady-state conditions are required for analysis of the circuit and are computed in the conventional manner. The parallel circuit $C \parallel R_L$ is converted into an equivalent series circuit to determine the angles Θ and Θ' by which $\tilde{i}_s$ is leading $\tilde{e}$ and $\tilde{e}_c$, respectively. The steady-state condenser voltage $\tilde{e}_c$ in the parallel circuit equals the voltage across the equivalent circuit as shown by the vector diagram in Fig. 15b.

The diode opens the circuit at the instant $i_d = 0$. For circuit constants as in Fig. 15, the diode current i_d substantially equals $\tilde{i}_s$ at the time of circuit interruption because the transient component i_t' of the current has decayed to a negligible value. Point 0 is thus easily located. In circuits with large series resistance, however, $i_d = 0$ does not coincide with $\tilde{i}_s = 0$ due to slow decay of the transient. In both cases the condenser voltage $\tilde{e}_{c(o)}$ equals the voltage $\tilde{e}_{(o)}$ at the time 0, because $i_d = 0$ and consequently there is no potential difference on R_S. The condenser voltage decays exponentially on R_L from its initial value at 0, as discussed for circuits with $R_S = 0$, and meets the supply voltage $\tilde{e}$ again at point C. At this instant (t_o), the diode closes the circuit. Current and voltage, however, do not rise to their steady-state values as in circuits with $R_S = 0$, because the steady-state voltage $\tilde{e}_{c(o)}$ differs from the line voltage $\tilde{e}_{(o)}$ by the amount $\tilde{e}_{c(o)} = \tilde{i}_{s(o)} \times R_S$. This demanded potential change on the capacitance causes transients. It is possible to determine the value of these transients analytically but the method is too involved to discuss in this lecture. The transients e_t and i_t in Fig. 15a prevent voltage and current from following the steady-state waveforms. When R_S is large, they do not decay to zero within one cycle and, therefore, require additional steps in the graphic solution. The oscillogram shown in Fig. 16 checks the graphical construction of Fig. 15.

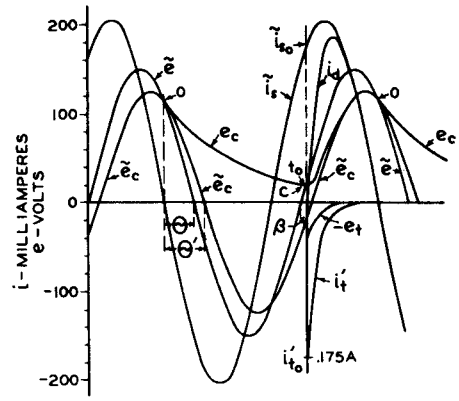


Fig. 15a

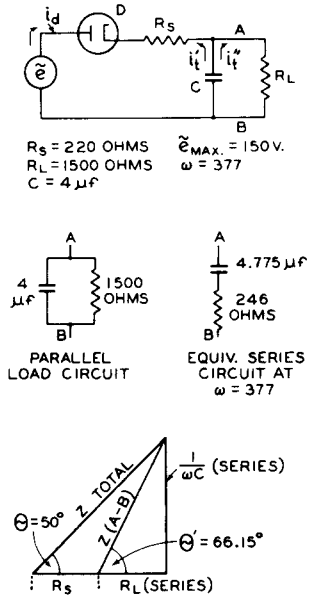


Fig. 15b

Figs. 15a and 15b - Half-wave, condenser-input circuit with series resistance.

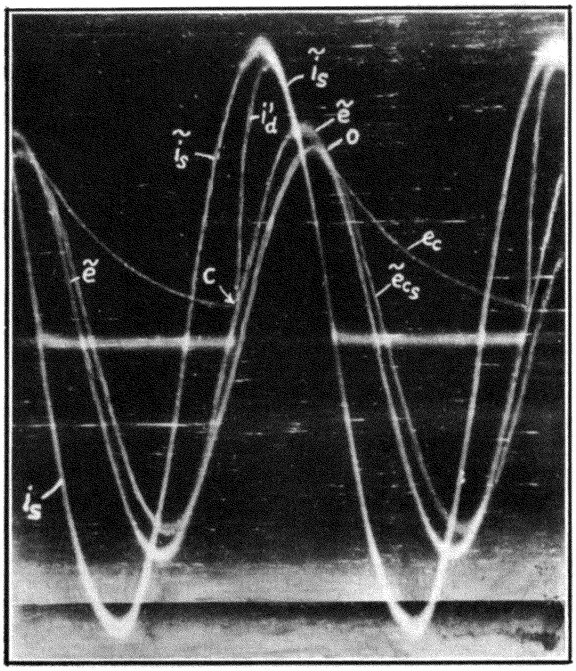
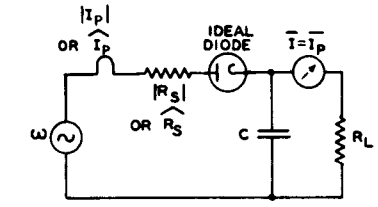


Fig. 16 – Oscillogram verifying graphical construction of Fig. 15.

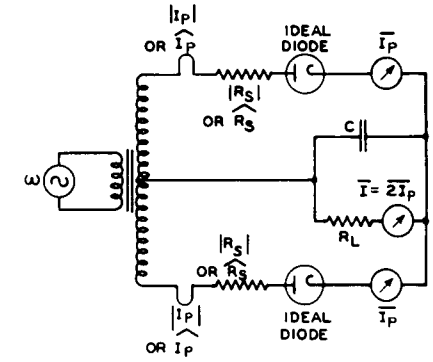
c. Generalized Operation Characteristics (Steady-State Operation) — It has been shown that the conduction angle φ is a function of the circuit constants in condenser-input circuits. The section of the energizing voltage $\tilde{e}$ utilized during conduction time has, therefore, no fixed value as in choke-input circuits where φ = 180° and where the voltage $\tilde{e}$ during φ is a half-sine wave. It is, therefore, not possible to derive a general equivalent circuit for condenser-input circuits which contains a voltage source of fixed wave shape and magnitude.

Steady-state conditions as well as transients are controlled by the circuit constants, which are contained in the product ωCR_L. The angle φ depends on the relative magnitudes of R_L and $\bar{R}_S$ and is, therefore, described in general if also the ratio $\bar{R}_S/R_L$ is known. General curve families may thus be evaluated which show the dependent variables $\bar{E}$, $\hat{i}$, and $\bar{I}$ in terms of ratio versus the independent variable ωCR_L for various parameter values $\bar{R}_S/R_L$. The series resistance $\bar{R}_S$ includes the equivalent diode resistance which is evaluated by means of equation (6), because the current wave is periodic in the final operating state. The reasoning leading to equation (6) is not applicable to a single transient, as obtained for starting conditions of rectifier circuits.

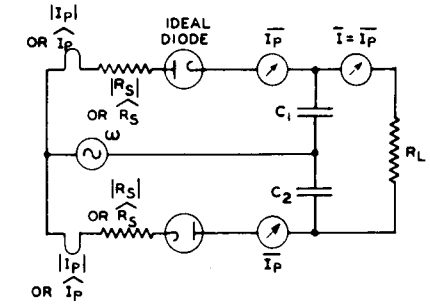
Generalized characteristics have been evaluated for the three types of circuits shown in Fig. 17. The characteristics in Figs. 18, 19, and 20 show the average voltage $\bar{E}$ across the load resistance R_L as a function of ωCR_L and $\bar{R}_S$ in per cent of the applied a-c peak voltage (sine wave) for half-wave, full-wave, and voltage-doubling circuits. They permit the solution of



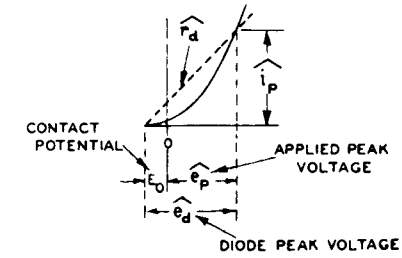
HALF-WAVE CIRCUIT



FULL-WAVE CIRCUIT



VOLTAGE-DOUBLING CIRCUIT



$\hat{r}_d = .88\bar{r}_d = .935|r_d| = \frac{\hat{e}_d}{\hat{i}_P}$

$\hat{R}_S = R_S + \hat{r}_d$

$\bar{R}_S = R_S + \bar{r}_d$

$|R_S| = R_S + |r_d|$

R_S = EXTERNAL RESISTANCE

$\hat{r}_d$ = PEAK DIODE RESISTANCE

$\bar{r}_d$ = EQUIVALENT AVE. DIODE RESISTANCE

$|r_d|$ = EQUIVALENT R M S DIODE RESISTANCE

Fig. 17 – Evaluation of characteristics for half-wave, full-wave, and voltage-doubling circuits.

RELATION OF APPLIED AC VOLTAGE TO DC OUTPUT VOLTAGE
IN HALF-WAVE RECTIFIER CIRCUIT.

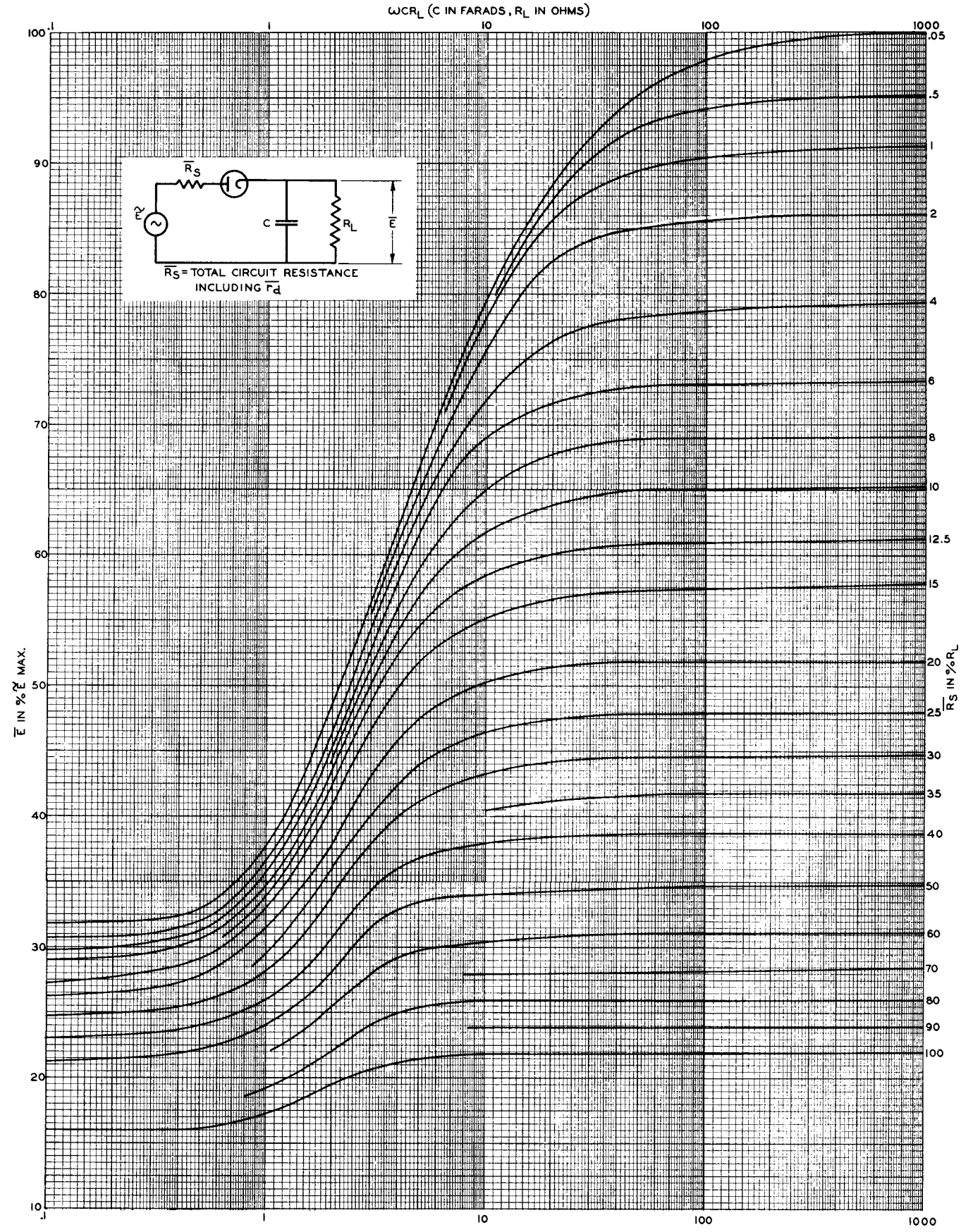


Fig. 18

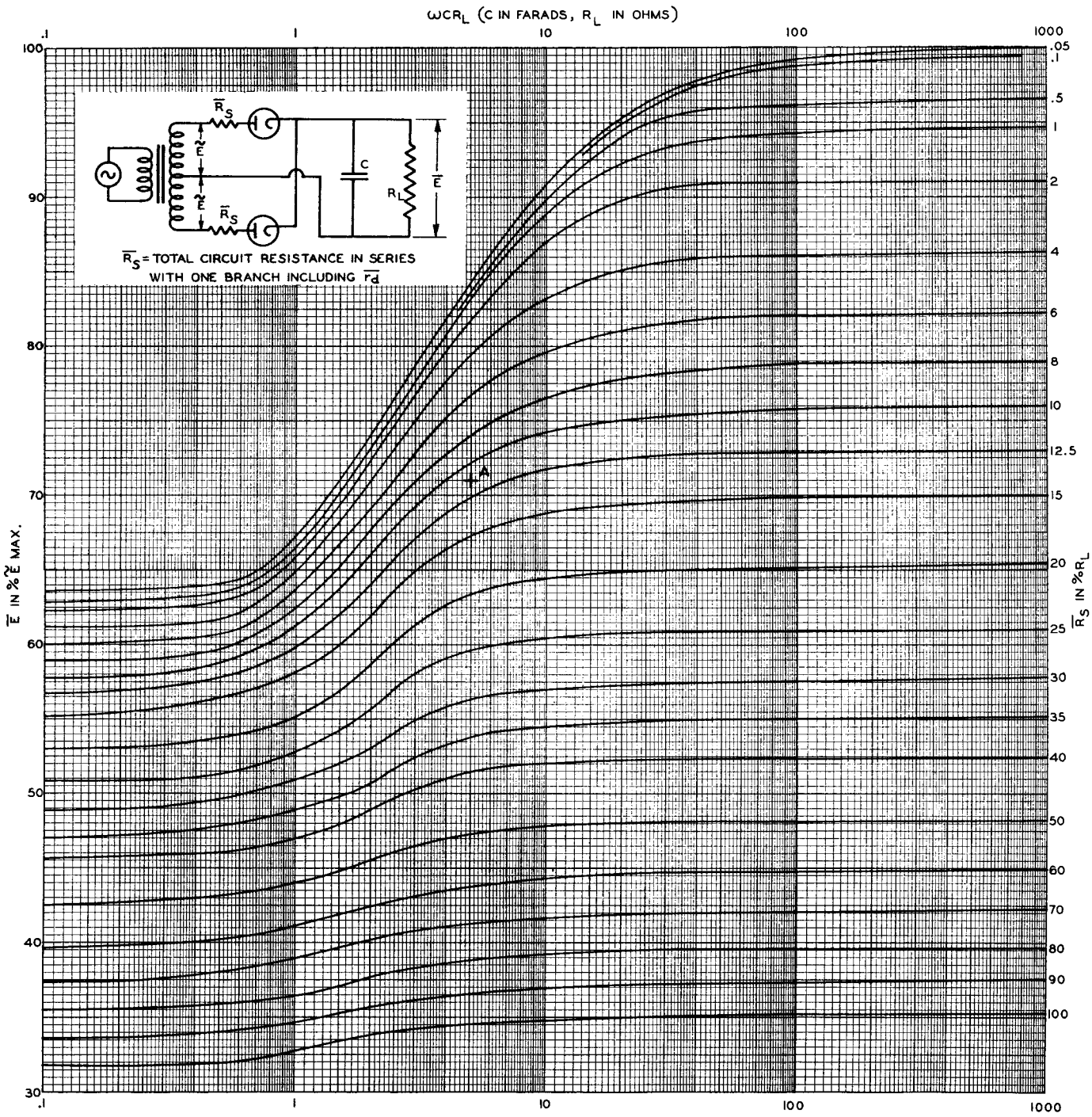

RELATION OF APPLIED AC VOLTAGE TO DC OUTPUT VOLTAGE
IN FULL-WAVE RECTIFIER CIRCUIT
ωCR_L (C IN FARADS, R_L IN OHMS)
.1
1
10
100
1000
100
90
80
70
60
50
40
30
.05
.1
.5
1
2
4
6
8
10
12.5
15
20
25
30
35
40
50
60
70
80
90
100
Ē IN % Ẽ MAX.
R̄S IN % RL
R̄S
Ẽ
C
RL
Ē
A
R̄S = TOTAL CIRCUIT RESISTANCE IN SERIES
WITH ONE BRANCH INCLUDING r̄d

Fig. 19

RELATION OF APPLIED AC VOLTAGE TO DC OUTPUT VOLTAGE IN VOLTAGE-DOUBLING CIRCUIT

ωCR_L (C IN FARADS, R_L IN OHMS)

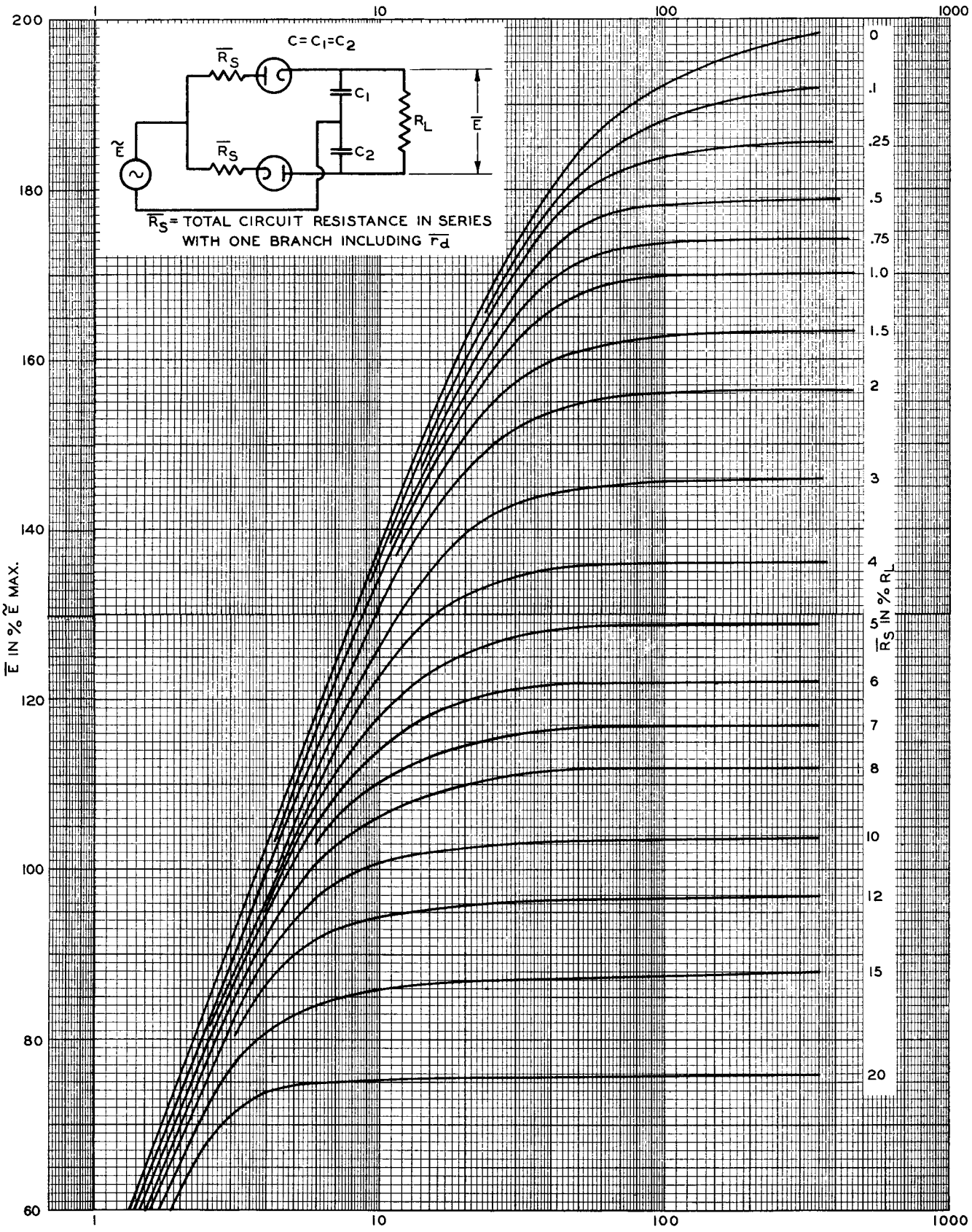


Fig. 20

the reversed problem to determine the magnitude of the applied voltage necessary to give a certain average voltage output for a given load. The series-resistance value $\bar{R}_S$ includes the equivalent average resistance $\bar{r}_d$ of one diode and the power-transformer resistances as reflected into one secondary winding. The characteristics were plotted from accurately measured values as their complete calculation required too much time. The measurements were made on circuits of negligible inductive reactance. Series-resistance values were determined accurately by the method shown in Fig. 2. Table II gives a number of calculated values which show the accuracy of the curves to be approximately 5 per cent or better.

In compiling the data for the current-ratio characteristics in Fig. 21, it was found that the three rectifier-circuit types could be shown by a single family after a "charge factor" $\underline{n}$ was added to the product of the circuit constants ωCR_L and to $\bar{R}_S$ as shown in Table II. The factor $\underline{n}$ is unity for the half-wave circuit. For the full-wave circuit, $\underline{n}$ is 2 because the condenser C is charged twice during one cycle. For the voltage-doubling circuit, $\underline{n}$ is 1/2 because the two condensers require together twice the charge to deliver the same average current at double voltage. The values in the table indicate that the factor $\underline{n}$ is actually not a constant. The mean value of the current ratios does, however, not depart more than approximately 5 per cent from the true value, the error being a maximum in the steep portion of the curves and decreasing to zero at both ends. The upper section of Fig. 21 shows the ratio of rms current to average current per diode plate. This family is of special interest in the design of power transformers and for computation of diode plate dissipation.

Fig. 22 shows the rms value of the ripple voltage across R_L in per cent of the average voltage.

The voltage-doubling circuit shown with the other two condenser-input circuits in Fig. 17 may be regarded in principle as a series connection of two half-wave rectifier circuits. Each condenser is charged separately during conduction time of one diode, but is discharged in series with the other condenser during the time of non-conduction of its associated diode. The analysis of operation is made according to the method discussed but will not be treated.

Table II

TYPE OF CONDENSER-INPUT CIRCUIT	$n\omega CR_L$	$\frac{\bar{R}_S}{n\ R_L}$	Θ°	φ°	$\frac{\bar{E}}{\hat{e}_{max}}$	$\frac{\hat{i}_d}{\bar{I}_p}$	$\frac{\lvert I_p \rvert}{\bar{I}_p}$
Half-Wave n = 1	0.5	0	26.5	153.5	0.335	3.33	1.69
	1.	0	45.0	134.0	0.384	3.68	1.81
	2.	0	63.4	111.6	0.486	4.61	2.00
	2.26	0	66.15	106.4	0.503	4.91	2.02
	4.	0	75.9	87.1	0.623	6.60	2.24
	8.	0	82.9	65.1	0.742	9.86	2.60
	16.	0	86.4	48.6	0.862	13.92	3.00
	32.	0	88.2	35.3	0.930	19.90	3.51
	64.	0	89.1	25.1	0.996	27.5 ?	4.16
	2.	0.10	-	121.	0.434	4.48	1.9
	2.26	0.147	50.	123.	0.428	4.42	1.88
	4.	0.05	65.1	99.3	0.632	5.28	2.1
	4.	0.10	56.	108.4	0.537	5.14	2.0
Full-Wave n = 2	1.	0	26.5	142.5	0.644	3.47	1.75
	2.	0	45.0	121.0	0.678	4.17	1.90
	4.	0	63.4	92.6	0.740	6.06	2.17
	4.52	0	66.15	86.8	0.744	6.55	2.24
	8.	0	75.9	67.0	0.816	9.30	2.55
	16.	0	83.0	49.0	0.885	13.74	3.00
	32.	0	86.4	35.6	0.945	19.70	3.50
	64.	0	88.2	25.4	0.999	27.1 ?	4.15
	4.	0.05	-	104.	0.671	5.43	2.05
	4.52	0.0735	50.	105.	0.636	5.35	2.04
	8.	0.05	56.	90.	0.710	6.20	2.20
	30.2	0.10	17.9	100.6	0.646	5.39	2.08

RELATIONS OF PEAK, AVERAGE, AND RMS DIODE CURRENT

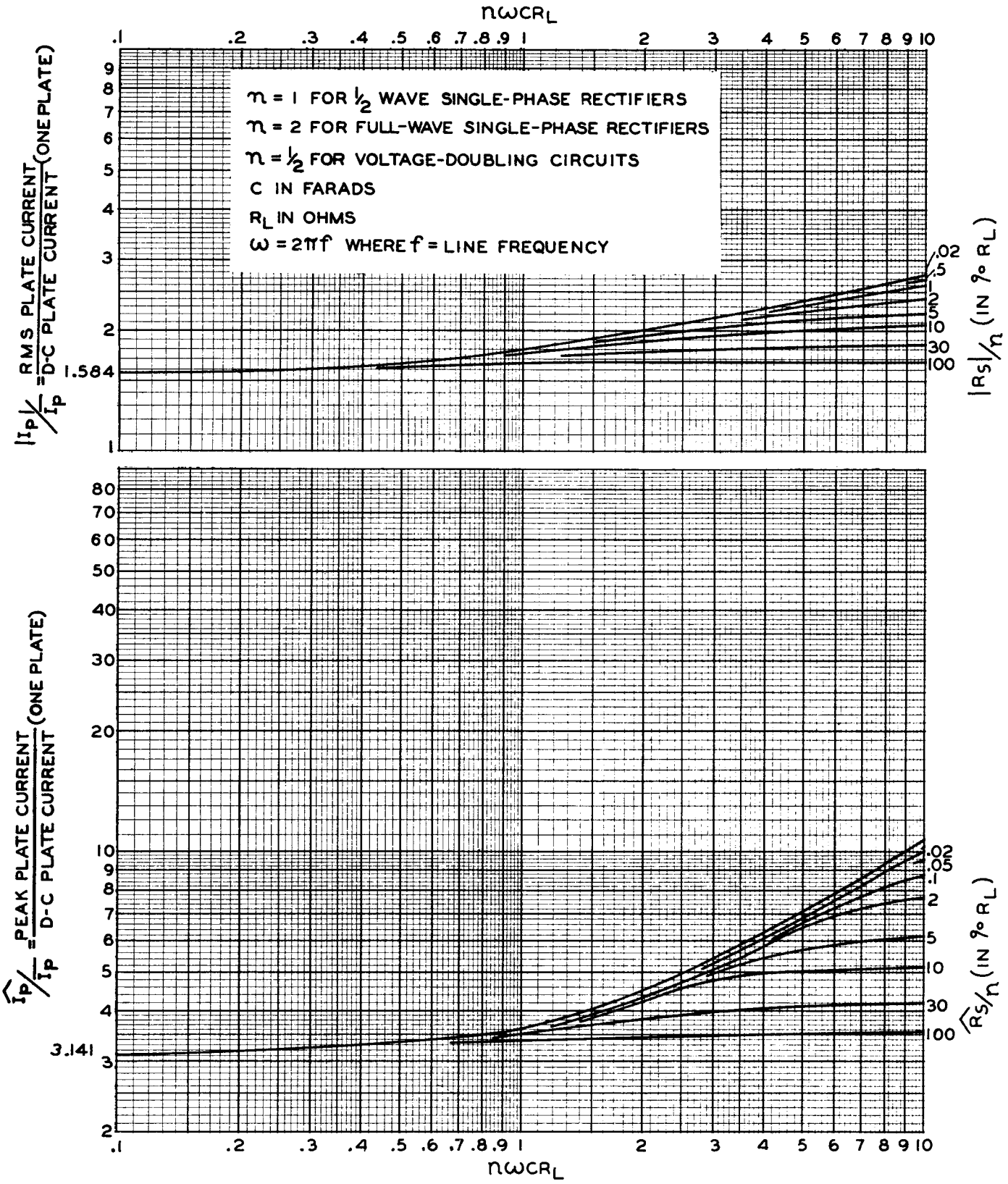


Fig. 21

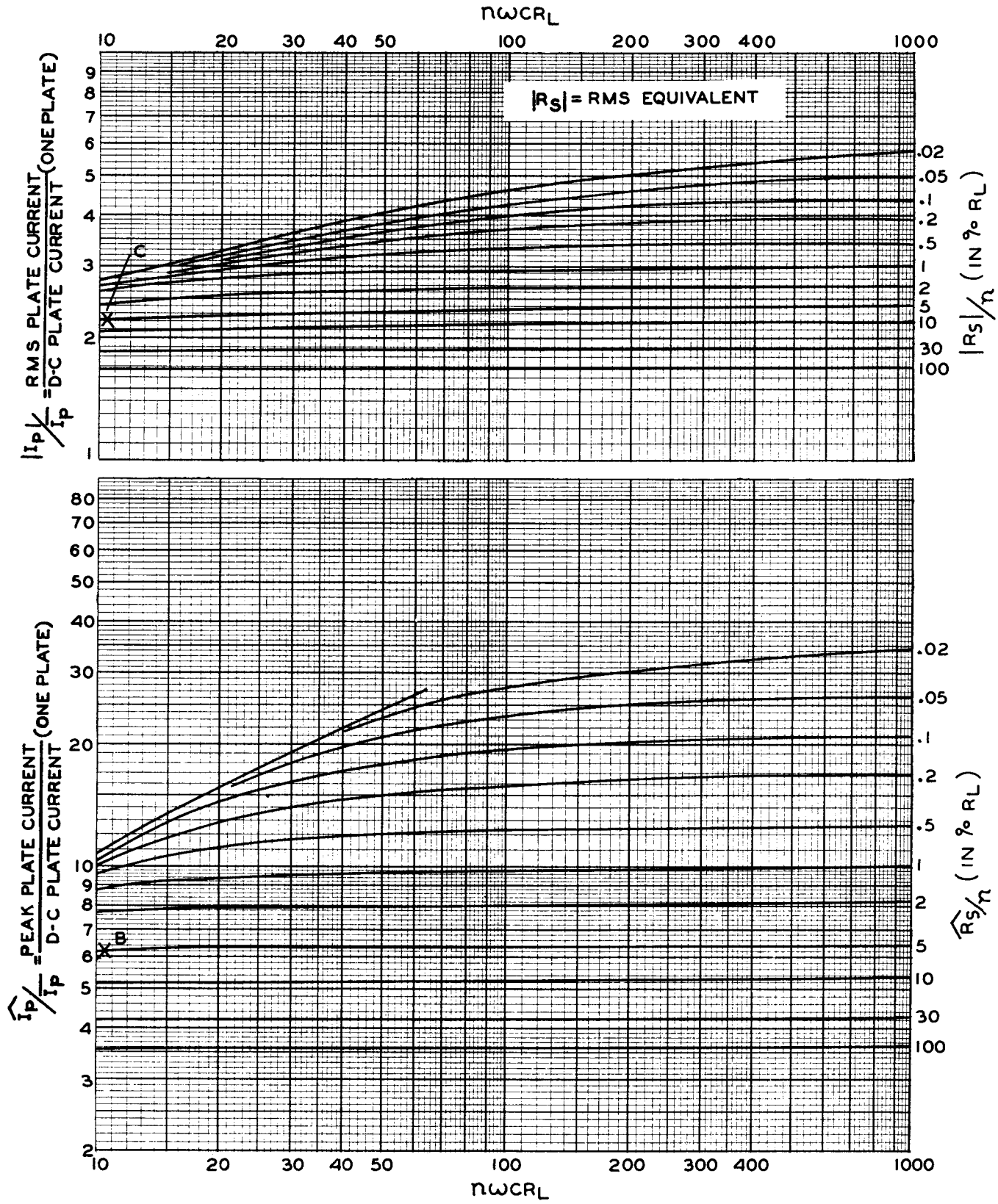

nωCRL
10 20 30 40 50 100 200 300 400 1000
|RS| = RMS EQUIVALENT
|Ip|/Ip = RMS PLATE CURRENT / D-C PLATE CURRENT (ONE PLATE)
|RS|/n (IN % RL)
.02 .05 .1 .2 .5 1 2 5 10 30 100
C
B
Îp/Ip = PEAK PLATE CURRENT / D-C PLATE CURRENT (ONE PLATE)
ÎRS/n (IN % RL)
nωCRL

Fig. 21 (cont'd)

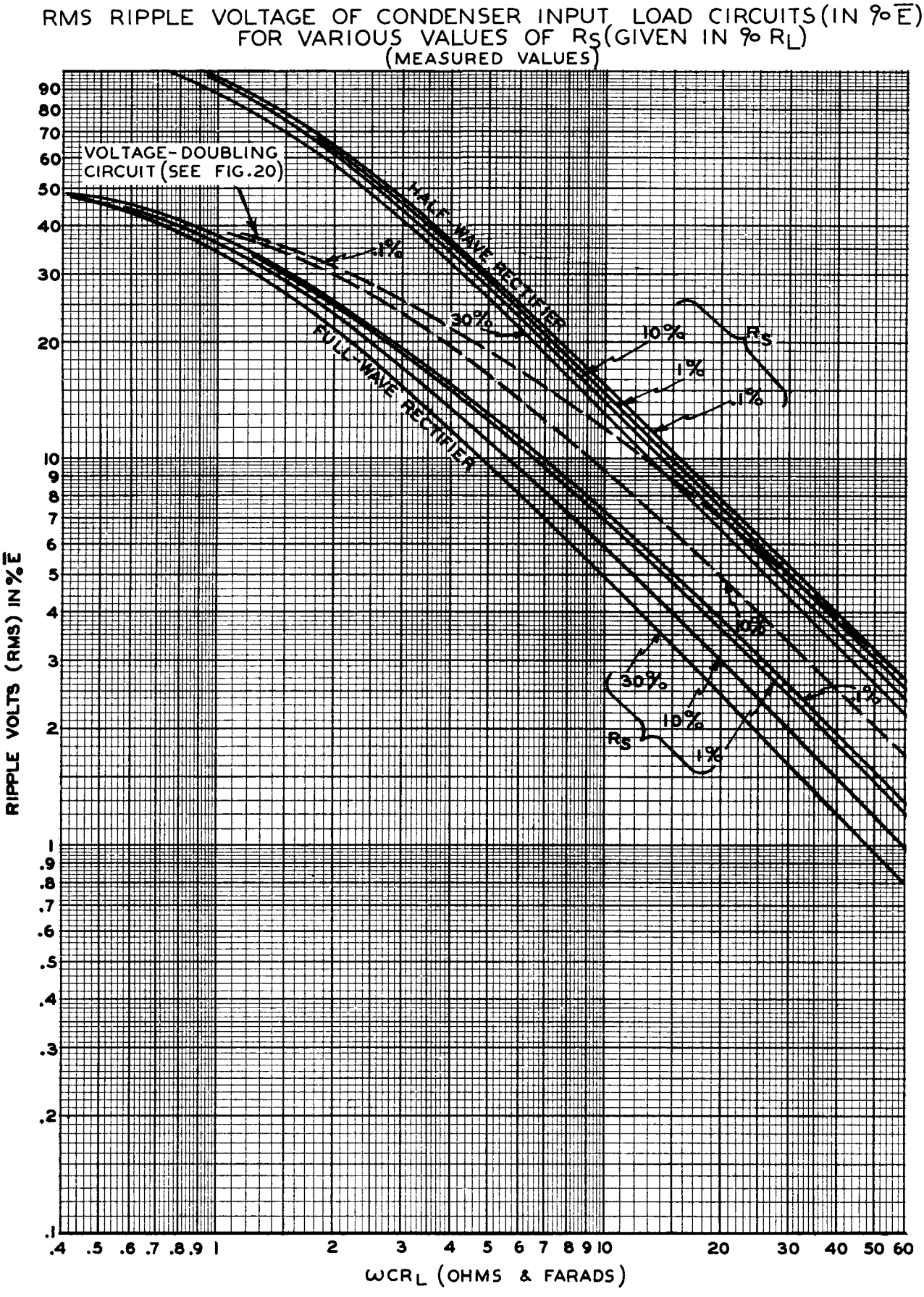

RMS RIPPLE VOLTAGE OF CONDENSER INPUT LOAD CIRCUITS (IN % Ē)
FOR VARIOUS VALUES OF RS (GIVEN IN % RL)
(MEASURED VALUES)
VOLTAGE-DOUBLING CIRCUIT (SEE FIG. 20)
HALF-WAVE RECTIFIER
FULL-WAVE RECTIFIER
1%
30%
10%
RS
RIPPLE VOLTS (RMS) IN % Ē
ωCRL (OHMS & FARADS)

Fig. 22

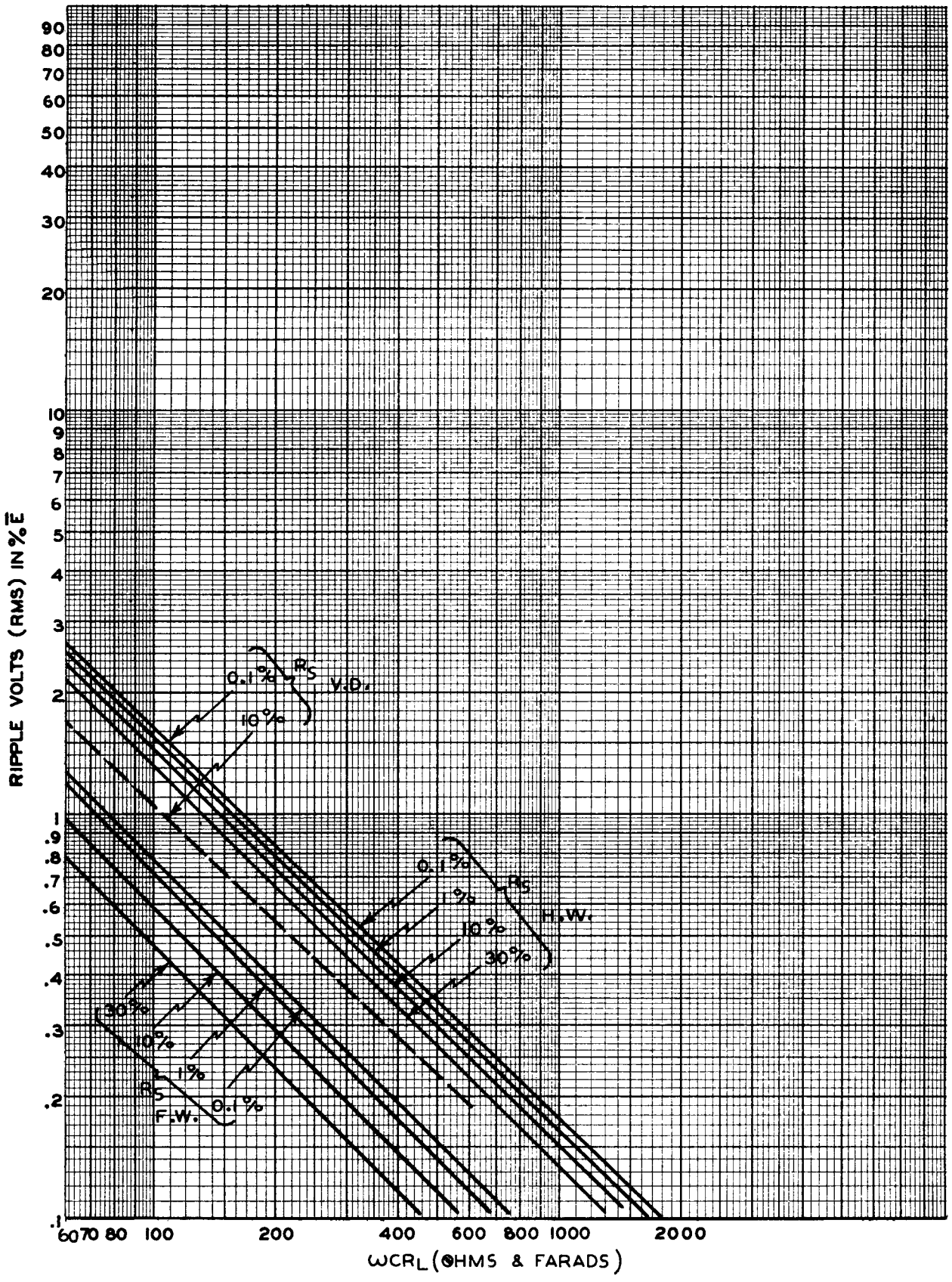

RIPPLE VOLTS (RMS) IN % Ē
90
80
70
60
50
40
30
20
10
9
8
7
6
5
4
3
2
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
60 70 80 100 200 400 600 800 1000 2000
ωCR_L (OHMS & FARADS)
0.1%
10%
R_S
V.D.
0.1%
1%
10%
30%
R_S
H.W.
30%
10%
1%
R_S
F.W.
0.1%

Fig. 22 (cont'd)

Lecture 20

THE DESIGN OF AUDIO AMPLIFIER AND POWER OUTPUT TUBES

S. W. Dodge

INTRODUCTION

During the past ten to fifteen years, the efficiency of receiving tubes has been increased by the development of special types for specific purposes. Until 1924, triodes with a filamentary cathode were the only types in common use. They were designed with high or low amplification factor according to the application. The principal requirements of such tubes were sufficient emission and a gas current less than a few microamperes. Because of the inherent limitations of the loudspeaker and the poor characteristics of the inter-stage transformers, the distortion introduced by the tube itself was relatively unimportant.

A tetrode tube with a fourth electrode in the form of a grid between the control grid and the plate was patented by Siemens and Halske in 1916. The addition of this fourth electrode, or screen grid, operating at a fixed positive potential, introduced secondary-emission difficulties. In 1926, two methods of suppressing the secondary current between plate and screen grid were described by N. V. Philips' Gloeilampenfabrieken of Holland. In the first method, the plate was spaced a sufficient distance from the screen, while in the second an open-mesh grid, or suppressor grid, was introduced between the plate and the screen grid. Tubes constructed with cathode, control grid, screen grid, suppressor grid, and plate were designated as pentodes and are now used in different forms for high-frequency as well as low-frequency amplification. The design of the suppressor grid will be discussed later in this lecture.

It is not within the scope of this lecture to describe the many circuits used in conjunction with various types of tubes but rather to outline some of the methods of tube design and to give some simplified formulas which may be readily used by the factory engineer for adjusting characteristics of tubes already in production.

CALCULATION OF CHARACTERISTICS

It has not been found possible to calculate with great accuracy the characteristics of receiving tubes of given dimensions or, conversely, to determine the dimensions necessary to give the required characteristics. Some of the numerous reasons for this situation are found in end effects, grid side-rod effects, and "Insulbildung".

The ends of the cathode are in contact with the mica spacers, the cathode tab, and the heater leads. At the points of contact, the temperature of the cathode is reduced with the result that the emission from the cathode in these and adjacent areas is also reduced. Then, too, the mica spacers acquire charges which alter the field near the ends of the cathode. The extent to which the emission is reduced and the field altered is neither accurately known nor readily calculated.

The influence of grid side-rods on characteristics is another factor which is not readily calculated, but we do know that as the spacing between grid side-rods and cathode is decreased, or as the diameter of the side-rods is increased, the error in calculation is increased.

The lack of uniformity of the electric field near the cathode surface is called "Insulbildung" by the Germans. It increases with the ratio of grid pitch to grid-cathode spacing, but the extent of the increase is not accurately known.

On the basis of certain assumptions, however, formulas have been developed which aid in determining the effect of the various electrode dimensions on the characteristics of tubes.

1. Amplification Factor (μ) in Triodes

Vogdes and Elder[1] deduced the following expression for the amplification factor (μ) of a triode. This expression neglects end effects.

$$\mu = K_1 S - K_2$$

where,

$$K_1 = \frac{2\pi N}{-\log \tanh 2\pi N b}$$

$$K_2 = \frac{\log \cosh 2\pi N b}{-\log \tanh 2\pi N b}$$

$$S = \begin{cases} \text{grid-to-plate spacing in inches for a parallel-plane structure,} \\ r_g \log_e \dfrac{r_p}{r_g} \text{ in inches for a cylindrical structure.} \end{cases}$$

N = grid turns per inch
b = grid-wire radius in inches
r_g = grid radius in inches
r_p = plate radius in inches

A plot which gives K_1 and K_2 directly when the turns per inch of the grid and the grid-wire diameter are known is shown in Fig. 1. Fig. 2 shows a

[1] Vogdes and Elder, "Formulas for the Amplification Constant for Three-Element Tubes in which the Diameter of Grid Wires is Large Compared to the Spacing," Physical Review, Vol. 24, p. 683; 1924: also, abstract, Vol. 25, p. 255; 1925.

plot which gives S for a cylindrical structure directly when the grid diameter and the plate diameter are known.

For small changes in μ, such as would be needed for the adjustment of characteristics of an existing tube, the approximate formula is

$$\mu = K_3 N^2 d S$$

where,

K_3 = a constant
N = grid turns per inch
d = grid-wire diameter
S = grid-to-plate spacing

It is not necessary to determine or know K_3. To raise μ by 10 per cent, for example, merely increase d or S by 10 per cent, or N by 5 per cent.

The error introduced by the use of this approximate formula is quite small. It increases with a decrease in the turns per inch. For a change in μ of 10 per cent, the error is 2.4 per cent at 20 turns per inch and drops rapidly to 0.4 per cent at 40 turns per inch.

2. Plate Current (I_b) and Transconductance (g_m) in Triodes, Tetrodes, and Pentodes

The equation for the plate current of an ideal parallel-plane triode is

$$I_b = 2.33 \times 10^{-6} \frac{A}{d_{gk}^2} \left[\frac{E_c + \frac{E_b}{\mu} + \varepsilon}{1 + \frac{1}{\mu} + \frac{4}{3} \frac{d_{gp}}{\mu d_{gk}}} \right]^{\frac{3}{2}}$$

where,

A = cathode area
d_{gk} = grid-to-cathode spacing
E_c = bias applied to grid
E_b = plate voltage
μ = amplification factor
ε = contact potential difference in volts between cathode and grid
d_{gp} = grid-to-plate spacing

This equation for plate current of a triode applies also to parallel-plane tetrodes and pentodes when certain of the terms are defined as indicated below.

I_b = plate current + screen current
E_b = screen voltage
μ = amplification factor from control grid to screen grid
d_{gp} = control-grid to screen-grid spacing

For all heater-cathode types of tubes, the following approximate formulas are convenient for making small adjustments in the plate current and transconductance of existing tubes.

$$I_b = \frac{K_4}{d_{gk}^2} \left(E_c + \frac{E_b}{\mu} + \varepsilon \right)^{\frac{3}{2}}$$

$$g_m = \frac{1.5\, K_4}{d_{gk}^2} \left(E_c + \frac{E_b}{\mu} + \varepsilon \right)^{\frac{1}{2}}$$

where K_4 is a constant and the other terms are the same as given above. Combining these two equations, we have

$$\frac{g_m}{I_b} = \frac{1.5}{E_c + \frac{E_b}{\mu} + \varepsilon}$$

For the filament-type tubes, the following equations give greater accuracy.

$$I_b = \frac{K_5}{d_{gk}} \left(E_c + \frac{E_b}{\mu} + \varepsilon \right)^2$$

$$g_m = \frac{2\, K_5}{d_{gk}} \left(E_c + \frac{E_b}{\mu} + \varepsilon \right)$$

where K_5 is a constant.

By inspection, it can be readily seen that the error introduced by the use of the approximate formulas decreases with an increase in μ and with a decrease in the ratio d_{gp}/d_{gk}.

With μ and g_m known, the plate resistance r_p is readily found from the relation $r_p g_m = \mu$.

3. Power output and Distortion

Class A Operation — The most important characteristics of power output tubes are the power output and distortion. For class A operation, they may be calculated from the plate-voltage vs plate-current family using the grid biases as indicated in Fig. 3. $E_c = E$ is the normal operating bias. The undistorted power output is

$$U.P.O. = \frac{1}{2} (I')^2 R_p$$

where

$$I' = \frac{1}{4} \left[I_{max} - I_{min} + \sqrt{2} (I_1 - I_2) \right]$$

and

$$R_p = \text{Load Resistance} = \frac{E_{max} - E_{min}}{I_{max} - I_{min}}$$

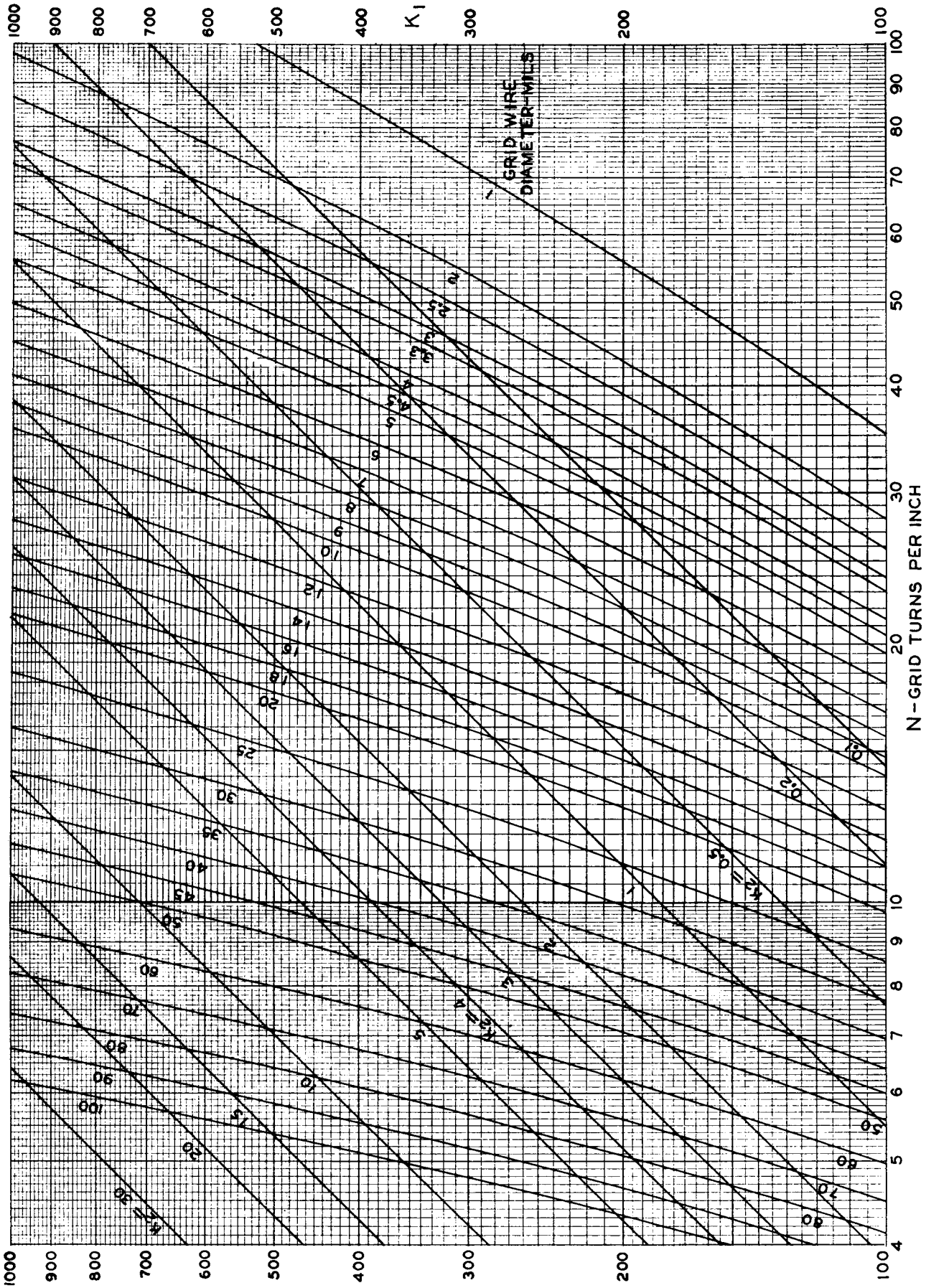


Fig. 1

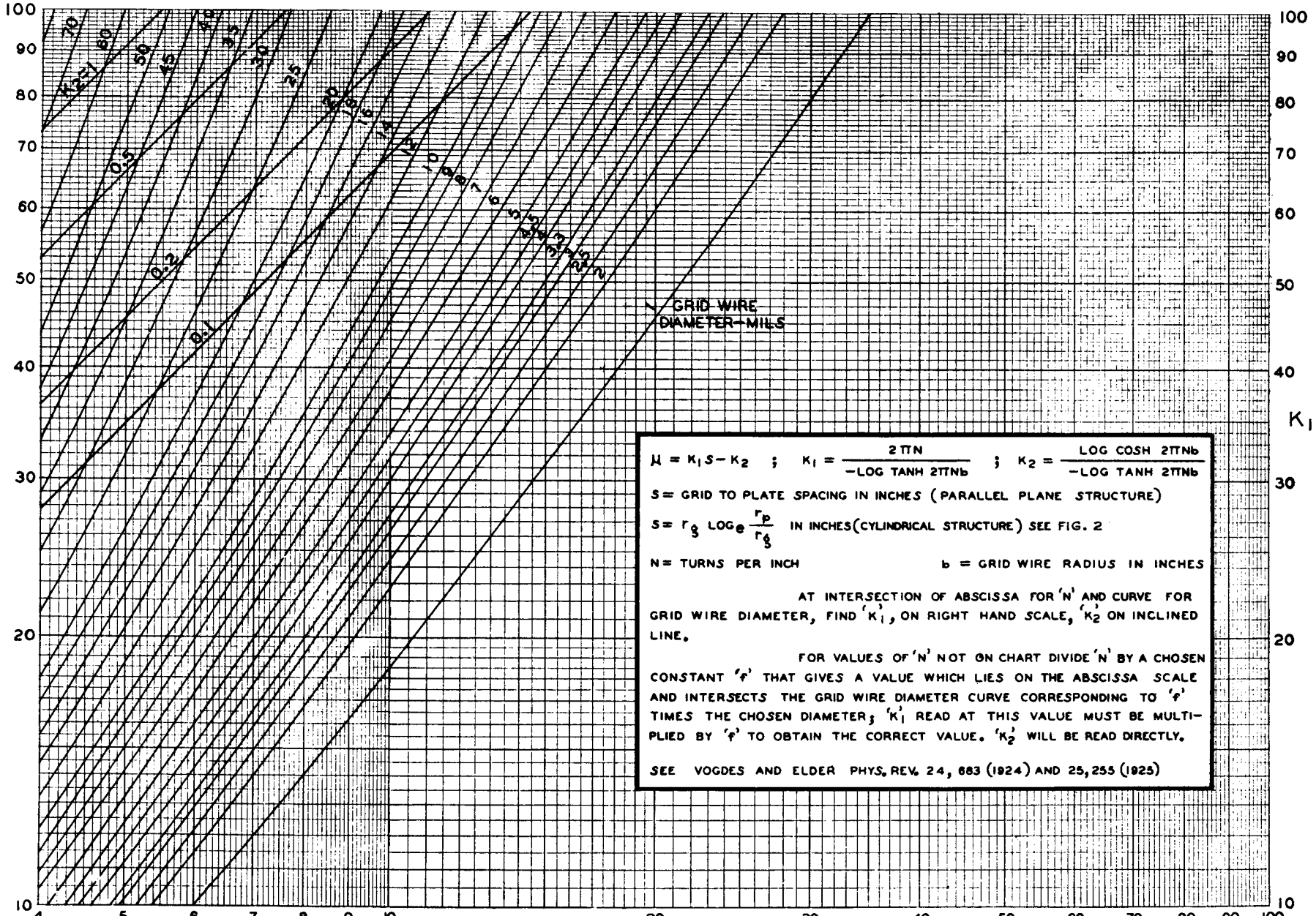

GRID WIRE DIAMETER-MILLS
N-GRID TURNS PER INCH
K1
$\mu = K_1 S - K_2$; $K_1 = \frac{2\pi N}{-\text{LOG TANH } 2\pi Nb}$; $K_2 = \frac{\text{LOG COSH } 2\pi Nb}{-\text{LOG TANH } 2\pi Nb}$
S = GRID TO PLATE SPACING IN INCHES (PARALLEL PLANE STRUCTURE)
$S = r_g \text{ LOG}_e \frac{r_p}{r_g}$ IN INCHES (CYLINDRICAL STRUCTURE) SEE FIG. 2
N = TURNS PER INCH
b = GRID WIRE RADIUS IN INCHES
AT INTERSECTION OF ABSCISSA FOR 'N' AND CURVE FOR GRID WIRE DIAMETER, FIND 'K1', ON RIGHT HAND SCALE, 'K2' ON INCLINED LINE.
FOR VALUES OF 'N' NOT ON CHART DIVIDE 'N' BY A CHOSEN CONSTANT 'f' THAT GIVES A VALUE WHICH LIES ON THE ABSCISSA SCALE AND INTERSECTS THE GRID WIRE DIAMETER CURVE CORRESPONDING TO 'f' TIMES THE CHOSEN DIAMETER; 'K1' READ AT THIS VALUE MUST BE MULTIPLIED BY 'f' TO OBTAIN THE CORRECT VALUE. 'K2' WILL BE READ DIRECTLY.
SEE VOGDES AND ELDER PHYS. REV. 24, 683 (1924) AND 25, 255 (1925)

Fig. 1 (cont'd)

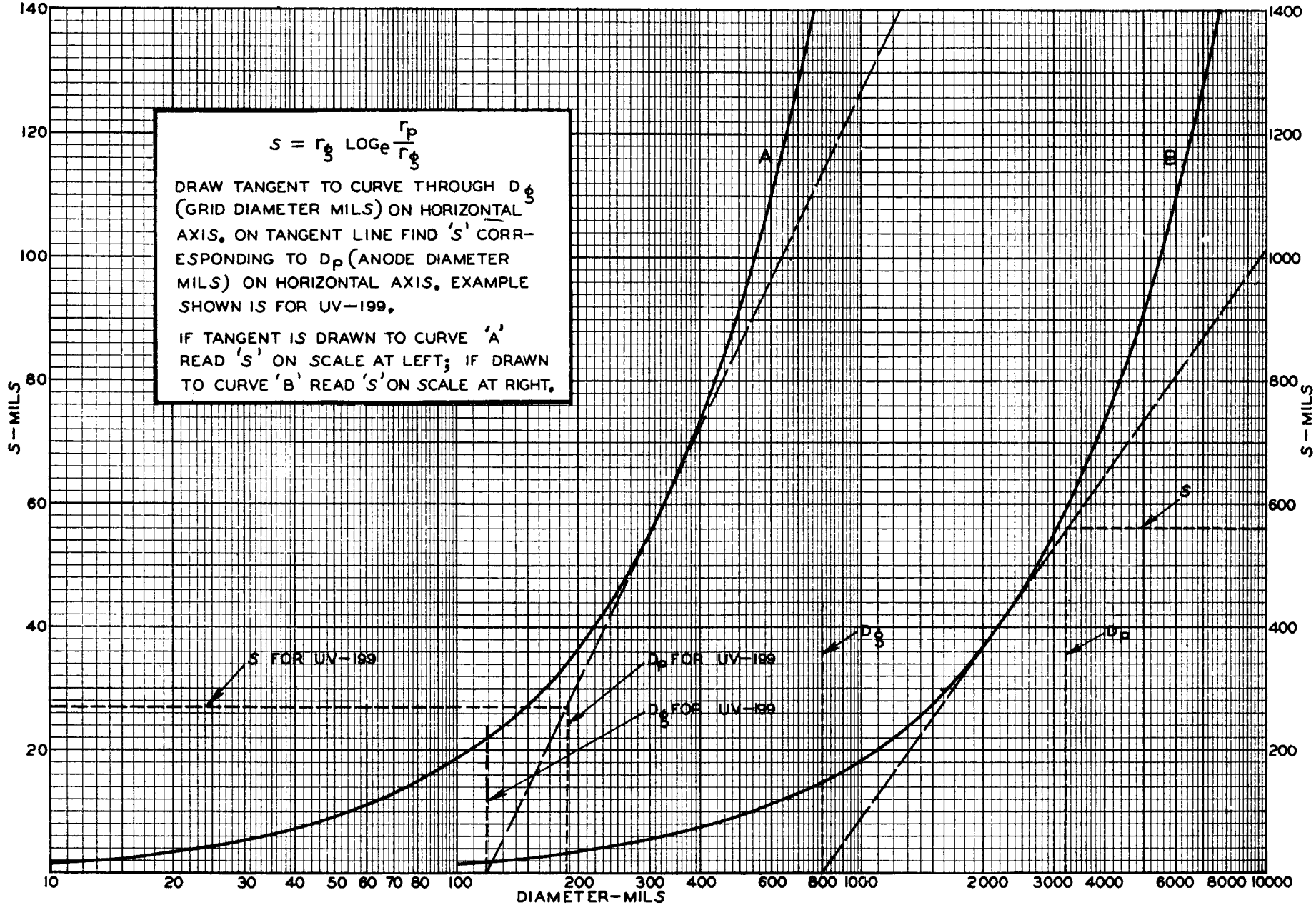

$S = r_{\phi} \log_e \frac{r_p}{r_{\phi}}$
DRAW TANGENT TO CURVE THROUGH D_{ϕ} (GRID DIAMETER MILS) ON HORIZONTAL AXIS. ON TANGENT LINE FIND 'S' CORRESPONDING TO D_p (ANODE DIAMETER MILS) ON HORIZONTAL AXIS. EXAMPLE SHOWN IS FOR UV-199.
IF TANGENT IS DRAWN TO CURVE 'A' READ 'S' ON SCALE AT LEFT; IF DRAWN TO CURVE 'B' READ 'S' ON SCALE AT RIGHT.
A
B
S
S FOR UV-199
D_p FOR UV-199
D_{ϕ} FOR UV-199
D_{ϕ}
D_p
S-MILS
S-MILS
DIAMETER-MILS
140
120
100
80
60
40
20
1400
1200
1000
800
600
400
200
10
20
30
40
50
60
70
80
100
200
300
400
600
800
1000
2000
3000
4000
6000
8000
10000

Fig. 2

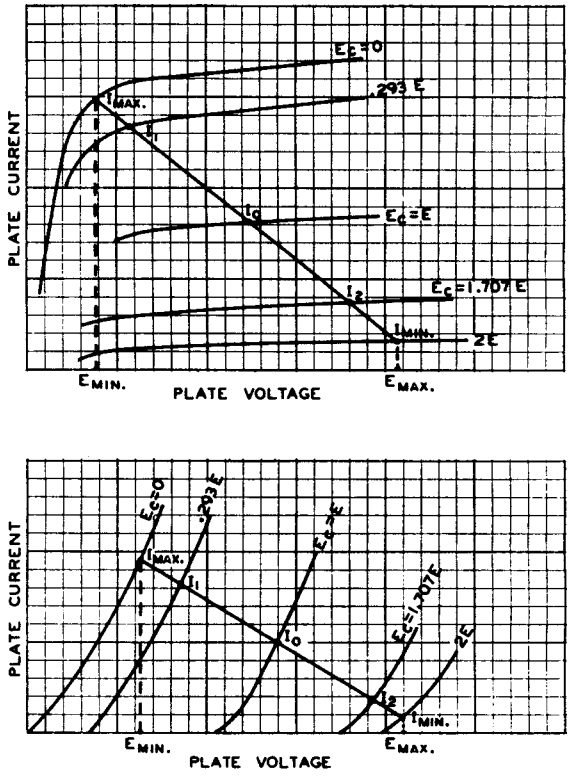


Fig. 3 - (top) Pentode family.
(bottom) Triode family.

The power output will be obtained in watts when the values for E_{max}, E_{min}, I_{max}, I_{min}, I_1, and I_2 are expressed in volts and amperes, respectively.

Distortion can be calculated with the following formulas.

% 2nd Harmonic :

When 3rd harmonic is present =

$$-100\,\frac{I_{max}+I_{min}-2I_0}{I_{max}-I_{min}+\sqrt{2}(I_1-I_2)}$$

When 3rd harmonic is absent =

$$-100\,\frac{\frac{I_{max}+I_{min}}{2}-I_0}{I_{max}-I_{min}}$$

% 3rd Harmonic =

$$-100\,\frac{I_{max}-I_{min}-\sqrt{2}(I_1-I_2)}{I_{max}-I_{min}+\sqrt{2}(I_1-I_2)}$$

The proper operating bias for a single class A triode is equal to $0.675\,E_b/\mu$. Further information on triode operation is given in a paper by Callendar.[2]

Class B Operation — With the advent of class-B output tubes, it became necessary to develop a simple factory test to reject poor tubes. A direct class-B test would consist of two tubes in push-pull driven by a suitable tube and operating into the rated load. The power output and distortion would then indicate the performance. Such a test is not suited for regular production testing.

The elimination of the distortion test and the substitution of an a-c signal in series with a resistance equal to the effective series resistance which would be introduced by the driver tube would simplify the test. However, such a test would not provide a method for testing the power output of a single tube. So, to do this, we insert into the plate circuit of the tube a simple series resistance equal to one quarter of the normal plate-to-plate load and measure the direct current through this resistance. In the case of two tubes, the fundamental power output in watts is

$$P.O._{two\ tubes} = \frac{I_m^2}{2}R$$

and in the case of one tube,

$$P.O._{one\ tube} = \frac{I_m^2}{4}R$$

where I_m is the peak value of current in amperes through the load resistance R in ohms.

The problem then resolves itself into the determination of the peak current. In the case of zero quiescent current and an assumed straight-line characteristic as indicated in Fig. 4a, the output of a single tube is a series of half-sine waves in which the average current is I_m/π and the rms current is $I_m/2$. In the case of a practical tube, the output consists of a sine wave of current slightly distorted and with part of the lower peak cut off as shown in Fig. 4b. E. W. Herold[3] gives an equation for I_m in terms of I_{dc} and I_{bo}, as follows:

$$I_m = \pi\,(I_{dc} - 0.25\,I_{bo})$$

Substituting this equation in the one above for the single-tube power output, we obtain

$$P.O._{one\ tube} = \frac{\pi^2}{4}(I_{dc} - 0.25\,I_{bo})^2 R$$

---

[2] M. V. Callendar, "A Theory of Available Output and Optimum Operating Conditions for Triode Valves," Proc. I.R.E., Vol. 21, No.7, p. 909; July, 1933.

[3] Ref. 20-3.

This formula leads to an error not exceeding 3 per cent in all cases where the distortion is less than 10 per cent and where the range for the ratio I_{dc}/I_{bo} is between 1.1 and infinity. For values ordinarily encountered, the error is probably under 1 per cent.

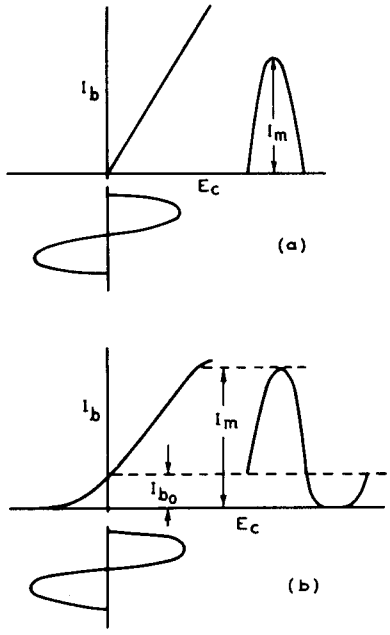


Fig. 4

SUPPRESSOR-GRID DESIGN

As has been stated previously, the screen-grid tube was the first advance over the simple triode. The screen grid, placed between the control grid and the plate and maintained at a fixed positive potential served two purposes. These were (1) to reduce grid-to-plate capacitance and (2) to maintain a fixed positive field independent of plate-voltage changes resulting from the load. In output tubes, the reduction of grid-to-plate capacitance is not particularly significant, but having the screen grid at a fixed positive potential allows a larger current to flow to the plate at low plate voltages. The first commercial tubes using a screen grid were the types 22 and 24. The plate family for each of these types was similar to that of the type 24-A which is shown in Fig. 5. It can be seen from this family of curves that at plate voltages less than that of the screen, secondary electrons from the plate flow back to the screen. If an output tube were made with characteristics similar to this with no precautions taken to reduce the effect of secondary emission from the plate, distortion and probably oscillation would result.

There are several methods available which may be used either singly or in combination to reduce the effect of secondary emission.[4] One method is to roughen the plate surface so as to present a number of small pits which tend to trap secondary electrons. The most satisfactory roughened surface for this purpose is a heavy carbon coating produced by a sooty flame. Such a carbon coating in addition to its trapping action also is effective in reducing secondary emission because of its low emissivity.

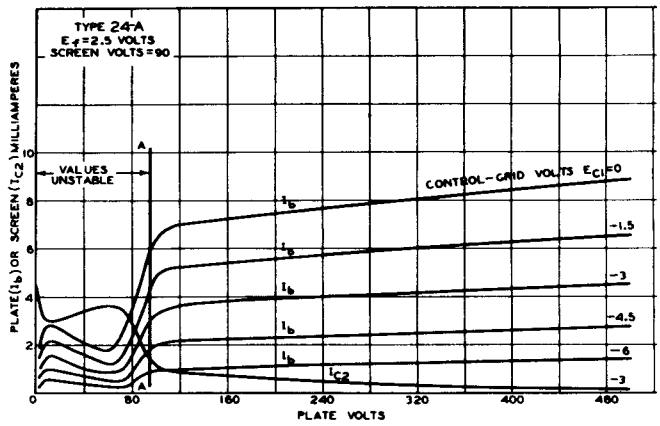


Fig. 5

A second method consists of placing fins on that surface of the plate toward which the primary electrons are accelerated. This method is effective because many of the released secondary electrons leave the surface at such an angle as to strike the fins on which they are collected. The cavities formed by the fins are similar in their action to the pits described in the previous method.

A third method is to prevent the secondaries from hitting the accelerating electrode by using a construction so designed that the high velocity of the secondaries causes them to trace an orbit which fails to touch the screen grid. A tube with a large diameter plate and a small diameter coaxial screen grid accomplishes this purpose.

A fourth method and one which is now being used more, involves the creation of a virtual cathode between the screen and plate, as in the 6L6 and similar beam power tubes.

A fifth method utilizes a suppressor grid between the screen and plate. Such a grid, which is held at a low potential (usually tied to the cathode) serves to reduce the field at the plate, and thereby effectually retards all the secondaries except those with sufficiently high initial velocity. This method will be discussed in detail.

If the space charge between screen and plate is neglected, the factor governing the number of secondaries which can pass to the screen is the retarding potential, which is the difference in potential between the plate and suppressor grid. The potential of this grid is a function not merely of its applied potential but also of the space potential at the position of this grid, if this

[4] Ref. 20-4.

grid were not present. It can be shown from electrostatic principles that the effective potential E_g of a grid between two other electrodes (see Fig. 6), if space charge is neglected, is given by the expression

$$E_g = \frac{\frac{E_{p1}}{\mu_1} + E_g + \frac{E_{p2}}{\mu_2}}{\frac{1}{\mu_1} + 1 + \frac{1}{\mu_2}}$$

where E_{p1} and E_{p2} are potentials of plates No.1 and No.2, respectively, and μ_1 and μ_2 are the amplification factors of the grid to plates No.1 and No.2, respectively.

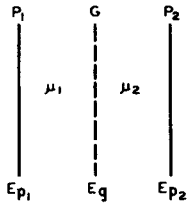


Fig. 6

Suppose that it is desired to suppress secondary electrons from plate No.2 and to impede the flow of primary electrons from plate No.1, which we will assume is the screen grid, to plate No.2 as little as possible. The retarding potential of the grid to secondaries from plate No.2 is then:

$$E_{p2} - E_g = E_{p2} - \frac{\frac{E_{p1}}{\mu_1} + E_g + \frac{E_{p2}}{\mu_2}}{\frac{1}{\mu_1} + 1 + \frac{1}{\mu_2}}$$

$$= E_{p2}\,\frac{1 + \mu_1}{1 + \mu_1 + \frac{\mu_1}{\mu_2}} - \frac{E_{p1} + E_g\mu_1}{1 + \mu_1 + \frac{1}{\mu_2}}$$

For this quantity to be as great as possible

$$E_{p2}(1 + \mu_1) \gg E_{p1} + E_g\mu_1$$

and in the usual case when the suppressor grid is at zero potential

$$E_{p2}(1 + \mu_1) \gg E_{p1}$$

From this it follows that μ_1 should be as large as possible to prevent the flow of secondary electrons from plate No.2 to plate No.1 which we have assumed is the screen grid.

By a similar analysis, it is easily shown that in order to allow primary electrons to pass readily from plate No.1 (the screen grid) to plate No.2, μ_2 should be as small as possible. Geometrically, these two conditions can be satisfied by placing the suppressor grid as close as possible to plate No.2. The actual required μ from suppressor grid to plate is best determined by trial. In Fig. 7, curve (a) shows the effect of a suppressor grid with a μ lower and (b) shows the effect of a grid with μ higher than required for optimum suppression of secondary electrons from plate to screen grid.

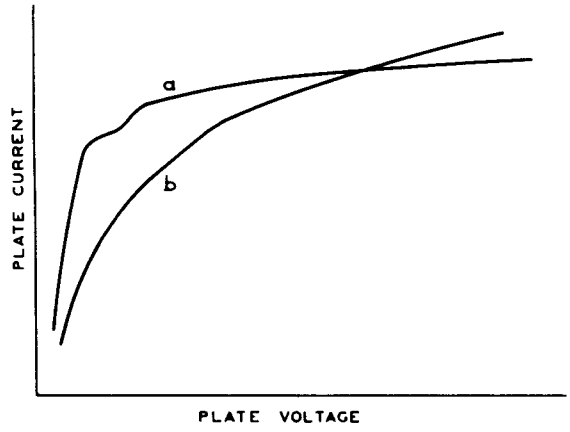


Fig. 7

BEAM POWER TUBES

A paper by O. H. Schade[5] gives the theory and the design of beam power tubes with space-charge suppression of secondary-emission effects. It also gives the general requirements of an ideal output tube as:

1) Low distortion consisting mostly of second harmonic with very little third and negligible higher-order harmonics.

2) Good power sensitivity to permit low-level operation of the preamplifier stage.

3) High power output with self-bias and supply circuits having the voltage regulation of conventional receivers.

4) Maximum efficiency in both tube and associated circuits with respect to power dissipation as well as cost.

5) Effective damping of resonant load.

These requirements are more nearly met by the new beam power tubes than by any of the other output tubes.

[5] O. H. Schade, "Beam Power Tubes," Proc. I.R.E., Vol. 26, No.2, p. 137; February, 1938.

EFFECT OF CONTROL-GRID WIRE DIAMETER

It is possible to obtain the same μ value by using either a heavy wire with open pitch (solid curves in Figs. 8 and 9) or a fine wire with closer pitch (dashed curves in Figs. 8 and 9). The dashed curves indicate much more desirable elec-

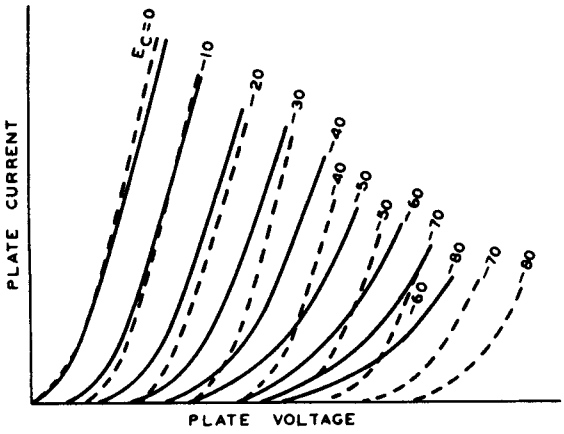


Fig. 8

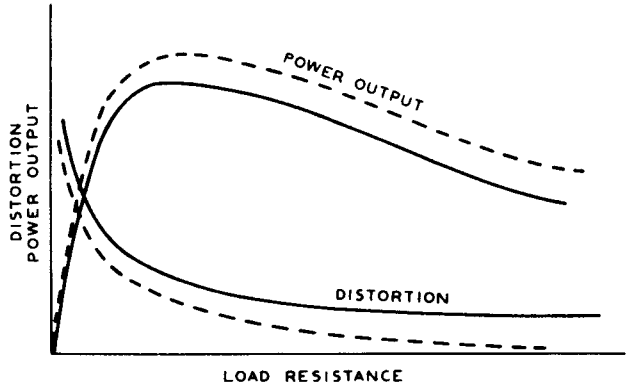


Fig. 9

trical characteristics with greater power output and reduced distortion. However, as in most features of tube design, a compromise must be made between mechanical strength of the grid and excellence of the electrical characteristics.

CONTACT POTENTIAL

Variations in grid-to-cathode contact potential are most serious in tubes with low operating bias (high-mu tubes) and high transconductance. In such tubes, this contact potential is probably the greatest single cause of variation in characteristics. For example, in a tube designed to operate with a plate current of 2 milliamperes and a transconductance of 1200 micromhos, a change of 0.2 volt in the grid-to-cathode contact potential results in a deviation of 12 per cent in the plate current.

In high-mu tubes, such as type 75 which operates normally at a bias of -1.5 volts or less, variations in contact potential not only alter the plate current but also materially affect the magnitude of the signal which may be applied to the control grid before grid current starts to flow. A change in contact potential from 0.3 to 1.1 volts would reduce the peak input signal from 1.2 volts to 0.4 volt, or to one-third of the largest signal which may be applied to the grid of a tube with the lowest contact potential.

Contact potential has been the subject of considerable investigation but at the present time there has been no practical method put into use which economically controls it closer than ± 0.3 volt.

GRID EMISSION

In conventional tubes with indirectly heated cathodes and in power-output tubes, grid emission is difficult to prevent. Carbonization or other treatment of the plate surface to increase radiation, and grid-radiator collars are used to remove heat from the grid. Grid side-rods of copper and extra leads in the press welded to the grid are means used to conduct heat away from the grid wires. Low-temperature cathodes, increased spacing between cathode and grid, and grids plated with gold and some other metals, all tend to reduce grid emission. Benjamin, Cosgrove, and Warren [6] state that the only sure method at present known for preventing grid emission is to keep the grid-wire temperature below 320°C.

---

[6] Benjamin, Cosgrove, and Warren, "Modern Receiving Valves: Design and Manufacture," Jour. I.E.E., Vol. 80, No.484, p. 412; April, 1937.

Lecture 21

THE DESIGN OF RADIO-FREQUENCY AMPLIFIER TUBES

T. J. Henry

INTRODUCTION

For good receiver performance, it is desirable that r-f and i-f amplifier tubes have high transconductance, low signal grid-to-plate capacitance, high plate resistance, and low r-f distortion. For economic reasons, we are also interested in low heater power, low cathode current, small tube size, and ease of tube manufacture. At very high frequencies, low input and output capacitances and low transit-time loading are also essential. The relative importance of these various factors depends on particular circuit requirements and an engineering compromise must be made among them to determine a satisfactory design.

In this lecture we will discuss briefly several of the above characteristics and indicate approximate methods of obtaining them. Exact design formulas are, in general, not available and so the designer must proceed by the cut-and-try method using his past experience and the available formulas to tell him what alterations to make in order to obtain the desired characteristics.

The desired characteristics are most nearly obtained in the r-f pentode.[1] This type can be made with high transconductance, high plate resistance, low grid-plate capacitance, and low r-f distortion. The tetrode can also be made to meet prescribed characteristics provided the plate voltage is not permitted to swing below the screen-grid voltage into the region where secondary electrons flow from plate to screen grid. When the plate voltage of the tetrode is higher than the screen potential, secondary electrons flow from the screen grid to the plate. The magnitude of this secondary-emission current varies with plate voltage and reduces the plate resistance below that which would obtain if no secondary emission occurred. Due to differences in the surface conditions of screen grids, there are wide variations in screen-grid current between tubes and in individual tubes during life. These variations make it impractical to use series screen-grid resistors with tetrodes in set design. In the case of pentodes, however, these variations do not exist, and so the use of series screen-grid resistors is practical. Since the pentode has higher plate resistance and more uniform screen current than the tetrode and can be operated with equal screen and plate voltages, it is practically the only type used for r-f work in the modern receiver.

[1] E. W. Ritter, "The Radio-Frequency Pentode: A New Use of the Suppressor Grid," Electronics, pp. 10-11; January, 1932.

TRANSCONDUCTANCE

The equation for plate current in an ideal parallel-plane triode is

$$I_b = G\left[\frac{\frac{E_b}{\mu} + E_c + \varepsilon}{1 + \frac{1}{\mu} + \frac{4}{3}\frac{b}{\mu a}}\right]^{\frac{3}{2}}$$

where,

I_b = plate current (when E_c is negative)
E_b = plate voltage
E_c = applied grid bias
ε = internal grid bias
μ = amplification factor (grid to plate)
a = distance between grid and cathode in centimeters
b = distance between grid and plate in centimeters

$$G = \frac{2.33 \times 10^{-6} \times \text{cathode area in sq cm}}{(\text{cathode-to-grid distance})^2} = \frac{2.33 \times 10^{-6} A}{a^2}$$

Designating $\left[1 + \frac{1}{\mu} + \frac{4}{3}\frac{b}{\mu a}\right]$ as α,

and differentiating with respect to E_c, we have

$$I_b = \frac{G}{\alpha^{\frac{3}{2}}}\left[\frac{E_b}{\mu} + E_c + \varepsilon\right]^{\frac{3}{2}}$$

$$\frac{dI_b}{dE_c} = g_m = \frac{3}{2}\frac{G}{\alpha^{\frac{3}{2}}}\left[\frac{E_b}{\mu} + E_c + \varepsilon\right]^{\frac{1}{2}}$$

It can be seen from the above formulas that the g_m can be made larger by decreasing the cathode-to-grid distance, or by increasing the cathode area, or the effective voltage. A decrease of n % in a increases g_m to $[100/(100-n)]^2$%. An increase of n % in the cathode area increases g_m by n %. In both cases the I_b increases linearly with g_m. An increase of n % in effective voltage increases g_m to $(100 + n)^{1/2}$ % and I_b, to $(100 + n)^{3/2}$ %. Increasing the effective voltage is the least efficient method of the three. The supply voltages and cathode wattage (and area) are limited by the set requirements.

The value chosen for the amplification factor may be determined by a number of considerations.

It is usually made fairly high (about 20) in r-f tubes in order to obtain a high ratio of g_m/I_b, since

$$\frac{g_m}{I_b} = \frac{3}{2} \frac{1}{\frac{E_b}{\mu} + E_c + \varepsilon}$$

However, the choice may also depend on the cut-off voltage desired or on the magnitude of g_m required.

The grid-to-cathode distance can be decreased only to a certain point as determined by difficulty of manufacture and lack of uniformity. The usual practice in ordinary receiving tubes having a grid length of approximately one inch is to use a grid-to-cathode distance of from .008" to .015", but spacing less than .010" increases the mounting difficulties considerably. In special types, such as the acorn tube and 1851 for use at very high frequencies, this spacing is reduced to .005" or less to obtain high g_m and to decrease transit-time loading. In these types, the characteristics of the tubes begin to depart considerably from the 3/2-power relation owing to the non-uniformity of the potential along the cathode and possibly to the effect of the position of the virtual cathode.

In tube design[2,3] it is desirable to keep the ratio of the spacing between adjacent turns of the grid wire to the grid-to-cathode distance less than unity. This is done to improve the uniformity of the field acting on the cathode. In the case of tubes having a very small grid-to-cathode distance, the ratio can be kept less than unity only by the use of finer grid wire which increases the mechanical problems of grid winding and mounting. Wires down to .001" in diameter have been used for grids, but .0033" is the lowest size which is ordinarily used.

By decreasing the distance from grid to plate and increasing the grid turns per inch to keep the same μ, b/a and consequently α can be reduced with a resultant gain in g_m. The increase in turns per inch also decreases the ratio mentioned in the previous paragraph, and increases the uniformity of the field at the cathode if the plate-to-grid spacing is not made too small. The reduction in this spacing has the disadvantage that a .001" change will cause a larger percentage change in μ, and hence smaller tolerances for these parts may be required.

For the calculation of g_m and I_b in a pentode, the following approximate formulas are usually satisfactory:

$$I_k = G'(V)^{\frac{3}{2}}$$

$$I_b = SG'(V)^{\frac{3}{2}}$$

$$g_m = \frac{3}{2} SG'V^{\frac{1}{2}} + \left[G'V^{\frac{3}{2}} \frac{dS}{dE_{c_1}} \right] \approx SG'V^{\frac{1}{2}}$$

$$\frac{g_m}{I_b} = \frac{3}{2} \frac{1}{V}$$

where,

I_k = cathode current

$G' = p \frac{G}{\alpha^{\frac{3}{2}}}$

p = the useful fraction of the calculated perveance < 1

$S = \frac{I_b}{I_k}$

V = effective potential $\approx \frac{E_c}{\mu_{1-2}} + E_{c_1} + \varepsilon$

$\alpha = 1 + \frac{1}{\mu_{1-2}} + \frac{4}{3} \frac{b}{\mu_{1-2}\, a}$

A more accurate formula for the ideal tetrode or pentode effective potentials may be derived but since the geometry of actual tubes is seldom either cylindrical or plane, the increase in accuracy is generally not sufficient to justify the use of a more complicated formula.

The quantity dS/dE_{c1} is negligible,[4] since the distribution of space current between screen and plate is not appreciably affected by the No.1 grid potential provided there are no focusing effects or secondary space charges.

The factor p depends upon the percentage of the cathode area cut off by the field of the grid side-rods and on the deviations from the perveance as calculated with the formula for the parallel-plane or cylindrical case due to the variation in grid-cathode spacings such as occur when an elliptical grid is used with a round cathode. The value of p can be roughly estimated from a study of the cross section through grid and cathode and from experience with constructions having similar grid and cathode shapes. The use of wider center-to-center distances of the No.1 grid side-rods and of smaller diameter side-rods will to a certain extent increase p.

The above formulas are derived from the parallel-plane case but can be used for tubes with a

---

2 Benjamin, Cosgrove, and Warren, "Modern Receiving Valves: Design and Manufacture," Jour. I.E.E., Vol.80, No.484, pp.401-431; April, 1937.

3 L. Oertel, "On the Theory of Valves whose Grid-Cathode Spacing is Smaller than the Pitch of the Grid," Telefunken-Röhre, No.12, pp. 7-17; April, 1938.

4 H. Rothe and W. Kleen, "Current Distribution: Theory in Multi-grid Tubes," Telefunken-Röhre, pp. 1-23; March, 1936.

cylindrical cathode, and with elliptical, cylindrical, or flat grids with fair accuracy, except that the perveance will vary considerably from that of the ideal case.

The value of S, the plate-to-cathode current ratio, may be roughly approximated by

$$S = 1 - (TPI \times d)$$

where,

TPI = turns per inch of screen grid

d = screen-grid wire diameter in inches

CAPACITANCE FROM SIGNAL GRID TO PLATE

The effect of feedback capacitance in a tuned-grid, tuned-plate amplifier has been treated in a number of articles[5,6] by assuming that the tube characteristic was linear. For a given grid-to-plate capacitance, the increase in gain we can obtain through an improvement in circuit constants or from an increase in g_m is limited in practice, when we reach the maximum amount of regeneration which can be tolerated. The form of the equation derived in the references for a single-stage amplifier is

$$C_{gp} = \frac{2g_o G}{\omega g_m}$$

where g_o and G are the respective equivalent conductances of the output and input circuits, ω is 2π times the frequency, and g_m is the transconductance of the tube.

This formula is based on the starting point of oscillation and it is usually necessary to use a C_{gp} of 1/4 or less of the value from the formula, depending on the particular circuit requirements. As an example, in ordinary receiver design, a tube with a plate resistance of 1 megohm and a g_m of 2000 micromhos should have a C_{g1p} of less than .005 μμf. With this capacitance, an increase in g_m above 2000 micromhos will not permit any appreciable increase in gain, unless some method of neutralization is employed. The allowable value is best determined by application tests. In higher frequency circuits where the load impedances are not usually as great, larger values of C_{g1p} can be tolerated, and higher g_m tubes can be used to advantage.

Low capacitance from No.1 grid to plate is obtained by using screen-grid tubes with closely wound screens at zero r-f potential. The No.1 grid lead and plate lead must also be carefully shielded outside of the mount itself. The capacitance within the active tube electrodes can be made as low as desired by winding the screen grid more closely. However, the proportion of cathode current which is absorbed by the screen increases with the turns per inch, and the g_m of the tube is reduced proportionately. Some further reduction in C_{g1p} may be obtained by cutting down the area of the plate to the useful portion in the path of the electron stream.

In pentodes, the suppressor grid also contributes to the reduction in capacitance. In a parallel-plane type of structure, the suppressor should be located about half-way between screen grid and plate for minimum capacitance. No optimum position has been determined for the cylindrical case, but since variations in the diameter of the suppressor do not seem to be critical, a spacing nearly half-way between No.2 grid and plate will be satisfactory.

In the conventional r-f pentodes of the unipotential-cathode type made in this country, the screen current is about .20 of the total cathode current and the capacitances are in the order of .005 μμf in glass tubes, and somewhat lower in metal tubes because of the better shielding. Some European pentodes have a screen-to-cathode current ratio of 0.25 to 0.30 and somewhat lower capacitances, but this reduction may be due in part to other variations in construction than the turns per inch on the screen. It has been observed that an increase in the ratio of screen current to plate current in a pentode causes an increase in the fluctuation noise in the plate circuit of the tube.[7]

It is not very practical to calculate the actual capacitance of a tube structure because the field through the mount is non-uniform due to the side-rods. The best way is to determine the capacitances on actual tubes and make such changes in the parts as are necessary.

PLATE RESISTANCE

In the simple linear theory of the vacuum tube, with capacitances neglected, the voltage amplification is:

$$V.A. = \frac{e_{output}}{e_{input}} = g_m \frac{R}{1 + \frac{R}{r_p}}$$

where R is the load resistance and r_p is the internal plate resistance, dE_b/dI_b. In order to realize high gains, it is desirable to keep R/r_p small. In i-f amplifiers, the external load impedance may be as high as 0.2 megohm and r_p should be 0.8 megohm or higher. If the external load is relatively low, as in television circuits, r_p need not be as high as 0.8 megohm.

---

[5] B. J. Thompson, "Oscillations in Tuned Radio-Frequency Amplifiers," Proc. I.R.E., Vol. 19, No.3, p. 421; March, 1931.

[6] Bernard Salzberg, "Notes on the Theory of the Single-Stage Amplifier," Proc. I.R.E., Vol. 24, No.6, pp. 879-897; June, 1936.

[7] H. Rothe and W. Engbert, "Noise in Receiving Tubes in the Region from 150 Kc to 15 Mc," Telefunken-Röhre, No.11, p. 183; December, 1937.

The value of r_p[4,8] is affected by the screen and plate voltages at which an r-f or i-f amplifier tube is operated. In modern a-c receivers, E_{c2} is about 100 volts and E_b is about 250 volts. In a-c/d-c receivers, the screen voltage equals the plate voltage and is approximately 100 volts. Regardless of whether the plate and the screen are operated at the same voltage or whether the plate is operated at a higher voltage than the screen, it is important to keep r_p as high as possible.

A high value of r_p is indicated in a pentode when its E_b vs I_b curves show a steep rise as E_b is increased from zero, make a sharp knee, and continue almost parallel to the horizontal axis. As shown in Fig. 4 of Lecture 20, if the suppressor grid has insufficient turns per inch, a secondary-emission dip may occur in the E_b vs I_b curve just after the knee; and if the suppressor has too many turns per inch, the knee of the curve will not be sharp. In either case, too low a value of r_p will occur at low plate voltages. It should be noted that it is not advisable to place the suppressor grid close to the plate as in output pentodes. In r-f pentodes, the value of C_{g1p} is important, and so a position half-way between screen and plate is chosen. In order to design a suppressor grid which will provide the desired characteristics, it will be found helpful to consider the suppressor-screen mu-factor as explained in Lecture 20. An investigation of the values of μ_{g3p} and μ_{g3-2} on several types of r-f pentodes designed for 100-volt screen-grid operation showed that μ_{g3p} and μ_{g3-2} were roughly equal and had an average value of about 7.5, varying from 4 to 10. This average figure might well be used as an initial trial value in design work when the construction is such that a standard suppressor grid cannot be used.

The design of a suppressor grid is facilitated by making a trial tube in which the suppressor lead is brought out. On this trial tube, it is convenient to determine the magnitude and direction of any necessary change in the turns per inch to obtain the best shape of E_b-I_b curves. This can be done by varying the bias on the suppressor and observing the effect of this variation on the E_b-I_b family as seen on the curve-tracer (or by measuring r_p under various voltage conditions). After the best bias on the suppressor grid has been found, it is possible to calculate a new grid which when operated at zero bias with respect to the cathode, will provide the same effective potential in the suppressor plane as the suppressor grid of the trial tube with "best" bias. This method may give inaccurate results if the suppressor grid is very far wrong in initial design.

Other factors besides optimum plate resistance and capacitance may determine the design of the suppressor grid. In some types, such as the 6K7 in which the suppressor grid is brought out to a separate base pin, over-suppression is employed so that an oscillator voltage may be applied to this grid where the tube is used as a mixer in a superheterodyne circuit, without requiring an excessive oscillator voltage for good performance.

Another factor which lowers plate resistance is the so-called "bulb effect." When an insulator with good secondary-emission properties, such as a glass bulb-wall, is near the plate and if through some transient effect it is struck by stray electrons, more electrons may be emitted from the insulator than strike it. As a result, a positive charge will be built up on the insulator surface. The potential of the insulator will approach that of the plate until an equilibrium condition is reached between the number of electrons striking and leaving the insulator per unit of time. The plate current then consists of two components, i_b through the grid array from the cathode, and i'_b from the insulator. When the plate voltage fluctuates at radio frequency, dE_b/di'_b may be low enough to reduce the total plate resistance considerably, sometimes to 1/5 of the value due to dE_b/di_b alone. This reduction in r_p does not always appear in the conventional r_p bridge measurement and should be checked in actual r-f circuits.

The secondary emission from glass bulbs is eliminated or reduced to a negligible amount by spraying the inside of the bulb with carbon. Secondary emission can, however, occur with other insulators than the glass bulb-wall. In one developmental construction, a very marked effect was obtained from a mica near the plate and in the path of the electron stream.

RADIO-FREQUENCY DISTORTION

It has been shown[9,10,11] that if the plate current of an r-f amplifier tube is other than a linear or quadratic function of the grid voltage, several forms of r-f distortion may occur. When an r-f carrier modulated by an a-f signal of modulation factor m is applied to the grid of a tube with other than a linear or quadratic char-

---

4 Loc. cit.

8 O. H. Schade, "Beam Power Tubes," Proc. I.R.E., Vol. 26, No.2, pp. 137-181; February, 1938.

9 S. Ballantine and H. A. Snow, "Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-Mu Tetrodes," Proc. I.R.E., Vol. 18, No.12, pp. 1377-1387; December, 1930.

10 R. O. Carter, "The Theory of Distortion in Screen-Grid Valves," Wireless Engineer, Vol. 9, No.8, pp. 429-438; August, 1932.

11 C. S. Bull, "Non-Linear Valve Characteristics; A Brief Discussion of Their Use," Wireless Engineer, Vol. 10, No.2, pp. 83-88; February, 1933.

acteristic, the resulting a-c output current (or voltage across a small load) will contain harmonics of the audio modulation frequency. A rise in the modulation factor will also occur.

When $\underline{m}$ is small, the ratio of second harmonic to the fundamental in the a-f modulation is

$$\frac{3}{16}\ m\ \frac{g_m''}{g_m}\ E^2$$

where E is the maximum amplitude of the carrier voltage, and

$$g_m'' = \frac{d^2 g_m}{de_g^2}$$

where e_g is the signal voltage. When $\underline{m}$ approaches unity, we have

$$\frac{3}{16}\ \frac{g_m''}{g_m}\ E^2$$

Similarly, a phenomenon known as cross-modulation is produced by the interaction of a modulated unwanted carrier and a wanted carrier. The wanted carrier E_w becomes modulated with the audio frequency of the unwanted carrier E_u and no subsequent selective circuits can remove this interference. This cross-modulation generally occurs in the first tube, because unwanted carriers are greatly attenuated in the later stages. When E_u is small compared with E_w, the modulation of the wanted by the unwanted carrier is approximately

$$M = \frac{g_m''}{g_m}\ E_u^2\ m_u$$

In order to obtain constant cross-modulation at any bias when E_u and m_u are constant, g_m''/g_m must be constant, and

$$g_m = g_{m_o}\varepsilon^{b e_g}$$

where g_{m_o} and b are constants, $\varepsilon = 2.718$, and e_g is the grid voltage. If we plot log g_m vs e_g, we obtain a straight line with a slope $\underline{b}$ having an intercept at the Y-axis of $g_m = g_{m_o}$, as shown in Fig. 1.

As demonstrated by Ballantine and Snow,[9] an exponential characteristic will give high modulation rise at high signal strengths, unless excessive values of plate current are used. It is, therefore, usual to design the e_g vs g_m characteristic so that it is concave downwards when plotted on a semi-logarithmic scale, and so that

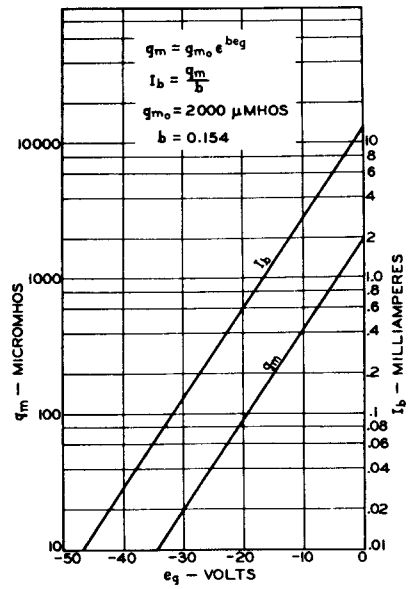


Fig. 1

g_m''/g_m is smaller at the higher grid biases at which the tube will be working with a large signal in a set with an avc system.

Let it be assumed that the screen grid is supplied from a high voltage through a series resistor of the proper value to give the usual screen voltage at minimum negative control-grid bias. Then, when the bias is made more negative, the screen voltage will rise due to the regulation and the cut-off bias of the tube will be increased. This method may be used to obtain a lower value of g_m''/g_m without increasing the plate current at minimum grid bias. It has the disadvantage that higher avc biases are necessary to obtain the same decrease in gain.

In Fig. 2 are shown the curves of g_m vs E_{g1} on a semi-logarithmic scale for a remote cut-off r-f pentode with and without a series screen resistor. In Fig. 3 are shown the carrier voltages which may be applied to the control grid before 4.5% second-harmonic audio distortion occurs plotted against the percentage of maximum transconductance at which the tube is working. It can be seen from these figures that actual tubes depart considerably from the exponential law. There are a number of methods of determining the r-f distortion of actual tubes. A number of these are outlined by Strutt.[12]

The design of a tube to have a remote cut-off characteristic is accomplished by varying the turns per inch of the control grid along its length so that various sections of the grid cut off at different biases. The grid may be considered as made up of a number of constant-mu sections of different lengths and with per-

[9] Loc. cit.

[12] M. J. O. Strutt, "Moderne Mehrgitter-Elektronenröhren," Julius Springer, Berlin, pp. 18-20.

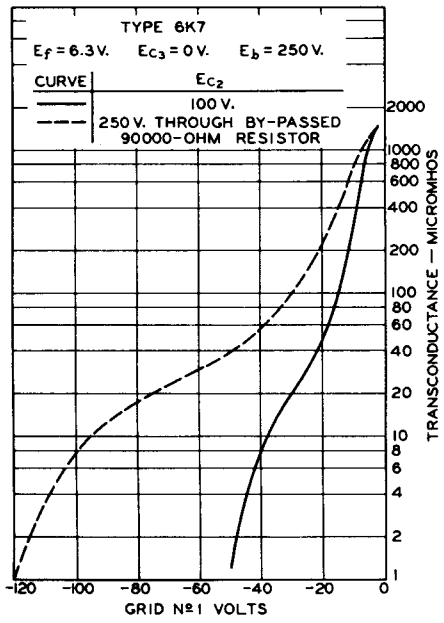


Fig. 2

veances proportional to their lengths. The e_g vs g_m curves may then be calculated and added together to give a resultant curve. By changing the lengths and mu's of the various sections, it is possible to obtain a curve approaching the desired characteristic. An experienced engineer can generally obtain a reasonable design in one or two trial tests.

Several articles[13,14] have been written on the design of variable-pitch grids by mathematical analysis, but these methods are tedious and, in general, not as useful as the simpler cut-and-try method.

ULTRA-HIGH FREQUENCIES

At very high frequencies, the input and output capacitance of a tube are a considerable portion of the tuned-circuit capacitance and limit the possible L/C ratios and the tuning range. These capacitances may be reduced[15,16] by reducing all tube dimensions proportionately without altering the I_b, g_m, etc. until initial velocities of the electron become an important factor. The acorn tubes have been designed on this principle

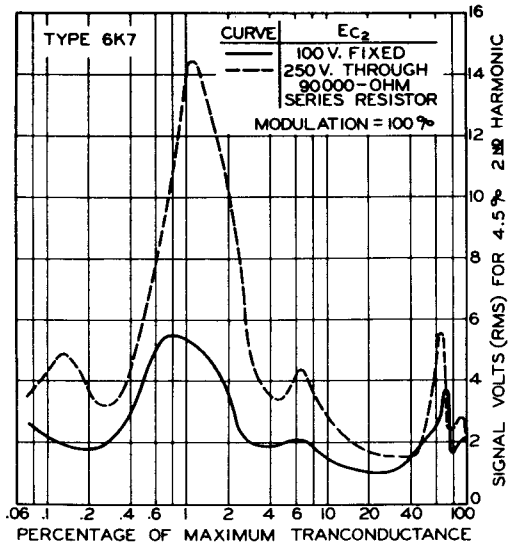


Fig. 3

although all the dimensions could not be reduced in proportion because of mechanical limitations.

In larger tubes, the output capacitance may be reduced by cutting down the plate area until only the electron collecting area is left. Reduction of input capacitance is a difficult problem because the area is largely determined by the size of the cathode, and because the spacings must be close not only to reduce transit time[17] but also to provide high g_m. High g_m is necessary if any gain is to be realized, because, as mentioned previously, load impedances are usually small at very high frequencies. The use of secondary-emission multiplication[18] has been suggested as a means of obtaining high g_m without large cathode areas and extremely small spacings.

Other factors such as the effect of the inductances of the electrode leads and of the electrodes themselves become important at ultra-high frequencies and have been discussed by Strutt and Van der Ziel.[19]

---

[13] G. Jobst, "On the Relation between 'Durchgriff' ($1/\mu$) and the Space-Charge Law in Tubes with Variable 'Durchgriff'," Telefunken-Zeitung, Vol. 12, pp. 29-44; 1931.

[14] J. E. Scheil and F. Marguerre, "On the Theory of Electron Tubes with Varying 'Durchgriff' along the Axis of the System," Arch. Elektrotechn., Vol. 28, pp. 210-233; 1934.

[15] B. J. Thompson and G. M. Rose, "Vacuum Tubes of Small Dimensions for Use at Extremely High Frequencies," Proc. I.R.E., Vol. 21, No.12, pp. 1707-1722; December, 1933.

[16] B. Salzberg and D. G. Burnside, "Recent Developments in Miniature Tubes," Proc. I.R.E., Vol. 23, No.10, pp. 1142-1157; October, 1935.

[17] W. R. Ferris, "Input Resistance of Vacuum Tubes as Ultra-High-Frequency Amplifiers," Proc. I.R.E., Vol. 24, No.1, pp. 82-105; January, 1936.

[18] J. L. H. Jonker and A. J. W. M. van Overbeek, "The Application of Secondary Emission in Amplifying Valves," Wireless Engineer, Vol. 15, No.3, pp. 150-156; March, 1938.

[19] M. J. O. Strutt and A. Van der Ziel, "The Causes for the Increase in the Admittances of Modern High-Frequency Amplifier Tubes on Short Waves," Proc. I.R.E., Vol. 26, No.8, pp. 1011-1031; August, 1938.

Lecture 22

THE DESIGN OF DETECTORS AND CONVERTERS

T. J. Henry

DIODE DETECTORS

The discussion of detectors will be confined to diodes, since these are almost universally used as detectors in radio receivers to-day.

The function of the detector is to demodulate r-f or i-f signals. Consider an amplitude-modulated signal, $e = E(1 + m \cos \alpha t) \cos \omega t$, where ω is 2π times some radio frequency, α is 2π times an audio frequency, and m is a modulation factor less than 1. If this signal is applied to the circuit shown in Fig. 1, in which the condenser C_2 has a finite impedance to frequencies of the order $\alpha/2\pi$, and a negligible impedance to the frequency $\omega/2\pi$, the circuit can be regarded as that of a simple half-wave rectifier with supply voltage $E \cos \omega t$ varying in amplitude at the rate $\cos \alpha t$. Across the load resistance R will appear a d-c voltage with a superimposed a-c voltage $A \cos \alpha t$. This simplified explanation will serve for most tube-design considerations. More detailed treatments will be found in the literature.[1]

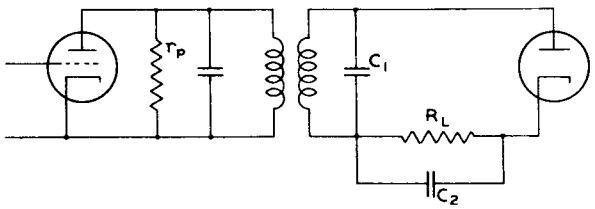


Fig. 1

In tube design we are interested in making a tube that will not only furnish the peak currents which occur with maximum signal input, but that will also give good rectification efficiency. The rectification efficiency (D) may be defined as:

$$D = \frac{\text{d-c voltage across } R_L}{\text{peak a-c voltage across } C_1} = \frac{E_o}{\hat{E}_s} \qquad (1)$$

If the tuned circuits in Fig. 1 are in resonance, we can represent them approximately by the equivalent circuit shown in Fig. 2. The value of R_s, the equivalent resistance in series with the diode, would be rather difficult to evaluate, but by assuming that it is small compared with r_d, the equivalent diode resistance, it is possible to calculate rectification efficiencies, peak plate currents, and average currents for any circuit values and applied voltages, as explained in Lecture 19.

In the diode detector stage used in conventional broadcast receivers, the value of R_L, the load resistor, is made high compared with $(R_s + r_d)$. This is done to obtain high rectification efficiency and also to decrease the power which must be supplied to the diode by the preceding stage. However, too high a value of R_L can cause distortion due to the difference in the a-c and d-c impedance of the $R_L - C_2$ combination.[1] For most receivers the value of R_L is between 0.1 and 0.5 megohm.

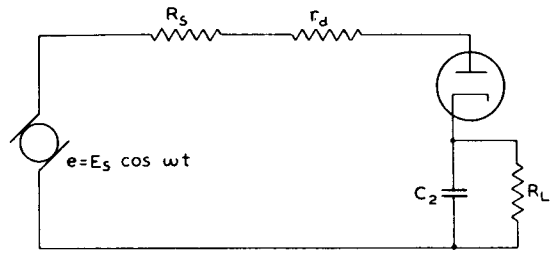


Fig. 2

On the basis of the resistance range for R_L, experience shows that a diode with an oxide-coated cathode which will give a plate current of over 1 milliampere at 10 volts on the plate and which will not be emission-limited below 2 milliamperes, is satisfactory. If we express this condition mathematically, we find that the diode must have a perveance of

$$0.032 \frac{\text{plate milliamperes}}{(\text{plate volts})^{\frac{3}{2}}}$$

A design to satisfy this relation can be very readily obtained in practice. Any increase in perveance over the value given by this relation will not add appreciably to the detector efficiency when the value of R_L is over 0.1 megohm.

Some television circuits may require a much lower load and in these special cases the use of rectifier design charts may be necessary for proper tube design. Some of these circuit requirements can be met by use of available power rectifiers, but special tubes with low capacitances and high perveance would be useful.

For ordinary broadcast receivers, the required perveance is so easily obtained that it is customary to combine the diode detector tube with an amplifier tube in one bulb. In this combination, two diodes are placed near the bottom of the cathode and either a triode or pentode above

[1] H. A. Wheeler, "Design Formulas for Diode Detectors," Proc. I.R.E., Vol. 26, pp. 745-780; June, 1938.

them. These double-function tubes effect a considerable saving in tube and socket cost. They do, however, have a critical feature not present in simple diodes.

If we have a diode and a triode on the same cathode in one envelope (as in Fig. 3), and an a-c signal is placed on the diode, some electrons during the time when the diode plate is positive will be accelerated towards the diode plate. The number depends upon the amplitude of the a-c signal voltage. Instead of being collected by the diode plate, these electrons may continue out into the space beyond the diode plate where they can be collected by the triode plate (or by its plate lead which comes down past the diode structure). The result is that a sort of triode is formed with the diode plate as the grid and the triode plate as the anode and that a distorted audio signal results across the load in the triode plate circuit. This distortion is particularly objectionable in receivers where the volume is controlled by a potentiometer arrangement across the diode load resistor R_L because the signal does not go to zero with the slider completely to the right, and the phase relations may be such that a null point occurs at some intermediate position of the slider.

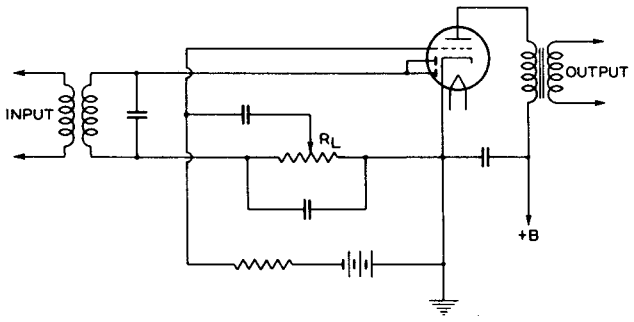


Fig. 3

This diode coupling effect may be effectually eliminated by shielding the diode section of the tube with a grounded shield so that very few stray electrons can escape. It might be thought that an ungrounded shield such as a mica would suffice, but from the results of experiments on a developmental type of diode-triode, it appears that grounded shielding is the safer method. In a circuit similar to that in Fig. 3 except that the grid is grounded, the amount of diode coupling may be determined by applying an a-c signal to the diode plate and reading the resultant a-c voltage across the triode plate load. Any new design of combination tube should be checked in this manner before the design is fixed.

Difficulties arising from coupling between the two diode units when each is used in a different circuit are apparently of little general importance.

Since the diodes are generally located on one end of the cathode where temperature conditions are none too uniform, and changes in processing may have considerable effect, it is wise to make the theoretical perveance at least twice that necessary to give the 1 milliampere at 10 volts mentioned above.

In addition, due to the position of the diodes and the variations in processing, variations in contact potential occur which may be a source of trouble. In order to have any sensitivity at low signal voltages, it is necessary that the contact potential be such that the diode draws a small amount of current with no applied plate voltage. In some types of circuits such as that of a diode-biased triode, it is important that the diode start to draw current before the triode grid. This circuit is used very little in present set design.

These diode-current starting points should be controlled by a periodic quality check to detect process changes which materially affect the diode-current cut-off.

In combination tubes the temperature of the diode plates is so low that no trouble has been experienced with primary emission from the plates. It would produce a reverse current on the negative cycle of the applied a-c voltage. However, in tubes where considerable cathode power is used, the diode plate temperature may be high enough to cause some primary emission and a reverse current will flow. This reverse current need not be large to cause decreased sensitivity on small signals and should be considered in the tube design. Process control or the use of plates with better radiating qualities, i.e., plates made of blackened, roughened material or with larger area, will prevent the reverse emission.

DESIGN OF CONVERTERS

In superheterodyne receivers, the incoming modulated r-f signal voltage of given frequency is applied to a stage in which a local oscillator voltage of some other frequency is produced. The difference between the radio and oscillator frequencies gives rise to an intermediate frequency which is amplified by the i-f stages. This intermediate frequency contains the same modulation components as does the r-f signal as well as some added distortion components.

If the frequency of the local oscillator is varied, any particular r-f signal within a given band may be used to produce an i-f signal of some constant frequency to which the i-f stages are tuned. The i-f stages can be carefully designed for the i-f carrier and its side bands so that a more efficient amplifier with better control of the selectivity can be obtained than one in which variation of the circuit constants of each stage is necessary for tuning. The tuning of the set is then accomplished by varying only the constants of the r-f circuit and the local oscillator circuit.

When two voltages of different frequencies f_o

and f_s are applied to any non-linear circuit, a number of resultant frequencies occur among which is the difference frequency $(f_o - f_s) = f_i$. If the output impedance of the non-linear circuit is large for f_i and negligible for the others, a voltage of frequency f_i alone is obtained.

The stage in which the intermediate frequency f_i is obtained is sometimes called the first-detector or modulator stage but is also called the converter stage. Tubes which act as the non-linear element in the circuit but are supplied by a separate oscillator tube will herein be called mixer tubes. Those types which perform both the function of oscillator and mixer will be called converter tubes.

Before the introduction of special mixers and converters, various circuits using triodes, tetrodes, or pentodes were used. With these, the oscillator voltage was generally introduced in the cathode circuit. These circuits had the disadvantage that the oscillator voltage was always capacitively coupled to the signal circuit so that when the signal circuit was not of negligible impedance to the oscillator frequency, a voltage of oscillator frequency occurred across the signal circuit and, if sufficiently high, produced such effects as degeneration, grid current, and radiation in the antenna circuit. Since the oscillator voltage was applied across the heater-cathode circuit, hum modulation was also introduced if heater-cathode insulation was poor. Furthermore, if the heater was not tight in the cathode, microphonics occurred due to varying heater-cathode capacitance.

Through the use of a screen grid to separate the oscillator and signal elements, this capacitive coupling can be reduced. Such a circuit can be obtained with an r-f pentode by introducing the signal on the No.1 grid and the oscillator voltage on the No.3 or suppressor grid. As will be shown later, in order to obtain maximum conversion gain, it is necessary to cut off completely the plate current during approximately one half of the oscillator-voltage cycle. Since the suppressor grid of the conventional r-f pentode has a low mu-factor, a fairly high amplitude of oscillator voltage is required to obtain the proper operating conditions. The use of a high oscillator voltage is undesirable because it is difficult to obtain uniformly over the operating range, and because it may induce undesired effects in other circuits. Another disadvantage of the r-f pentode for mixer use is that the plate resistance of this type of tube is low when a high negative bias is applied to the suppressor. The average plate resistance over a cycle of oscillator voltage is considerably lower than the load impedance normally used in the plate circuit, with the result that there is loss in gain.

1. Outer-Grid-Injection Mixers (Class I)

The pentode with suppressor-grid injection led to the design of the class of mixer tube which is the simplest from the viewpoint of design, namely, that in which the oscillator signal is introduced or injected by means of a grid further from the cathode than the signal grid (No.1). An example of this class is the 6L7. These outer-grid-injection-type mixers will be called class I mixers. As can be seen from Fig. 4, the class I mixer is similar to the pentode with oscillator

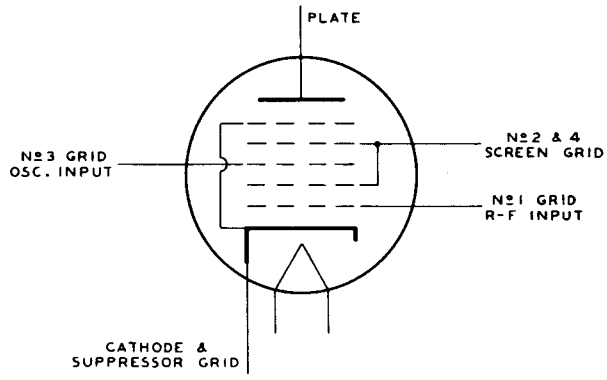


Fig. 4 - Outer-grid-injection mixer (6L7).

voltage on the No.3 grid except that an additional screen and suppressor have been introduced to increase the plate resistance.

A useful figure of merit for converter tubes is the conversion transconductance g_c, which is defined as follows:

$$g_c = \frac{i_{if} \text{ in the output circuit}}{e_{rf} \text{ in the input circuit}} \qquad (2)$$

when the r-f signal input voltage is small.

Conversion transconductance can be used to determine the conversion gain (C.G.) of the stage in the same manner as transconductance (g_m) is used for r-f pentodes.

$$C.G. = g_c \frac{R_L}{1 + \frac{R_L}{r_p}} \qquad (3)$$

where R_L is the load resistance and r_p is the effective plate resistance of the mixer tube with the oscillator voltage applied. The magnitude of r_p varies with the strength of oscillation, but under any given voltage conditions r_p may be defined as the inverse of the rate of change of d-c plate current with plate voltage. Expressed as an equation, this definition is

$$\frac{1}{r_p} = \frac{di_p}{de_p}$$

For the class I mixer the value of conversion transconductance can be shown to be[2]

$$g_c = \frac{1}{2\pi}\int_{\pi}^{\pi} g_{m_{1-p}} \cos \omega t \; d(\omega t) \qquad (4)$$

where g_{m1-p} is the transconductance from No.1 grid to plate and the voltage on the No.3 grid varies at the rate E cos ωt about a d-c bias E_o. In Fig. 5, g_{m1-p} is shown plotted as a function of E_{c3} for three different cases, as shown by curves A, B, and C. These cases will now be considered.

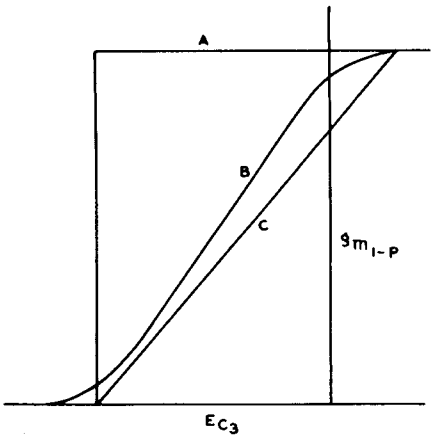


Fig. 5

In order for the integral determining g_c to be a maximum, the tube should operate only when cos ωt is positive. The g_m should be cut off at all angles of ωt between $\pi/2$ and $-\pi/2$ and should be a maximum during the interval $-\pi/2$ to $\pi/2$. The case represented by curve A fulfills these requirements when the d-c bias is at the value of E_{c3} at which g_{m1-p} vanishes. For this optimum case, the value of g_c is

$$\frac{g_{m_{1-p}(max)}}{\pi}$$

In practice, a curve of this shape does not appear to be possible, but a curve having the shape of curve B can be obtained. This curve will give a value of g_c between that of the case represented by curve A and that of the case represented by curve C, i.e., between

$$\frac{g_{m_{1-p}(max)}}{4} \quad \text{and} \quad \frac{g_{m_{1-p}(max)}}{\pi}$$

when $g_{m1-p}(max)$ occurs at the most positive excursion of the oscillator voltage and the d-c bias is approximately at g_m cut-off.

If a curve of the transconductance from signal grid to plate vs oscillator grid bias is available, it is possible to determine g_c for any bias and a-c amplitude on the oscillator grid by an approximate integration of equation (4).

Let $-E_o$ be the bias, and $E_c = (-E_o + E_m \cos \omega t)$ be the instantaneous voltage on the oscillator grid. Using the Trapezoidal Rule for approximate integration with 30° intervals for ωt, we obtain the following equation:

$$g_c = \frac{(g_5 + 1.73g_6 + g_7) - (g_1 + 1.73g_2 + g_3)}{12} \qquad (5)$$

where g_1, g_2, and so on, are the values of transconductance between signal grid and plate to be taken from the g_m-E_c curve at the values of E_c derived from the following table.

Signal-Grid-to-Plate Transconductance at $E_c = (-E_o + E_m \cos \omega t)$	cos ωt
g_1	1.000
g_2	0.866
g_3	0.500
g_5	-0.500
g_6	-0.866
g_7	-1.000

The voltage E_o in some mixer circuits is a fixed bias. In converter circuits and in a number of mixer circuits, E_o is the voltage drop (I_cR_c) across a grid leak in the oscillator grid. The rectification efficiency, E_o/E_m, in such a self-bias circuit depends on the circuit and the shape of the grid-current curve. In the usual converter circuits with oscillator grid leak of approximately 50000 ohms, efficiencies of 70 to 90 per cent are realized.

In order to make a mixer with high conversion transconductance, it is necessary to make g_{m1-p} high at or near the condition of the maximum positive oscillator excursion. The high value of g_{m1-p} is obtained in the same way as for r-f pentodes, namely, by increasing the perveance. The perveance can be increased by the use of a larger cathode area or by close cathode-to-No.1-grid spacing, and also by using close No.1-to-No.2-grid spacing. Since a high ratio of screen current to cathode current reduces the transconductance, and since we have two screen grids in the mixer, the reduction in transconductance is somewhat more than in the case of the r-f pentode. The choice of the turns per inch for the No.2 grid depends on the amount of shielding required for producing the desired capacitance as in the case of the r-f pentode. The turns per inch for

[2] C. F. Nesslage, E. W. Herold, and W. A. Harris, "A New Tube for Use in Superheterodyne Frequency Conversion Systems," Proc. I.R.E., Vol. 24, No.2, pp. 207-218; 1936.

the No.4 grid are selected to obtain high r_p by reducing the effect of plate voltage on the potential in the plane of the No.3 grid.

In order to keep the oscillator-voltage requirements within reason, the No.3 grid should have a fairly high mu-factor; and as a result, more current is returned to the No.2 grid than would return if the No.3 grid were designed only for suppression.

In mixer service at high frequencies, a transit time effect occurs which may cause electron current to flow to the No.1 grid even at negative potentials.[2,3] When the No.3 grid is swinging negative, some of the electrons approaching it are turned back towards the screen. While they are returning, the No.3 grid potential continues to increase negatively and additional force is exerted on the returning electrons. Some of these electrons do not strike the screen but circle around and return into the region between grids No.2 and No.3 where, if the period of transit is right, they acquire still more energy by the same means. Eventually they may have enough to overcome the retarding field due to the negative potential of the No.1 grid and will then flow through the No.1 grid circuit.

In order to keep this grid-current effect small, it is necessary to decrease the transit time by making the space between grids No.2 and No.3 as small as is consistent with good control action by the No.3 grid. The tube may also be designed to operate with a higher No.1 grid bias.

2. Inner-Grid-Injection Mixers (Class II)

The second class of mixers is that in which the oscillator voltage is applied to the No.1 grid and the signal voltage to an outer grid. Mixers in this class will be termed inner-grid-injection-type mixers. At present there are no class II mixers as such on the market, but the principle is employed in a number of converters in which such mixers are combined with oscillators.

The conversion transconductance (g_c) of a class II mixer may be determined from a curve of $g_{m_{3-p}}$ (or $g_{m_{4-p}}$ if G_4 is the signal grid) as a function of E_{c_1} using the formulas discussed for the class I mixer but substituting $g_{m_{3-p}}$ for $g_{m_{1-p}}$ and considering E_{c_1} as the oscillator voltage instead of E_{c_3} (see Fig. 6). The maximum g_c is approximately

$$\frac{g_{m_{3-p}}{}^{(max)}}{4} \quad \left[\text{or} \quad \frac{g_{m_{4-p}}{}^{(max)}}{4} \right]$$

when optimum oscillator conditions are used.

In this type of mixer, however, one is confronted with the problem of the design of a space-charge-grid tube to provide maximum g_m; and the conditions for optimum design of this type of tube have not been analytically determined to the degree that they have for the former types of tubes.

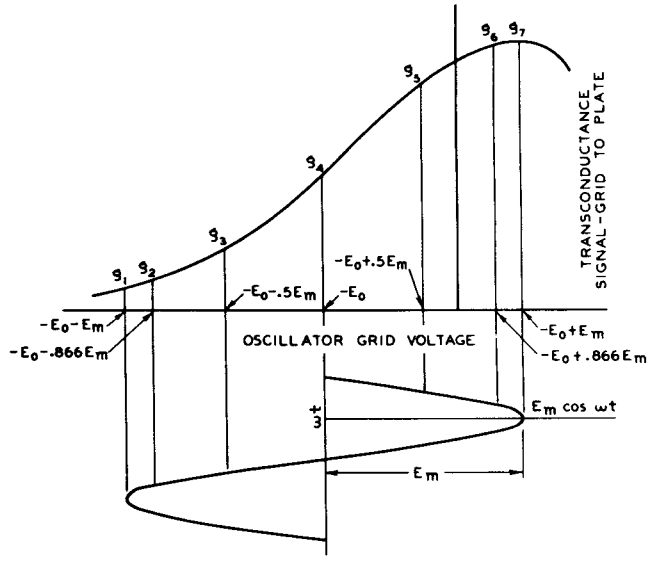


Fig. 6 - Method of determining numerical values for use in equation (5).

Salzberg and Haeff[4] and others have studied this problem assuming no transverse fields. Below,[5] Rothe and Kleen,[6] and Recknagel[7] in various German papers have considered the problem as an electron-optical one. By neglecting space charge and taking into consideration only the deflection of the electrons by the field of the various grid wires, they have derived expressions for the anode current in terms of the geometry of the tube and the potentials on the various electrodes. These papers are not particularly useful

---

[2] Loc. cit.

[3] M. J. O. Strutt, "Electron Transit Time Effects in Multigrid Valves," Wireless Engineer, Vol.15, pp. 315-321; June, 1938.

[4] B. Salzberg and A. V. Haeff, "Effects of Space-Charge in the Grid-Anode Region of Vacuum Tubes," RCA Review, Vol. 2, No.3, pp. 336-374; January, 1938.

[5] Frank Below, "Zur Theorie der Raumladegitterröhren (On the Theory of Space-Charge Grid Tubes)," Zeitschrift für Fernmeldtechnik, Vol. 9, pp. 113-118, 136-143; 1928.

[6] H. Rothe and W. Kleen, "Stromverteilurgssteuerung (Controlling of Current Distribution)," Telefunkenröhre, pp. 158-174; November, 1936.

[7] A. Recknagel, "Zur Intensitatssteurung von Elektronenstromen (On Intensity Control of Electron Streams) Hochf: tech.u. Elek:akus., Vol.51, No.2, pp. 66-72; 1938.

for the actual design of a specific converter, but give an interesting theoretical background.

If the screen grid (No.2) and the signal grid (No.3) were made of many turns of exceedingly fine wire so that the fields were very uniform and deflections were reduced, a high maximum g_m would be realized. Unfortunately, we cannot attain these conditions because mechanical strength determines the wire size, and the signal grid usually has a variable pitch (including a large gap) to reduce r-f distortion and cross-modulation.

The design of the space-charge-grid mixers is therefore, largely accomplished by trial and error. The dimensions of the grids are varied until an acceptable value of g_c is realized. During the course of the design of such a mixer, it has been found useful to study the E_{c1}-g_{m3-p} curves of one structure before designing the next trial structure. The E_{c1}-I_b curves with various E_{c3}'s, which can readily be obtained on the curve tracer, are also useful. In certain cases, the use of a rubber-membrane field model for tracing electron trajectories and of a probe for determining current distribution have also been used.[8]

An important feature of the class II mixer is that the curve of signal-grid bias (E_{c_s}) vs g_c may have a maximum at some negative value of E_{c3}. This maximum may be moved in a positive direction by increasing the turns per inch on the high-mu section of the signal grid. It is obviously desirable to have this peak occur at a bias less negative than the minimum value of $-E_{c3}$. With the present trend toward zero bias on r-f tubes and converters, the peak should be at or very close to $E_{c3} = 0$. Consideration should also be given to the range over which this bias for peak g_c may vary in production tubes.

3. Present Designs of Mixers and Converters

There are a number of different general types of mixers and converters on the market, each of which has its advantages and disadvantages. The general principles of each type will be described at this point and later comparisons will be made of some of the merits and defects of the various types.

In Fig. 4 is illustrated the schematic diagram of the class I mixer, which has been previously discussed. Fig. 7 shows a triode-hexode-type mixer, such as the 6J8-G. It consists of a class I mixer and a triode oscillator, both having a common cathode. The triode grid is tied internally to the No.3 grid of the mixer section.

The other two types of converters to be discussed each consist of a class II mixer plus an oscillator section. In each case a single grid structure is used for the oscillator grid and for the No.1 grid of the mixer, but in one construction separate electron streams are used for oscillator and mixer.

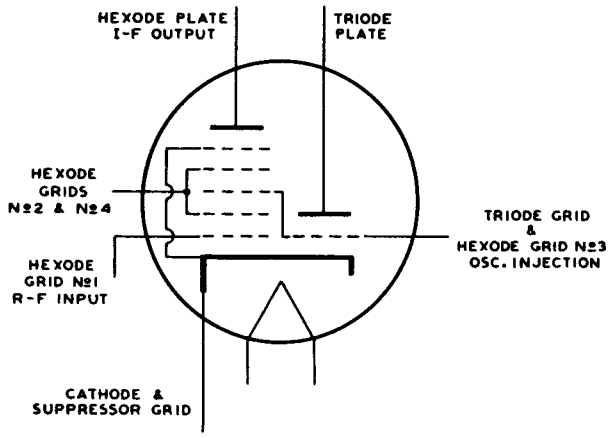


Fig. 7 - Triode-hexode converter (6J8-G).

Fig. 8 shows a pentagrid converter, such as the 6A8. In this type, a pair of rods serve as the oscillator anode. These rods are commonly called the No.2 grid and are located as shown in Fig. 8 in the plane of the grid side-rods between the No.1 grid and the first screen grid (G_3). Since they are shielded from the cathode by the

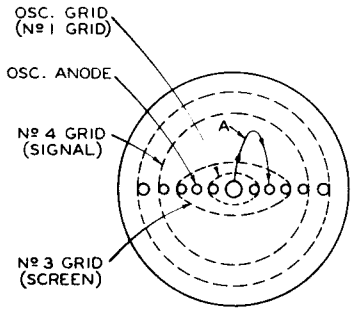


Fig. 8 - Cross-section of pentagrid converter (6A8).

No.1 grid side-rods, they derive their current mainly from electrons which turn back in front of the signal grid (G_4) and return through G_3. In order that there will be adequate oscillator-anode current, the rods must be at a considerably higher voltage than the screen voltage. The current and g_m of the oscillator are also dependent on the bias applied to the signal grid (G_4).

In Fig. 9 is shown a cross-section of the 6K8.[9] In this type, the flat cathode is surrounded by a flat grid which acts as the oscillator grid on the lower side of the figure and as the first grid of the mixer on the upper side. On the mixer side, a single grid serves as both inner and outer

[8] J. L. H. Jonker and A. J. W. M. van Overbeek, "A New Converter Valve," Wireless Engineer, Vol. 15, pp. 423-431; August, 1938.

[9] E. W. Herold, W. A. Harris, and T. J. Henry, "A New Converter Tube for All-Wave Receivers," RCA Review, Vol. 3, No.1, pp. 67-77; July, 1938.

screen grid. The screen grid and signal grid are surrounded by a shield which separates the oscillator and mixer sections and also functions as a suppressor grid between G_4 and P_{HX}.

A converter type EK3 has lately been announced by the N. V. Philips' Glowlampworks in Holland which also makes use of separate electron streams for the oscillator and mixer.[8] In addition, the

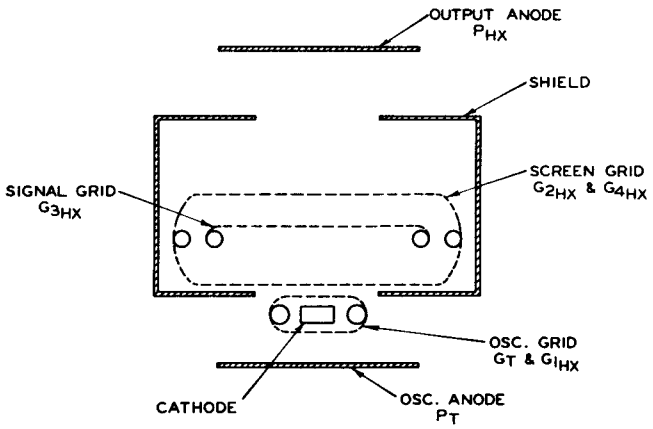


Fig. 9 - Cross-section of triode-hexode converter type 6K8.

construction is such that very few electrons can return from in front of the signal grid to the region of the No.1 grid.

4. Important Features of Converters

Desirable features in mixers or converters are:

1) High conversion transconductance (g_c)
2) High plate resistance of the mixer (r_p)
3) Low cost
4) Low cathode current
5) Low coupling effects between oscillator and input circuit
6) Low frequency drift
   a. with automatic volume control (avc)
   b. with supply voltage
   c. due to oscillator interlocking
7) Low input conductance
8) High oscillator transconductance.

The high g_c and r_p need no further comment. Cost is, of course, a factor in all tube design but in the case of the converters, it has had a great deal to do with the choice of design.

5. Coupling Effects

With an inner-grid injection type of mixer, the negative space charge in the vicinity of the control (signal) grid increases[10,11] as the oscillator grid becomes more positive. This increase in negative space charge tends to drive electrons out through the control-grid circuit. The capacitance between oscillator grid and control grid has the opposite effect, that is, a positive increment of voltage on the oscillator grid will tend to induce a negative charge on the control grid and draw electrons in through the control-grid circuit. If, as is customary, the oscillator frequency is higher than the signal frequency, the input circuit offers capacitive reactance to the oscillator frequency and the variation in space charge will produce a voltage out of phase with the oscillator voltage. This electron coupling acts as a negative capacitance from oscillator grid to signal grid. The effect becomes more important at high frequencies because, if we keep the intermediate frequency constant, the signal and oscillator frequencies have a lower ratio and, therefore, the input-circuit impedance will become appreciable at oscillator frequency so that higher voltages of oscillator frequency will be developed across the signal circuit. With higher oscillator frequency than signal frequency, this induced voltage of oscillator frequency on the signal grid tends to reduce g_c. With lower oscillator frequency than signal frequency, it tends to increase g_c.

The negative capacitance due to space-charge coupling can generally be neutralized for any particular frequency by connecting an external capacitance from oscillator grid to signal grid,[8,11] but at very high frequencies where transit time becomes a factor, the use of a resistance in series with the capacitance[3] may be required. This method of neutralization is not entirely satisfactory. It has been observed that the use of a neutralizing capacitance increases frequency drift in the 6K8. The Philips' Glowlampworks use a built-in neutralizing capacitance and resistance in the type EK3. They state that frequency drift is increased by this neutralization but is still low enough to be satisfactory.

Outer-grid-injection tubes also show some space-charge coupling due to electrons returning from the oscillator grid to the signal-grid region but the effect is slight. In the inner-grid injection type, space-charge coupling is inherently present because the space charge near the signal grid must vary at oscillator frequency.

---

[8] Loc. cit.

[3] Loc. cit.

[10] W. A. Harris, "The Application of Superheterodyne Frequency Conversion Systems to Multirange Receivers," Proc. I.R.E., Vol. 23, pp. 279-295; 1933.

[11] M. J. O. Strutt, "Performance of Some Types of Frequency Changers in All Wave Receivers," Wireless Engineer, Vol. 14, pp. 184-192; 1937.

6. Oscillator-Frequency Drift[8,9,11]

The oscillator frequency in most converters varies with the bias on the signal grid. This variation is of sufficient magnitude in the 6A8 and similar converters to cause considerable difficulty in the high-frequency bands where a percentage change in oscillator frequency will cause a much larger percentage change in the beat frequency between a given r-f signal and the oscillator frequency.

If a fading r-f signal is tuned in, the variation of the avc bias may be sufficient to cause the station to swing in and out of the band. A more important effect is the difficulty encountered in tuning.[9] The magnitude of the output will differ when one tunes in from the higher- or the lower-frequency end of the dial to a station with a large signal. Considerable readjustment may be necessary to obtain optimum tuning.

Change of frequency with change in signal-grid bias may be due to a change in the magnitude or the phase angle of the oscillator g_m, to a change in the internal capacitances of the tube, or to a combination of these effects. An out-of-phase component in the transconductance will act as additional capacitance or inductance in the oscillator circuit, and a change in phase angle, or in magnitude of g_m, will change the value of this effective capacitance or inductance.

In the pentagrid converter, the oscillator anode derives most of its electron supply from the electrons returning through the screen from the G_3-G_4 region. Changes in signal-grid bias affect both the number and the paths of the electrons returning to the oscillator anode and have considerable effect on the magnitude of g_m and also on the phase angle of g_m (due to transit time) at high frequencies.[3,8]

To reduce this oscillator drift with frequency, a designer can put the oscillator section on a separate electron stream from that of the mixer section. This arrangement was employed in the 6K8 with the result that the frequency drift was very much reduced.[9] Other factors, such as the change in input capacitance with variations of signal-grid bias, cause some residual frequency drift with avc in both inner- and outer-grid injection types of converters, but with separate shielded electron streams the effect is small.

Frequency drift may also occur when the supply voltage to the converter elements is varied. This may happen when the line voltage fluctuates; or in a set where the regulation of the rectifier filter circuit is not perfect, fluctuation in the voltages on the converter elements will occur when the average current drawn by the other tubes in the set changes due to signal. If the frequency drift is large enough, the signal strength and avc bias on the other tubes in the set vary, and a type of audio-frequency motor-boating known as "flutter" may occur at high input signals if the time constants of the circuits are favorable. This effect can be reduced by adequate by-passing of the screen and oscillator-anode supplies, but such by-passing, of course, causes extra expense. Measurements usually show less fluttering on the outer-grid injection type of converter but the reason is not well understood. It may be due in part to factors other than those discussed above.

7. Interlocking

At high signal inputs to the 6A8, an effect known as "interlocking" occurs. Interlocking is due to the signal-frequency pulses of electrons from in front of the signal grid which return to the oscillator anode and tend to pull the oscillator frequency into step with the signal frequency.[10] This effect is small in types where the oscillator is on a separate electron stream.

8. Input Conductance

High input conductance in a converter loads the tuned input circuit, broadens the tuning, and decreases the gain. In the case of the outer-grid-injection mixers, the input conductance increases with frequency in a manner similar to the increase of conductance in the conventional triode or r-f pentode. Moreover, it has been observed that under the oscillating condition the conductance is higher than might be expected from conventional tube considerations. Measurements of input conductance under static conditions show that it increases with increasingly negative oscillator-grid bias.

The increase in conductance is apparently due to the electrons which return from the oscillator-grid region to the No.1 grid-cathode region. When their return is prevented, this variation of input conductance with oscillator-grid bias may be greatly reduced. One method of reducing the number of returning electrons is by the use of beam principles similar to those used in the Philips EK3.[8] Electron current to the No.1 grid will also be reduced.

The input conductance of inner-grid-injection converters is generally low and may even be negative.[11]

9. Signal-Grid Current

Signal-grid current in outer-grid-injection mixers has been previously discussed. This phe-

---

[3] Loc. cit

[8] Loc. cit

[9] Loc. cit.

[10] Loc. cit.

[11] Loc. cit.

nomenon has also been observed in inner-grid-injection converters to a smaller degree, although the reason for its occurrence is not clear.

10. Oscillator Transconductance

The magnitude of the g_m of the oscillator is important because at high frequencies the circuit losses become greater and a higher g_m oscillator is desirable in order to get adequate excitation.[9]

In the pentagrid converter, g_m can be increased by using a larger oscillator anode, but a loss of g_c or a higher cathode current will result.

In the 6K8, the oscillator g_m is quite high because of the large amount of cathode area devoted to it. With a rectangular cathode, one half of the cathode is used for the oscillator and with the other half it is possible to get enough perveance to operate the mixer section with a fair amount of g_c.

In the outer-grid-injection converters, the oscillator and mixer sections in practice are mounted one above the other; and in order to get adequate perveance for the mixer section, the oscillator uses less than half of the cathode length. Since it is convenient for stem connections to mount the oscillator section at the bottom, the end-losses are high in this section and variations in g_m due to emission changes are to be expected. It is possible to improve the g_m by the use of a rectangular cathode or by the use of a larger cathode with more heater power. The rectangular cathode would increase the cost of an already expensive construction, and increased heater power is not very acceptable in tubes for use in both a-c and d-c sets.

11. Comparison of the Various Types of Converters

The pentagrid converter offers the advantages of being relatively low in cost, but its large frequency drift, lower oscillator g_m, and electron coupling are undesirable for high-frequency use. It can be used with an external oscillator tube to eliminate the first two objections but this increases set cost.

The 6K8 is reasonably cheap, has low frequency drift, and good oscillator g_m, but still has electron coupling. Because the 6K8 lacks symmetry, distortion of its parts during manufacture produces greater changes in characteristics than occur in types having symmetrical structures. The principle of separate electron streams for oscillator and mixer could be applied to symmetrical structures (as in the Philips EK3) but the cost would probably be high.

The outer-grid-injection mixer in which beam principles are used can be designed to have all the desirable features listed above except the one of low cost for the converter stage. Low cost is precluded since a separate oscillator tube is required. It does not appear practical to build a tube combining an outer-grid-injection mixer with an oscillator having as high a g_c and oscillator g_m as are obtainable with the inner-grid-injection converter using the same heater power.

The choice of a converter or mixer design, therefore, depends largely on properly balancing economy and performance.

12. Application Tests

All new designs of tubes should have application tests when there is any doubt about their performance in actual circuits; but in converters and mixers, so many factors are involved that a thorough study of their performance under actual operating conditions is extremely desirable, especially in the high-frequency bands. Other difficulties which have not been discussed above but which should be investigated in application tests are hum, hiss, off-resonance microphonics, blocking, and variation in performance over the expected range of oscillator excitation.

---

[9] Loc. cit.

Lecture 23

THE DESIGN AND CONSTRUCTION OF TRANSMITTING TUBES

E. E. Spitzer

Transmitting tubes are used as audio amplifiers, modulators, linear amplifiers, frequency multipliers, power amplifiers, plate-modulated amplifiers, and for a host of miscellaneous purposes. Audio amplifiers are generally operated class A, while modulators are operated class B. Linear amplifiers are also operated class B and are used to amplify modulated waves. Transmitting tubes in the other classes of service are generally operated class C.

These three classes of service will be considered briefly. Fig. 1 shows a schematic circuit of a class A amplifier and how the plate-current

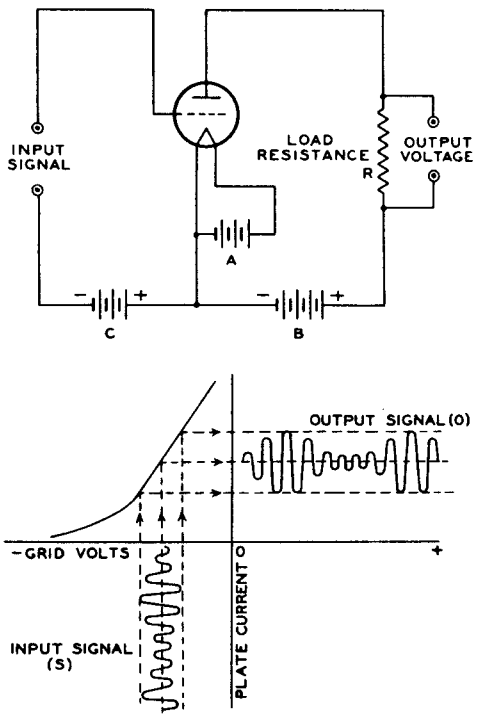


Fig. 1

wave is produced by a grid-voltage wave. Operation of the tube on the linear portion of its characteristic produces an output wave which is substantially undistorted. The maximum theoretical efficiency of a class A amplifier is 50 per cent. The grid is generally not driven positive; hence, the power gain is high.

Fig. 2 shows the circuit diagram and formation of the plate-current wave in a class B audio amplifier. The grids are biased approximately to cut-off so that only one tube conducts at a time. Use of two tubes in push-pull permits reproduction of the full wave-form in the output circuit. As a result, generation of serious harmonic distortion is avoided, as shown in Fig. 3. The maximum theoretical efficiency of such an amplifier is 78.5 per cent.

When a class B amplifier is used to amplify a narrow band of frequencies, such as a broadcast signal, a tuned load circuit can be used. Due to

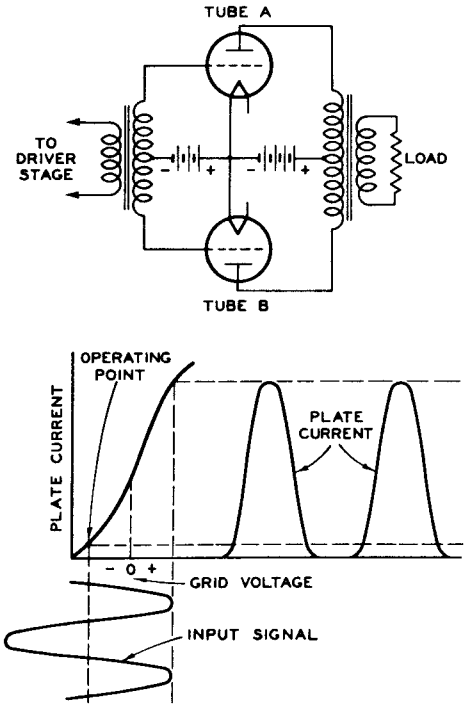


Fig. 2

the strong filtering action which is obtainable with such a circuit, it is not necessary to use two tubes in push-pull. The tuned circuit can be made to eliminate the harmonics.

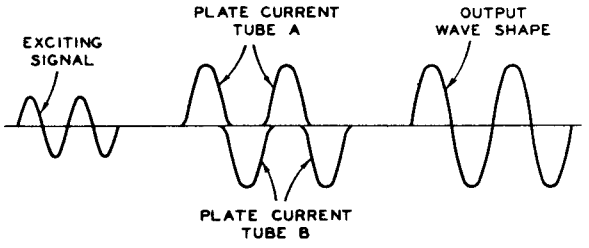


Fig. 3

Fig. 4 shows the characteristics of a class B linear amplifier of the type just described. It will be observed that the circulating current in the tuned circuit (I_T) varies linearly with the

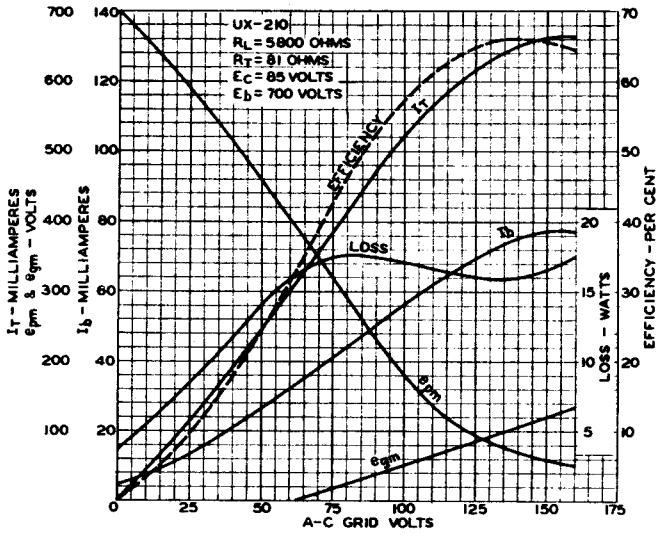


Fig. 4

amplitude of the a-c grid voltage up to a saturation point. Hence, if the a-c grid voltage is modulated over this range, the output current will also be modulated. If the a-c grid voltage is to be 100 per cent modulated, it is necessary to choose an operating point half-way up the linear portion of the I_T curve. As Fig. 4 shows, the efficiency also varies linearly with a-c grid voltage, so that at the carrier point the maximum theoretical efficiency is only 39.3 per cent. Practically, only about 33 per cent efficiency is attained. This is a convenient figure, as it means that the carrier output of a tube in class B linear operation is equal to half its plate-dissipation rating.

In general, the efficiency of an amplifier increases as the angle of each cycle, during which plate current flows, decreases. In a class C amplifier, high efficiency is desired, so the grid is biased well beyond cut-off. Current flows usually about 120 degrees of each cycle. The maximum theoretical efficiency is 100 per cent. The grid is driven far positive in order to get the maximum output. Fig. 5 shows the voltage and current wave-forms in such an amplifier.

When a sinusoidal, grid-exciting voltage is used, the plate-current wave-form is roughly a section of a sine wave. For such a condition, Fig. 6 shows the calculated ratio of peak plate current to d-c plate current as a function of the half-angle of plate-current flow. These high peaks must, of course, be considered in the design of the cathode for a tube to be operated in such service.

By means of Fig. 7, the plate efficiency of an amplifier can be calculated. As this figure shows, the efficiency depends on the product of a voltage and a current factor. Practically, efficiencies of the order of 70 per cent are easily attained.

Naturally, the plate-current wave-form of a class C amplifier is rich in harmonics, and these must be eliminated by a tuned circuit. Fig. 8 shows the characteristics of a simple tuned circuit. For an operating Q of 20, as is often used, the resistance of the circuit to the second harmonic is only 0.11 per cent of the resistance

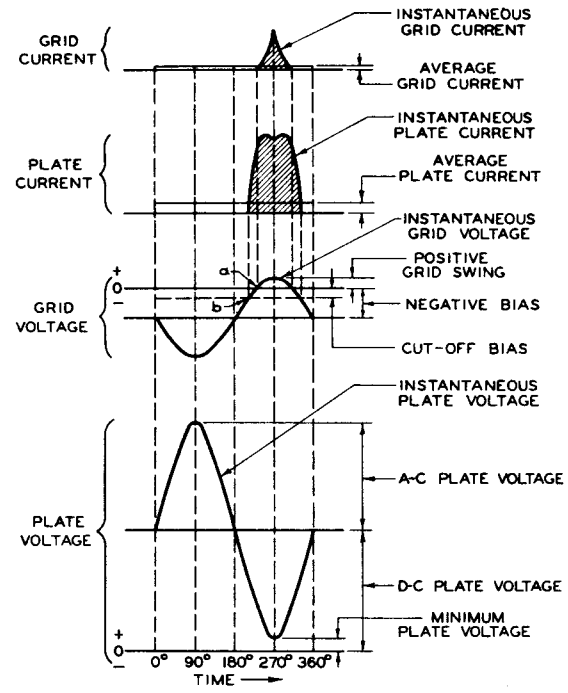


Fig. 5

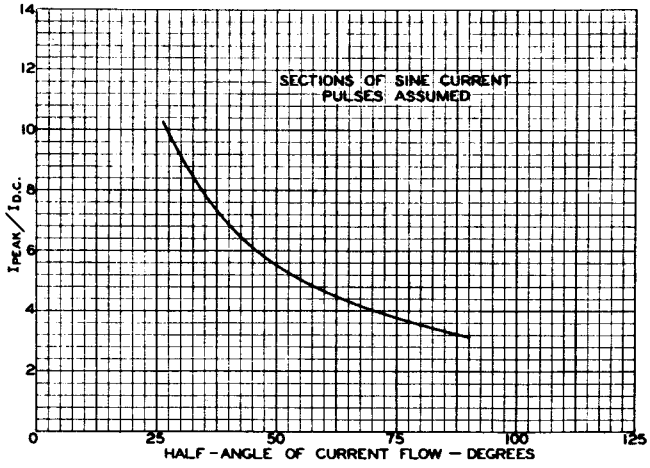


Fig. 6

presented to the fundamental component. The reactance to the second harmonic is 3.33 per cent. It is readily seen that even though there is a high percentage of harmonic in the plate-current wave-form, it is strongly reduced by this simple tuned circuit. In practice, a second stage of filtering is obtained by tuning the antenna circuit, so that the harmonic radiation can be reduced easily to any desired value.

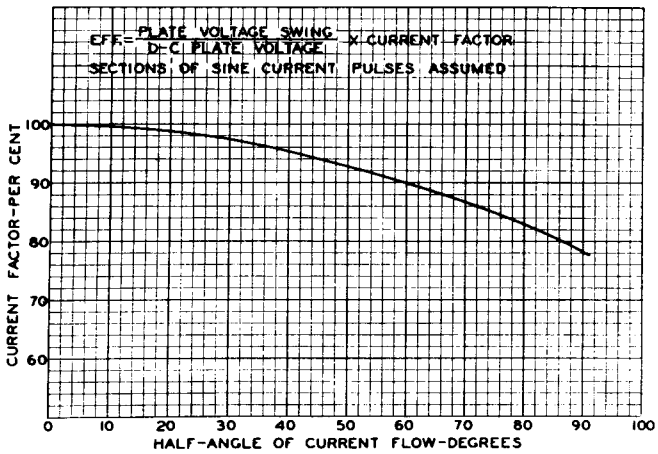
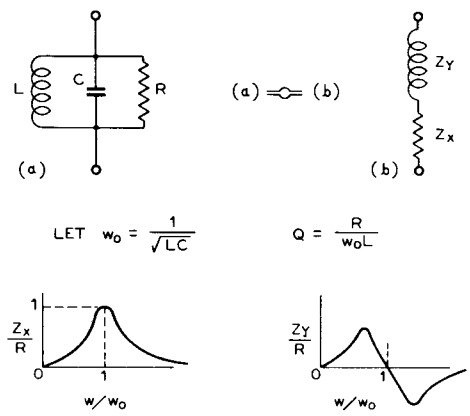


Fig. 7 - Class B and C amplifier efficiency.

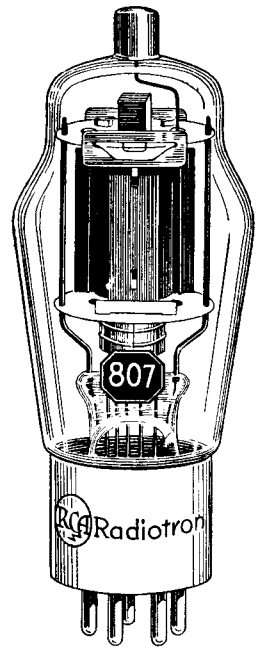
The maximum ratings of transmitting tubes for class C telegraph service are considered the basic ratings, as the ratings for other classes of service can be derived from them. The maximum ratings are given for d-c plate voltage, d-c plate current, plate input, plate dissipation, screen voltage and input, suppressor voltage and input, grid bias and grid current. The maximum plate ratings are determined by the original requirements and design. In any case, the plate input is not allowed to exceed four times the plate dissipation, as higher values would require more than 75 per cent efficiency, which would be very difficult for all users to obtain.



Q	FUNDAMENTAL		2nd HARMONIC		3rd HARMONIC	
	Z_x/R	Z_y/R	Z_x/R	Z_y/R	Z_x/R	Z_y/R
10	1	0	.0044	.0664	.0014	.0374
20	1	0	.0011	.0333	.00035	.0187
30	1	0	.00050	.0221	.00016	.0125

Fig. 8 - Characteristics of a simple tuned circuit.

Fig. 9 shows the maximum ratings of the RCA-807 for several classes of service. The fundamental class C telegraph ratings were established by means of life tests. In the case of class C telephone service, the ratings are given for carrier conditions, that is, with no plate modula-



RCA-807

Transmitting Beam Power Amplifier

As R-F Power Amplifier and Oscillator—Class C Telegraphy

Key-down conditions per tube without modulation

D-C Plate Voltage	600 *max.* Volts
D-C Screen Voltage (Grid No. 2)	300 *max.* Volts
D-C Grid Voltage (Grid No. 1)	−200 *max.* Volts
D-C Plate Current	100 *max.* Milliamperes
D-C Grid Current	5 *max.* Milliamperes
Plate Input	60 *max.* Watts
Screen Input	3.5 *max.* Watts
Plate Dissipation	25 *max.* Watts

As Plate-Modulated R-F Power Amplifier—Class C Telephony

Carrier conditions per tube for use with a max. modulation factor of 1.0

D-C Plate Voltage	475 *max.* Volts
D-C Screen Voltage (Grid No. 2)	300 *max.* Volts
D-C Grid Voltage (Grid No. 1)	−200 *max.* Volts
D-C Plate Current	83 *max.* Milliamperes
D-C Grid Current	5 *max.* Milliamperes
Plate Input	40 *max.* Watts
Screen Input	2.5 *max.* Watts
Plate Dissipation	16.5 *max.* Watts

As R-F Power Amplifier—Class B Telephony

Carrier conditions per tube for use with a max. modulation factor of 1.0

D-C Plate Voltage	600 *max.* Volts
D-C Screen Voltage (Grid No. 2)	300 *max.* Volts
D-C Plate Current	80 *max.* Milliamperes
Plate Input	37.5 *max.* Watts
Screen Input	2.5 *max.* Watts
Plate Dissipation	25 *max.* Watts

Fig. 9

tion applied. It is well known that the power in a 100 per cent modulated wave (single frequency modulation) is 50 per cent greater than in the carrier wave alone. The rms voltage is greater by a factor of $\sqrt{1.5}$, or roughly 1.25. These factors are used in reducing the plate ratings for class C telephone service. The plate-input and dissipation ratings are reduced by the factor 1/1.5 (= 66.7%) from the class C telegraph ratings. The plate voltage is reduced by the factor 1/1.25 (= 80%).

For class B telephone service it is necessary to reduce the plate-input and plate-current ratings because of the low efficiency under carrier conditions. An efficiency of 33 per cent is assumed for rating purposes, and this fixes the plate input at 150 per cent of the plate-dissipation rating. The plate- and screen-voltage ratings are not reduced, since the r-f and d-c voltage stresses on the tube are no more severe than in class C telegraph service. The plate-current rating is reduced because of the reduction of input. It should be noted at this point that the product of plate current and voltage is higher than the maximum plate input in some cases. The product is purposely higher to allow the use of full input at reduced plate voltage, if desired. Such operation is possible with tubes of high perveance. However, all ratings are independent, that is, none of them may be exceeded.

Where tubes are designed primarily for the amateur market, the plate-input watts per dollar of list price is made as high as possible for competitive reasons. Tubes designed for the last stage of broadcast transmitters must have a rating approved by the Federal Communications Commission. For the purpose of readily limiting the radiated carrier power of such transmitters, the FCC limits the input power to the last stage by assuming certain efficiencies. For high-level modulated stages, the efficiencies are shown in Table I.

Table I

Rated Carrier Power Watts	Plate Efficiency Per cent
100	50
250 - 1000	60
2500 - 50000	65

For example, the RCA-833, which was designed for a 500-watt, high-level, modulated carrier, has a carrier input of 500/0.6 = 835 watts.

Given the essential input ratings, the next decision is a choice of triode, tetrode, beam tetrode, or pentode construction. Triodes permit the highest watts per dollar but require neutralization. Table II gives some illustrative figures on comparative triodes and multigrid tubes.

Table II

WATTS PER DOLLAR

(Max.Class C Telegraph Input Watts/$ of List Price)

	Type	Watts	Price	W/$
Triodes	809	75	$ 2.50	30.0
	808	200	7.75	25.8
	805	315	13.50	23.3
	833	1250	85.00	14.7
	898	200000	1650.00	121.0
Tetrodes & Pentodes	807	60	3.50	17.1
	814	180	17.50	10.3
	803	350	34.50	10.1
	861	1200	295.00	4.0

Tetrodes avoid neutralization within limits. Above 45 megacycles very expensive construction is required to avoid feedback due to the inductance of the screen lead. Beam tetrodes and pentodes have the advantage of very low driving power. Whereas in triodes the grid must be driven highly positive to obtain the peak current pulses, in beam tetrodes and pentodes the necessary acceleration to the electrons is supplied by the screen, so that the grid need not be driven far positive and thus does not draw heavy grid current. Fig. 10 illustrates these statements very clearly.

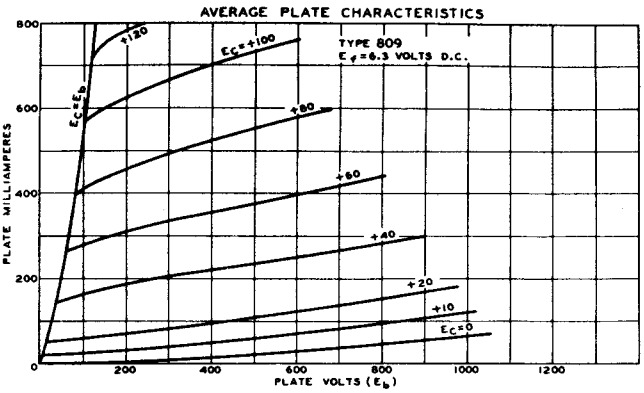
Pentodes have the additional advantage that they can be suppressor modulated. While suppressor modulation results in a big reduction of carrier output, the absence of heavy modulation transformers often offsets this loss in mobile installations. Fig. 11 shows a typical suppressor-modulation characteristic. It will be noted that the suppressor current is very low; hence, the modulation power required is very low. It is difficult to obtain more than about 80 per cent modulation due to the curvature of the characteristic.

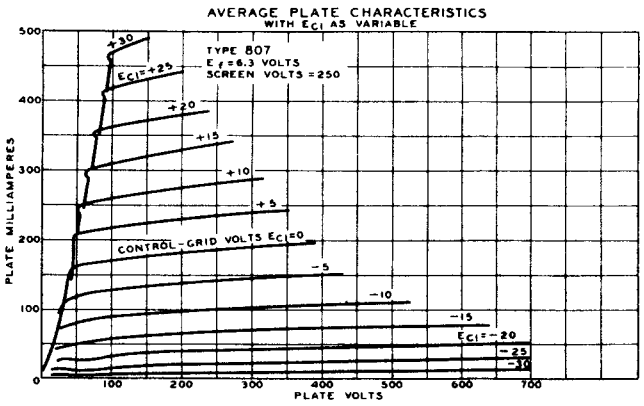
The next step in design is to determine the perveance needed. It is calculated from the desired plate current and plate efficiency. From Fig. 12 it can be seen that practically it is undesirable to swing the grid more positive than the plate because the ratio of plate to grid current drops rapidly beyond this point, and as a result high driving power is required. If the plate and grid voltage are made equal in the simplified expression for the characteristic of a triode, the mu-factor drops out. It will be noted that if plate current in milliamperes is read at $E_b = E_c = 100$ volts, the value obtained is the plate perveance of the tube in microamperes/volt$^{3/2}$.

Details for calculating the perveance of any structure will not be given here as they have

been adequately covered by Kusunose.[1] It is well to point out that perveance depends almost entirely on the effective cathode area and the grid-to-cathode spacing. The mu-factor of the first grid or the grid-to-plate or grid-to-screen spacing have practically no effect on the perveance.

Table III shows how the required perveance is calculated. The ratio of peak plate current to d-c plate current is determined from Fig. 6 for the assumed operating angle. The current factor





TYPICAL OPERATION - CLASS C TELEGRAPHY

	Triode RCA-809	Tetrode RCA-807	
D-C Plate Voltage	500	500	Volts
D-C Screen Voltage	-	250	Volts
D-C Grid Voltage	-50	-50	Volts
Peak R-F Grid Voltage	135	80	Volts
D-C Plate Current	100	95	Ma.
D-C Screen Current	-	9	Ma.
D-C Grid Current	20	2	Ma.
Driving Power	2.5	0.14	Watts
Power Output	35	30	Watts
Power Gain	14	214	-

Fig. 10

[1] Y. Kusunose, "Design of Triodes," Proc. I.R.E., Vol. 17, No.10; October, 1929.

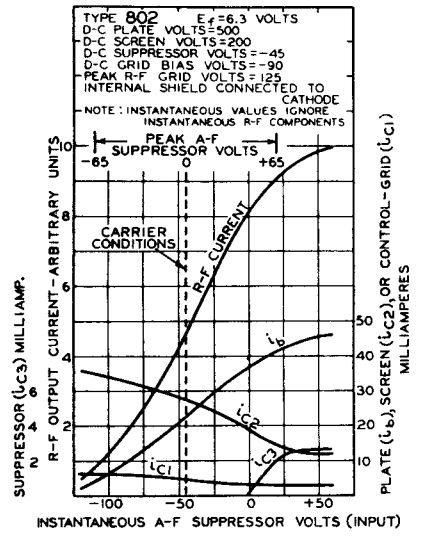


Fig. 11 - Typical suppressor-modulation characteristic.

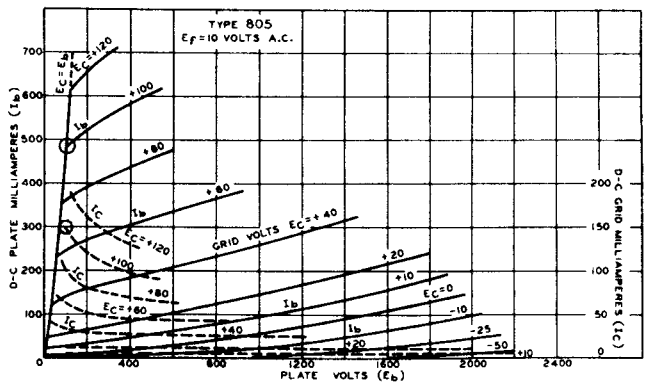


Fig. 12 - Plate and grid family for illustrating concept of plate perveance (G) in a triode for which

$$I_b = G\left(\frac{E_b + \mu E_c}{1 + \mu}\right)^{\frac{3}{2}}$$

$$\text{If } E_b = E_c \text{ , then } G = I_b/E_b^{\frac{3}{2}}.$$

is obtained from Fig. 7. From experience, we assume the peak grid current is 1/4 of the peak plate current. This assumption permits calculation of the minimum perveance, and the required grid and cathode structure can then be calculated using Kusunose's method.

The extension of the foregoing method to multigrid tubes follows along similar lines.

When tubes are designed for operation at high frequencies, the effect of transit time must be taken into account. An experience curve, shown in Fig. 14, which gives a plate efficiency factor as a function of the minimum grid transit angle, is used. This efficiency factor is applied to the normal efficiency determined from the opera-

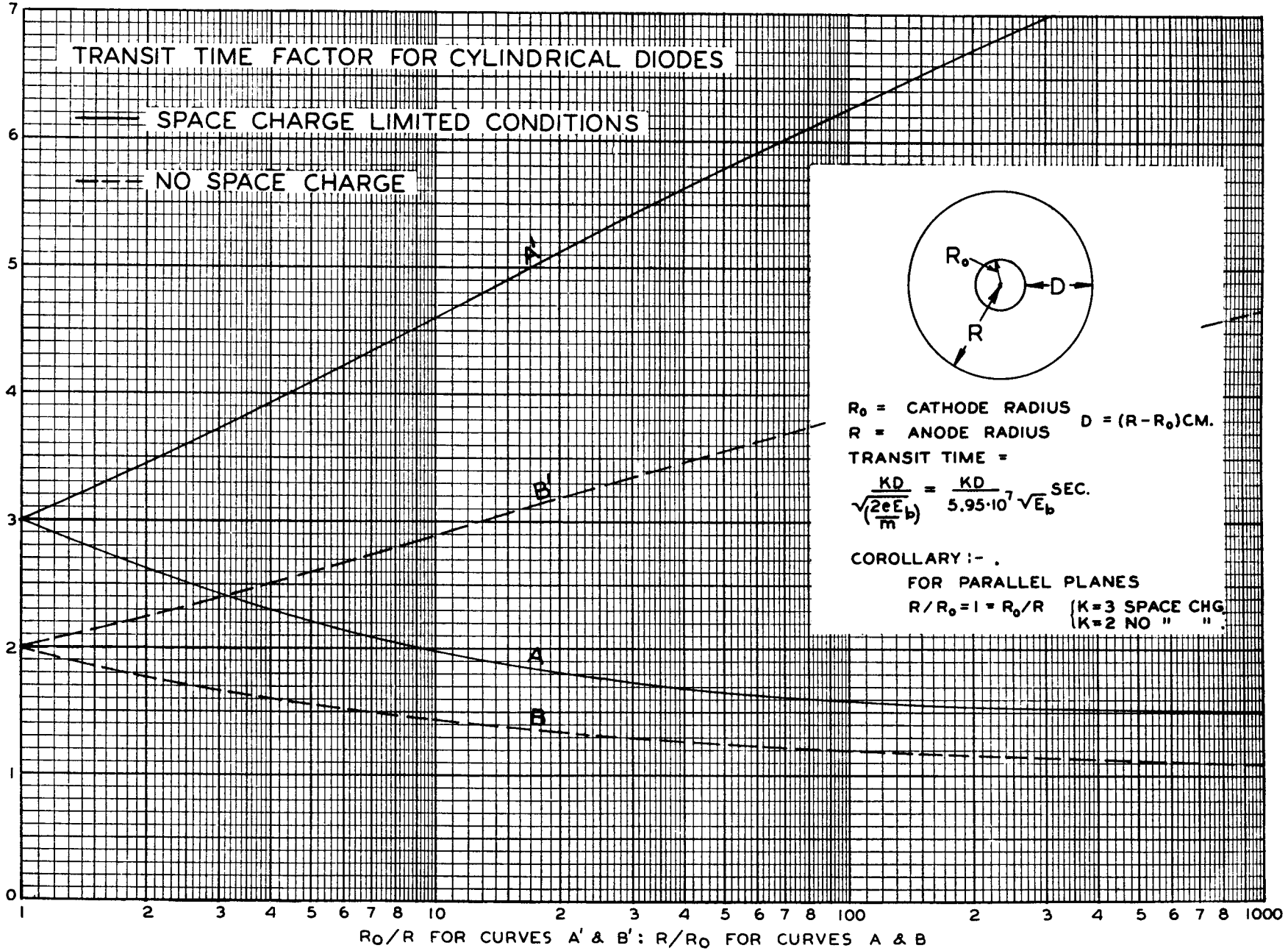

TRANSIT TIME FACTOR FOR CYLINDRICAL DIODES
SPACE CHARGE LIMITED CONDITIONS
NO SPACE CHARGE
A'
B'
A
B
K
R_o
D
R
R_o = CATHODE RADIUS
R = ANODE RADIUS
D = (R − R_o)CM.
TRANSIT TIME =
$\frac{KD}{\sqrt{(\frac{2eE_b}{m})}} = \frac{KD}{5.95 \cdot 10^7 \sqrt{E_b}}$ SEC.
COROLLARY :- .
FOR PARALLEL PLANES
$R/R_o = 1 = R_o/R$ {K=3 SPACE CHG. {K=2 NO " " .
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 10 2 3 4 5 6 7 8 100 2 3 4 5 6 7 8 1000
R_o/R FOR CURVES A' & B' : R/R_o FOR CURVES A & B

Fig. 13

ting angle and plate swing. The minimum transit angle is determined with the aid of curves in Fig. 13.

Table III

DETERMINATION OF MINIMUM PERVEANCE

Given: A triode to operate at
plate volts (E_b) = 750
plate ma. (I_b) = 100
plate efficiency = 75%

Assume: Half-angle of plate-current flow = 60°

Then: $I_{b\ peak}/I_{b\ d.c.} = 4.6$

$I_{b\ peak} = 4.6 \times 100 = 460$ ma.

Current Factor = 90%

Efficiency = 75%
$= E_{b\ swing}/E_{b\ d.c.} \times 90$

$E_{b\ swing} = 75/90 \times 750 = 625$ volts

$E_{b\ min} = 750 - 625 = 125$ volts

Peak Grid Current = $1/4 \times 460$
= 115 ma.

Peak Space Current = 460 + 115
= 575 ma.

Minimum Perveance = $575/(125)^{3/2}$
= 411 μamp./volt$^{3/2}$

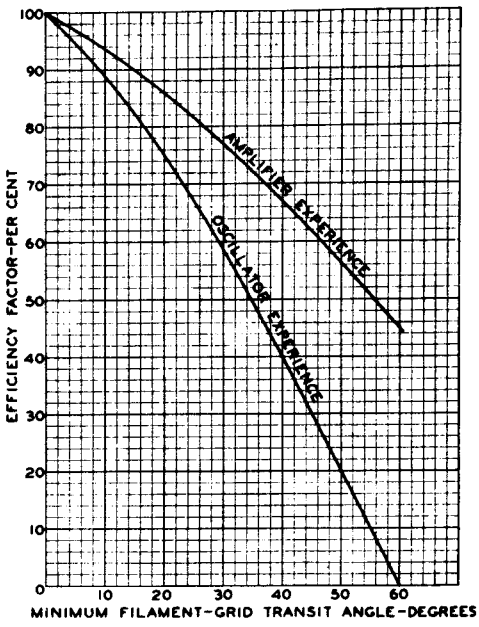


Fig. 14 - Experience curves.

The next step is to choose materials for the electrodes. The choice of any material is based on experience and cost. Oxide-coated cathodes and filaments are not used in amplifier tubes operated with plate voltage above 600 volts. The reason is largely because such tubes do not have the very considerable power and voltage overload capacities which American users are accustomed to have in tubes. In an oxide-coated-cathode type of tube operating above 600 volts, the danger of grid contamination and runaway is too great for a commercially acceptable tube. Thoriated-tungsten filaments are used up to 3500 volts, at which point loss of emission due to gas bombardment becomes an important factor. Beyond this voltage, pure tungsten filaments are used. The relative design emission efficiencies are: oxide cathodes, 130 milliamperes per watt; thoriated tungsten, 50 milliamperes per watt; tungsten, 7.5 milliamperes per watt.

Information on design of oxide-coated cathodes has been given in previous lectures. Thoriated-tungsten cathodes are readily designed with the aid of Fig. 15. Tungsten filaments are designed from the well-known Jones and Langmuir Tables.

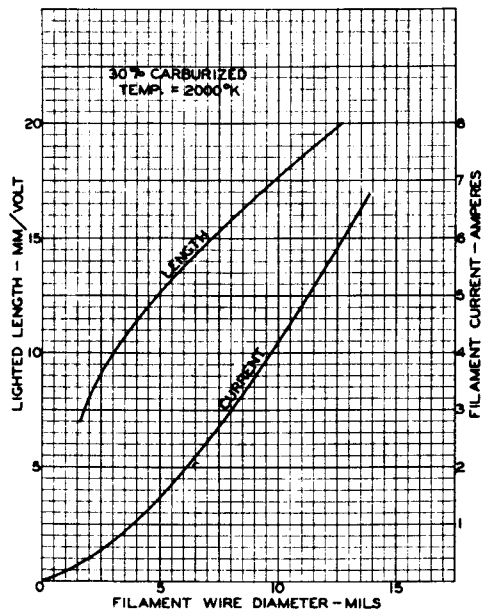


Fig. 15 - Design curves of thoriated-tungsten filaments.

Anode materials, arranged in order of cost, are: carbonized nickel, carbon, molybdenum, tantalum, external anode copper. The chief advantage of nickel is very low cost. Carbon has the advantages of uniform heating and absence of warping. In addition, expensive dies are not needed. Tantalum offers the advantage of getter action. The power which can safely be dissipated in these materials depends largely on the pro-

cessing. It is best determined from existing successful tube designs.

The glasses used are No.8 for low-power tubes; No.702-P and No.705-BA for high-power tubes. No.12 and No.814-KW are used for stems with No.8 bulbs. No.702-P is used with tungsten seals and No.705-BA with Fernico seals. The maximum dissipation per square inch of bulb surface has been held to 1.7 watts for No.8 and to 4 watts for No.702-P and No.705-BA.

Insulators most widely used are mica, lava, and 21R material which is a ceramic composed largely of alumina. The latter has been found to have the best mechanical and electrical properties. Some power-factor data are given in Table IV.

Insulators must be kept free of conducting deposits if power loss at high frequencies is to be avoided. Such deposits often will not show up when tests for d-c leakage are made, but measurement at high frequency on a Q-meter will quickly reveal the deposits. In the case of very-low-power tubes, such loss merely reduces the output. In the case of medium- or high-power tubes, insulator loss may result in cumulative heating of the insulator until it fuses and gases the tube.

Table IV

POWER FACTOR OF INSULATORS AT 60 MEGACYCLES

Material	% Power Factor		
	25°C	300°C	400°C
Fused Quartz	0.03	0.04	0.06
21R (1450)	0.06	0.08	0.08
Lava Grade I	0.16	0.33	0.60
Glass No.707DG	0.12	0.23	1.03
Glass No.702P	0.28	1.21	2.86
Isolantite	0.38	3.38	6.98
Glass No.8	1.06	12.50	-

Lecture 24

THE DESIGN AND CONSTRUCTION OF CATHODE-RAY TUBES

W. H. Painter

Electron guns, whether they be designed for use in oscillograph tubes or television tubes of the high-vacuum type, all depend on the fundamental theories of electron optics for an explanation of their mode of operation. As previous lecturers in this course have demonstrated, it is possible to calculate the approximate trajectory of an electron through various fields of force by using the principles of electron optics. Electron optics is a relatively new branch of science. While knowledge of the subject is expanding rapidly there are as yet no simple design formulas available to the development engineer, and much of the work remains in the realm of cut-and-try. An understanding of equipotential line plots of various types of structures will give an excellent idea of where to attack a development problem, but even if we were able to calculate cathode-ray tube performance completely on paper, the manufacturing technique has not sufficiently advanced to permit us to duplicate our paper designs.

While all of the electron guns which we make employ the same basic principles, there are differences in achieving the desired ends which justify dividing them into four classes for discussion. I have classified them arbitrarily as oscilloscope, Kinescope, projection Kinescope, and Iconoscope guns. These divisions are not rigid; indeed, the tendency is toward greater standardization.

Oscillograph tubes do not differ greatly from Kinescopes except that, in general, the quality requirements are less severe. Fig. 1 shows a relatively simple type of gun that is used in the

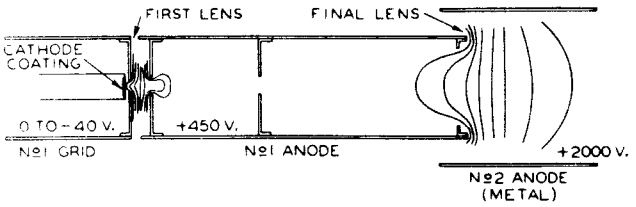


Fig. 1 - Electron gun used in type 905.

905, 907, and 909, a series of 5-inch oscillograph tubes. The source of electrons is an indirectly heated cathode consisting of a sleeve with a cap on one end. Cathode coating is sprayed into a depression in the cap and the excess is scraped off, leaving a smooth surface about 0.090" in diameter. The excess is generally removed with a razor blade; the cap may then be pressed against a piece of glass or similar smooth surface to remove bumpiness.

The No.1 grid is a metallic cylinder enclosing an apertured disk, the aperture being 0.040" or 0.048" in diameter. The cathode must be mounted within this grid, with the active surface behind the aperture and as close to it as possible. Application of a negative voltage to the grid exercises the usual control over the emitted electrons.

Co-axial with the grid is the No.1 anode, another cylinder enclosing one, two, or more apertured disks. The No.1 anode operates at a positive potential; its field supplies the initial acceleration to the emitted electrons. The difference in potential between grid and No.1 anode establishes an electron lens which converges the electrons toward the axis. When they reach the axis, their paths cross and they continue on down the No.1 anode in an expanding envelope. Those which stray too far from the axis are intercepted by the masking apertures and the rest continue on into the final focusing field. The No.2 anode of the 905 is a metallic cylinder, while on some other types it is formed by a conductive coating on the interior wall of the bulb neck. Regardless of the physical nature of the No.2 anode, the potential difference between the No.1 anode and the No.2 anode sets up another electron lens which again converges the electrons and, when the voltage ratio is correct, focuses them into a spot on the fluorescent screen.

In those cases where the No.2 anode is formed by a metallic coating on the bulb wall, it performs another function, i.e., the collection of secondary emission from the screen. Since the screen is a good insulator, the charge built up by the arrival of beam electrons can only be removed by secondary emission. Should the bulb be left uncoated, these secondaries would be collected by the deflecting plates and cause some distortion in the trace. Use of a conductive bulb coating removes the loading from the deflecting circuits. Until recently it has been thought that, by properly selecting a very dull-surfaced material, reflection from the bulb wall could be reduced. Some recent experiments indicate that most coatings cause considerably more reflection than the clear glass, provided no reflecting surface surrounds the bulb. However, since it is sometimes inconvenient to mount the tube in a blackened shield, the coating frequently performs useful service in this respect.

In a tube of the 905 type, the deflecting plates are rigidly mounted on heavy leads that project through the side walls of the bulb as illustrated in Fig. 2. The manufacture of this type of tube involves some expensive splicing of the bulb, and from a cost standpoint it is much better to mount the plates directly on the stem. To obtain the necessary number of leads to allow

direct connection to each of the plates involves more stem leads than we have had available until recently. It also involves some unconventional basing procedure. In the 906, which is a 3-inch tube, these problems were avoided by connecting one deflecting plate of each pair to the No.2 anode within the tube. This, of course, limits

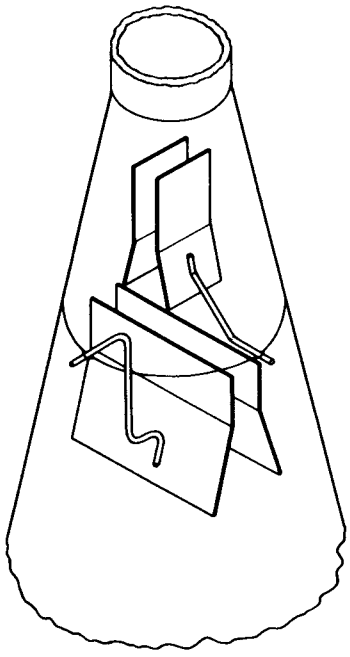


Fig. 2 - Deflecting-plate structure of type 905.

the deflecting circuit to the unbalanced type and introduces some distortion into the pattern. A study of equipotential line plots of the deflecting plate structure showed that by bending the ends of one of the plates to form a sort of box-like structure (see Fig. 3), less distortion resulted than with the use of flat parallel plates, in the case of the unbalanced circuit connection. However, with a balanced circuit, bending the end of one plate does not help materially and there still exists a considerable distortion of the spot in the corners of the pattern.

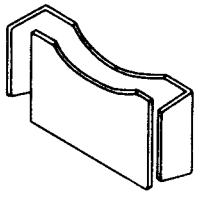


Fig. 3 - Upper plates of type 906.

As I have already mentioned, Kinescope guns must in general be capable of better focus than those used in oscillograph tubes. Further, the focus must be more uniform over the entire scanned pattern. Also, since a high-frequency signal is to be applied to the control grid, the modulation characteristic assumes a position of more importance.

In the RCA-1800, a 9-inch Kinescope, and in a developmental 12-inch Kinescope which has been used in the broadcast field tests, we have obtained the best results by the addition of a No.2 grid, or accelerating grid (see Fig. 4), to the type of gun already described. The No.2 grid is

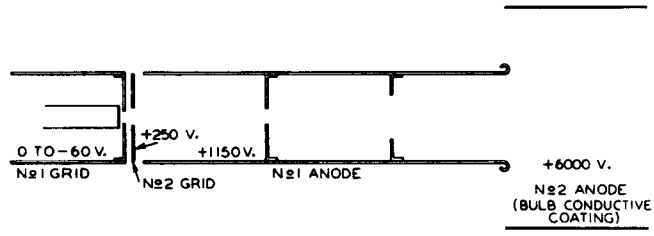


Fig. 4 - Electron gun used in type 1800.

interposed between the control grid (No.1 grid) and No.1 anode and commonly operates at about +250 volts. The No.2 grid cooperates with the control grid to form the first electron lens in this case. It also acts to shield this initial accelerating field from the effects of voltage variation on the No.1 anode. Thus, when the No.1 anode potential is varied to bring the spot to focus, the current drawn from the cathode, and thus the brightness of the picture, is affected much less than it would be were the No.2 grid not there. This is a matter of considerable convenience in operation of a television receiver.

Unfortunately, we have, as a rule, been unable to secure very good focus with this exact type of gun when the No.2 anode potential is much below 3000 volts, and for low-voltage Kinescopes it has been common practice to use modifications of the 905-type gun. Recently, however, a new factor arose in the design of receivers which forced us to modify the gun design. When the No.1 anode draws current, the voltage supply must have fairly good regulation to keep the beam in focus. The No.2 anode supply must also have good regulation, so that this type of gun requires two well-regulated supplies. In the interests of lower receiver cost, it seemed desirable that the No.1 anode draw no current. In the older type of gun this was obviously impossible if the masking apertures were to continue to perform their allotted function. Study of this problem has led to the use of inverse focusing, that is, the beam is first accelerated, then decelerated, and finally accelerated again to its final velocity. The No.2 grid has been lengthened and the masking apertures transferred to it from the No.1 anode (see Fig. 5). The current drawn by the No.2 grid is relatively unimportant because the No.2 grid is connected internally to the No.2 anode. This connection, of course, requires that the No.2 an-

ode supply have somewhat better regulation because of the higher drain, but the No.1 anode draws no current and only a single regulated power supply is needed. This type of gun is capable of giving well-focused spots and is now being used in the 5-inch Kinescopes.

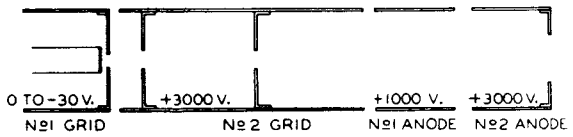


Fig. 5 - Electron gun with zero anode No.1 current.

When we come to the projection tube, we find that a different set of requirements faces the designer. A Kinescope for direct viewing operates at a No.2 anode potential of from 2500 to 6000 volts. A picture of reasonable brilliance can be obtained with peak beam currents of 200 microamperes or even less. The projection type requires a very small image of high brightness, suitable for transmission through a lens system onto a screen. Since optical systems of a suitable type are only from three to five per cent efficient, a major portion of the picture brightness is thrown away. It is, therefore, necessary to use much higher voltages, much higher beam currents and much finer focus in the projection tube than in any of the directly viewed types.

Modifications of conventional types of guns were able to produce pictures which were not sufficiently bright and were not too rich in detail. At 10000 volts, peak beam currents of 400 or 500 microamperes have been obtained with satisfactory focus. However, currents of this magnitude placed such a tremendous load on the cathode — since the No.1 anode current was always several times the No.2 anode current — that the tube life was extremely limited. A new design is currently being used which, while it has not transformed the projection tube into a success overnight, has given great improvements in the picture and shows promise for even better results in the future. In this type, which is shown in Fig. 6, the electrons are accelerated very rapidly over a very short distance, and the crossover occurs at a high potential. In this manner a large current can be converged into a very small crossover. A masking aperture is placed at this crossover, and this aperture serves as the object for the final lens system. Initially, this final lens was formed entirely by a magnetic coil placed outside of the tube. At present we are using this magnetic lens in conjunction with an electrostatic lens of the usual type. With this type of gun, peak beam currents as high as 1.5 milliamperes have been obtained with good focus at a No.2 anode potential of 15000 volts.

There are other problems besides that of the electron gun which make the problem of television projection very difficult. Assuming that we can make a gun which will provide sufficient current in a small enough spot, the efficiency and life of the fluorescent screen must be improved. Most of the fluorescent materials we have tried begin to saturate both with voltage and current at levels below those needed. Measurements of secondary emission from screens indicate that most materials "stick" at fairly low potentials, that is, with increasing voltage the ratio of secondary-to-primary electrons falls so low that the potential of the screen lags far behind the potential of the No.2 anode. So far, only green willemite of the type used in oscilloscopes and "magic eyes" has shown the ability to reach as high as 20000 volts. Most of the materials saturate with current, so that as the beam current is increased, the efficiency falls as much as 50 per cent in some cases. Not much data have been acquired on life, but previous experience indicates that 200 or 300 hours is about all we can expect at present without serious drop in light output.

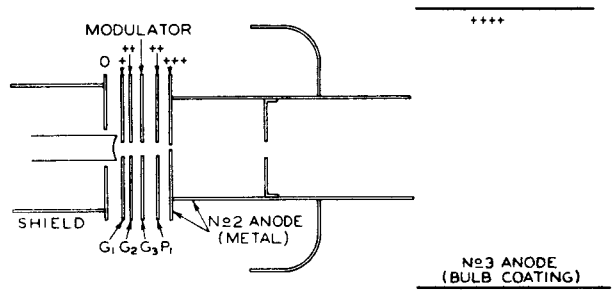


Fig. 6 - Projection-type electron gun.

A new type of trouble met in the projection tube is the actual discoloration of the glass due to electron bombardment. At 20000 volts and 1 milliampere, the end of a projection tube will acquire a dirty brown discoloration in 20 minutes. That this discoloration is not burning of the screen is shown by the fact that when the screen is washed out of the bulb, the stain remains in the glass. While this effect is possibly some form of electrolysis, no solution of the difficulty has yet been found.

The requirements of the Iconoscope gun are quite unique. Whereas in other types we are constantly striving for more focused current, we find that the Iconoscope tends to work better as the current is reduced. On the average, the beam current in an Iconoscope is about a quarter of one microampere. Some of the tubes operate at only 0.1 microampere. This low current operation, one might think, should make the design problem quite simple, but there are complicating factors. First of all, the mosaic is mounted at an angle (see Fig. 7) to the axis of the electron gun so that the spot will be elliptical even if the cross-section of the beam is circular. The beam must, therefore, be smaller than would be the

case for perpendicular intersection. Secondly, the tube operates at a maximum potential of 1000 volts, and low-voltage beams are more susceptible to distortion than high-voltage ones. Thirdly, the resolution must actually be quite superior to that of a Kinescope. Offhand, it would seem that if the beam were capable of resolving 500 lines, it could transmit all the information that the system is capable of passing. Experience has shown, however, that a tube possessing resolution capabilities of 600 or 700 lines actually gives a sharper, cleaner picture than one which will resolve only 500 lines. We, therefore, must attempt to secure the highest possible resolution even though it be in excess of that needed theoretically. All these requirements are satisfied through the use of a brute-force method. The gun utilizes an accelerating grid tied to the No.2 anode, the No.1 anode being at a lower potential as in the 5-inch Kinescope gun previously mentioned. The masking aperture in this case is only 0.030" in diameter, and the maximum beam current is limited to 5 or 6 microamperes. This same type of gun is used in Monoscope tubes.

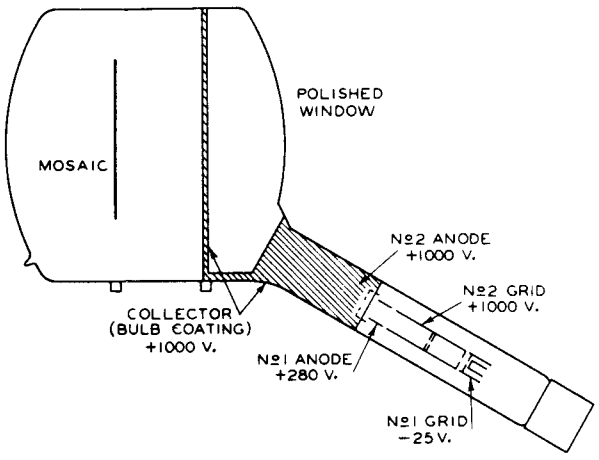


Fig. 7 - Iconoscope outline and electron gun.

Deflection of the electron beam may be accomplished by either magnetic or electrostatic fields, or by a combination of the two. From a tube manufacturing standpoint, magnetic deflection is by far the more preferable, because of its greater simplicity. Electrostatic deflection requires that four plates be mounted in the tube, either on independent leads brought through the side walls of the bulb, or on leads which pass through the stem. In the former case, an expensive sealing operation is involved, and the question of alignment is extremely troublesome. In the latter case, the addition of plates makes the mount quite long and heavy. It involves the use of multiple-lead stems which will carry quite high voltages, and brings up a rather complicated basing problem. Tube quality is more difficult to maintain because of the distortion which occurs in the corners of a scanned pattern. The same difficulties are encountered to a lesser degree when only a pair of plates is mounted in the tube, the other deflection being accomplished magnetically.

From a circuit standpoint, there are arguments both ways. Magnetic deflection requires a very carefully designed deflecting yoke and consumes considerable power. There is always a serious problem in designing the yoke to permit rapid enough return or flyback time. Electrostatic-deflecting circuits, on the other hand, do not consume power but do involve the use of high voltages. For oscillograph tubes, there is no question that static deflection is superior because of the far higher frequencies that can be used, but in television the question of picture quality has thus far kept magnetic deflection in the lead as far as any commercial planning is involved.

One advantage possessed by electrostatic deflection lies in the absence of the so-called ion spot in tubes employing this system. Many of you may have noted a small dark spot in the center of the screen of some of our Kinescopes. Customers quite frequently complain about this spot, claiming that our production department is unduly careless in allowing the spot to rest stationary on the screen, thereby burning it. Such is not the case.

As you know, all of the gas cannot be removed from a tube during evacuation. Even though we obtain pressures of less than 0.1 micron before tubes are tipped off, there is sufficient residual gas to be ionized by the passage of the electron beam. The positive ions thus formed are attracted to the cathode. Arriving there, they create negative ions which rush back up the tube, strike the screen with great force and permanently damage an area on the screen. Due to the mass of these ions, they are deflected by the magnetic deflecting fields only a small fraction of the distance through which the electrons are deflected. Thus, when a full scanned pattern is obtained on the screen, the ions are scarcely deflected at all and form a spot only slightly larger than the size of the electron beam.

Ions and electrons are affected alike by electrostatic fields. Thus, if deflection is half-magnetic and half-electrostatic, the ions will be deflected over the full pattern by the static field and will create a dark line on the screen, somewhat less intense than the spot because of the greater area involved. If full electrostatic deflection is employed, then the ions are deflected over the same area as the electrons, the damage done to the screen is diminished because of the greater area affected, and no effect is visible to the eye over the normal life span of the tube. Whether or not this ion spot will eventually force us to revert entirely to static deflection I do not know, but it is hoped that such will not be the case for the sake of ease of manufacture.

The bulb problem is one of the most serious

encountered with respect to the cost of cathode-ray tubes. As you well know, glass blowing by hand is expensive. Until we are able to order several thousand bulbs at a time and thereby justify the development of machinery for bulb manufacture, there seems no way of preventing the bulb cost from being a large percentage of total material cost. The problem is even more difficult than that; machinery which will handle bulbs with diameters of nine and twelve inches is not in existence, and it will probably take a year or two to design such equipment after such work is started.

Until recently, the majority of cathode-ray types have been made in Pyrex or Nonex bulbs. The 3-inch size has used a soft-glass bulb, and we are now changing over to soft glass in 5-inch and 7-inch sizes. The 9-inch and 12-inch tubes using soft glass have presented a serious gas problem when conventional exhaust schedules are used, and a great deal more work is needed in this respect. However, aside from the mere problem of having the glass works produce blanks of any sort, we impose a more serious problem by our quality demands. Kinescope bulb faces must be free from seeds and bubbles and of uniform thickness. Iconoscope bulb faces must be as nearly perfect optically as it is possible to make them. In fact, optical quality is so important that the face plates must be ground and polished on both sides before being sealed to the cylindrical part of the bulb. Despite these extreme steps, well over half of the plates we purchase are unfit for use.

Many types of cathode-ray bulbs have straight sloping sides joined to the face by a rather sharp curve. As bulbs became larger, we found that the bulb weight could be reduced without sacrificing strength by curving the sides and maintaining reasonably large radii on all curves. The trend in design of large tubes is more and more toward an "onion" shape; some of the European companies have gone even further than we have in this direction.

The construction of the 9-inch and 12-inch hard-glass tubes embodies a rather novel feature. It is difficult to blow a large bulb in one piece and obtain good enough optical quality. The 9-inch and 12-inch Nonex blanks are made in two parts. A face plate is stamped in a press mold, in the manner employed in making glass pie plates and cooking dishes. The conical part of the bulb is blown in the conventional way. After the face plate is polished, the two parts are sealed together on a glass lathe. This procedure provides good quality but is rather an expensive process.

The manufacture of cathode-ray tubes is not unlike other tube work in that it constitutes a continual series of difficult problems. A great many of these problems are due to the fact that there is an almost total absence of production volume and, therefore, a corresponding lack of opportunity for equipment development. However, as gun design stands at present, we are up against requirements for accuracy that are very hard to meet.

Let us take the cathode mounting for an example. A piece of 0.125" cathode tubing about 11 millimeters long is fitted with a cap over one end. The face of this cap is indented 0.002" and experience has indicated that the indentation depth should be held to ± 0.0002". A sleeve is fitted around the other end of the cathode and welded to the stem lead. During activation of the cathode the heat causes the tubing to expand some 0.009". When the heater is dropped back to operating temperature, the cathode contracts about 0.003". To prevent shorts during activation, we have to figure on a minimum operating spacing between cathode and grid of about 0.005" to 0.006".

I once ran some curves of cathode-grid spacing vs beam focus for a typical gun structure. The result was the curve shown in Fig. 8. You will note that the region of 0.004" to 0.006" spacing gives much poorer focus than a closer

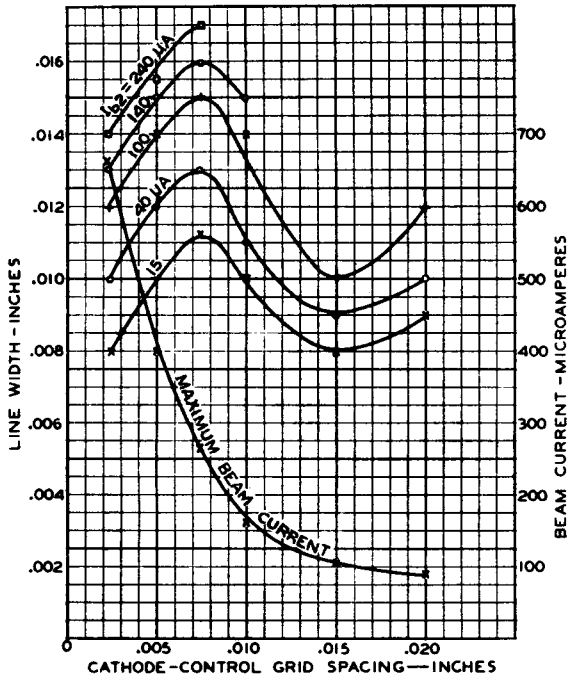


Fig. 8 - Variation of scanned line width with cathode-to-control-grid spacing.

spacing. A spacing of 0.015" again gives a good spot, but for this structure gave insufficient current. The current could be raised by opening up the grid aperture or by increasing the potential on the accelerating electrode, but this change would lower the mutual conductance. It seems, then, that we need a method of spacing the cathode very close to the grid. Furthermore, when the cathode expands, it seldom does so along the axis of the tube, with the result that the emitting area is no longer centered with respect

to the grid aperture and the surface of the cathode is not parallel to the grid.

Several methods of overcoming these difficulties have been suggested. The first step was to support the cathode by a ceramic disk located at about the middle of the sleeve as in Fig. 9. This disk was secured to the sleeve by means of two collars. The outer edges of the disk rested on a metal collar bearing on the grid. The spacing was maintained by accurate location of the collars on the cathode sleeve. This method did not eliminate all of the expansion trouble, nor did it completely center the cathode, but it did provide a mounting less susceptible to mechanical shock. Other ideas of a similar nature have been suggested, but most of them involve the disadvantage of requiring ceramics that must be held to tolerances of ± 0.001", a procedure which is not yet practical.

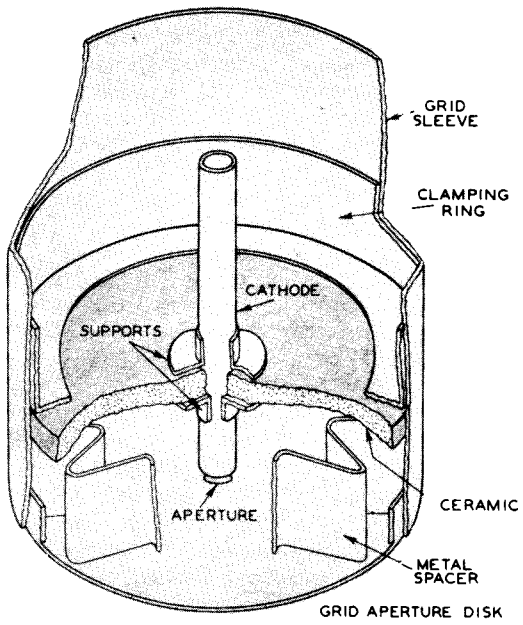


Fig. 9 - Ceramic-supported cathode.

Another source of never-ending trouble lies in the roundness, or rather the lack of it, of the anodes which form the final electron lens. The No.1 anode is conventionally made of either 1/2" or 3/4" seamless tubing. The shape of the end of this tubing contributes largely to the shape of the final lens. If the tubing is not round, we find that the field is so distorted as to produce a focused spot which is elliptical instead of round. Such a spot results in unequal resolution in the picture in horizontal and vertical directions. Because there are so many factors in electron-gun construction over which we have no control, it is extremely difficult to weigh the importance of the many factors which might contribute to making the spot elliptical. However, in our judgment, the end of the No.1 anode cannot vary more than 0.001" from a true circle. This variation means a diameter tolerance of only 0.0005". It is possible to obtain a certain percentage of parts which meet these limits, particularly if the end of the tubing is carefully shaped. However, if any strains are introduced into the material during the cutting operation, the chances are great that the tubing will warp out of shape during high-frequency treatment. Rolling over the end of the tubing has helped this situation some, but unless the material possesses just the right degree of malleability, the shrinkage on part-making is excessive.

One might logically suppose that the insertion of an aperture in the end of the tubing would be sufficient to insure a perfectly round edge. One of the reasons why we do not obtain as fine a focus as we desire is due to the fact that aberration occurs in electron lenses. If the curvature of the electric field is not such that the force exerted on any electron is proportional to the distance of the electron from the axis, all of the electrons will not come to focus in the same plane. This effect is analogous to spherical aberration in a glass lens. Since this aberration does occur in most practical lens structures, its effect can be minimized only by keeping the beam from passing through those non-uniform portions of the lens. It has been estimated that the beam diameter should be restricted to not more than about 30 per cent of the lens diameter to avoid serious aberration in our guns. With 1/2" tubing, this restriction means that the beam must be confined to a very narrow pencil. Insertion of an aperture with sufficient material left around it to provide any strength would restrict the beam diameter still further and reduce the amount of current that could reach the screen. We are somewhat better off with the 3/4" tubing in this respect, but even here the insertion of apertures has not helped a great deal. The use of tubing of much larger diameter has always appealed to us as one of the easiest ways out of this particular dilemma. However, increasing the No.1 anode beyond 3/4" would require a bigger neck, which in turn would require a larger diameter deflecting yoke. A larger yoke, we are told, would require too much power and render it difficult to attain a sufficiently rapid return time. There is a possibility that a yoke could be designed to button around a small constriction in the neck, leaving us free to use any desired size of tubing around the gun. Indeed, such yokes have been built, but are not considered practical for manufacture.

Even though we may attain the necessary roundness of the No.1 anode tubing, considerable distortion may be introduced by the glass neck which forms the No.2 anode. Greater tolerances are permissible here because of the larger diameter and the higher voltage. However, departure from a perfect circle by more than 0.0075" causes too much distortion in the spot. To obtain glass

tubing within such a limit is extremely costly. The logical solution for a No.2 anode seems to be to use a metallic cylinder supported by the mount. With this construction, however, we run into the problem of mounting insulation that will stand a potential difference of about 5000 volts without having objectionable leakage.

Since cathode-ray beams are quite susceptible to magnetic fields, it is essential that we use a non-magnetic material for our mount parts. It has been common practice in the past to use Monel, but we are trying to find an acceptable substitute. Too many lots of Monel are received which can be picked up with a magnet. Then, too, we frequently find that the material has acquired magnetic properties after it has undergone heat treatment. Nichrome has always been very satisfactory from a magnetic standpoint, but the material cannot be obtained in seamless tubing. At the present time, we are working with stainless steel and Inconel as the most likely substitutes.

Assuming for the moment that we have obtained a perfect cathode mounting, a perfectly round anode of non-magnetic material, and a perfectly round tube neck, we probably still would be able to turn out a sizable quantity of rejects because of poor mount alignment. As you noted earlier in the discussion, there may be from two to five or six apertures along the axis of the gun. These must be co-axial to within ± 0.001" if the focus is to be good. In the past we have depended largely upon mandrels projecting through these apertures to line them up. This procedure immediately imposes the limitation that the aperture sizes must all be the same or decrease as we progress toward the grid. Furthermore, when it is considered that 0.040" is a fairly common size and even 0.030" is found in some tubes, the maintenance problem on such slim mandrels can be appreciated. On low-voltage oscillograph tubes, where there is considerable tolerance on focus, we have adopted a ceramic mounting. The anodes are made of two formed halves and are welded together around two supporting ceramic rods. This arrangement appears to work successfully in the 913 and the new 902, but for tubes with higher voltage ratings there is a serious problem of leakage across the ceramic rods. To date, none of the results from this type of gun have given any indication that television quality could be maintained by this process. Our television tubes, as a result, still utilize a glass-bead type of mount. This differs from the conventional type in that, instead of using pre-formed beads, we weld all the supports to the various parts, support the assembly in a jig, and melt the glass down over the supports. This procedure prevents strains set up during welding from destroying the alignment when the mandrel is removed, but it does set up strains and warping during heating of the anodes and oxidizes the parts so that their appearance is poor.

On the projection-tube mount, we are forced to use a different technique. Here the mount consists of a series of apertures with no connecting tubing. Six apertures may be mounted in a total length of not over 3/4", the final aperture running at 6000 volts or over. Since the aperture in the final disk is only 0.004" diameter, the use of a mandrel is impossible. Instead, we try to hold very close tolerances on the outside diameter of the aperture disks and on the centering of the holes. The disks are then dropped into a V-shaped block and alignment depends upon the accuracy of the parts.

Exhaust of cathode-ray tubes until very recently has been done entirely on trolley positions. During the past two years, however, considerable progress has been made in the development of a machine. At present we have two such machines, one for 2-inch and 3-inch tubes, and one for 5-, 7-, 9-, and 12-inch types. They differ from conventional sealex machines chiefly in the size of sweeps used and in the slower indexing speed. Because of the extremely high vacuum needed, we have found it necessary to mount the mercury pumps on the turret and provide a pump for each port. The speed has been gradually increased until we can pump 3-inch oscillograph tubes at the rate of 60 per hour. Television tubes can be pumped in from 20 minutes to 1 hour per revolution of the machine, depending on the permissible baking temperature, the size of the bulb, and the type of fluorescent-screen material used. A baking oven is mounted over the track for the first six positions. Mr. Kaufmann dealt quite completely with the subject of screen materials in Lecture 5, so I will mention them only briefly. As you know, the most common type of screen is the so-called green willemite, which is zinc orthosilicate. The silicate screen materials are quite stable and easily handled and present no serious problems on exhaust. When the demand for a white screen for television arose, a great deal of work was done with mixtures of blue and yellow zinc sulfides. The yellow sulfide in particular was quite unstable and very difficult to handle in production. If the baking temperature was raised above 350°C, the screen was likely to assume any one of a number of undesirable shades. On the other hand, if the temperature was kept below 425°C, the tubes generally failed in five or ten hours because of gas. One possible solution involves the use of a mixture of blue sulfide and yellow silicate material. To date, this mixture seems to have most of the desirable characteristics of the silicate materials. Machine exhaust seems quite feasible and the chief problems are to develop further the process of making blue sulfide and to maintain sufficient purity in the processing. Extremely minute quantities of copper are sufficient to cause changes in the color of the screens; and the entire processing of the tubes must be free of any contamination. Both of these problems can be solved, we feel, and the white material is expected to become standard on television tubes before long.

There remains to be said a brief word about testing cathode-ray tubes. In addition to many of the electrical characteristics which are measured on receiving tubes, we are interested in the questions of focus and brilliance. We first try to determine "optimum conditions". These might be defined as those conditions which will give us the brightest possible picture of the required resolution. Resolution can, of course, be determined by measuring the size of the focused spot. However, any appreciable amount of current in a stationary spot would damage the screen irreparably, so that it is necessary to apply scanning voltages and measure the width of the scanned line by the use of a calibrated telescope. The widest line that will give the required resolution is chosen depending on the size of the tube. The grid bias is then increased until the scanned line becomes the predetermined width and the conditions of bias and focusing voltages are noted. Herein lies the first inaccuracy in the testing method. It has been shown that the distribution of electron density across the width of a scanned line is approximately Maxwellian in nature. The light derived from a fluorescent screen depends upon the number of electrons striking the screen. As the spot is scanned across the screen, it is obvious that any unit area traversed by the center of the spot will be bombarded by more electrons than a similar area near the edge of the spot. As a result the brightness of a line decreases from the center to the edge. When viewed through a telescope with a magnification of some 50 times, the edge of the line is very indistinct, and the judgment of the operator enters into the reading. Furthermore, if the spot is elliptical and the long axis of the ellipse coincides with the direction of scanning, an optimistic reading will result.

Since the apparent line width varies with the velocity of scanning, it has been considered necessary to scan the screen during testing with a velocity comparable to that of a television picture. Circuits generating saw-tooth voltages at about 13000 cycles per second are rather temperamental and are not suited to factory usage. Instead, we use a sine-wave scanning voltage of a much lower frequency and adjust the length of the scanned line so that the velocity of the spot at the center of the screen is the same as that achieved with the higher-frequency saw-tooth voltage. Sine-wave scanning offers another advantage. Since the spot must slow down and reverse its direction, the intensity of the ends of the lines is higher than elsewhere. The bright spot thus formed at the end of each line has virtually the same appearance as a stationary spot. By looking at the end of a line, the operator can judge the ellipticity of the spot. Specifications call for adjusting the focus, not for the narrowest line, but for that condition in which the spot at the end of a line has the most nearly symmetrical appearance — again a matter of judgment. In some types of television tubes, we have been focusing for the narrowest horizontal line, then rotating the yoke 90 degrees and measuring the line width in the vertical direction. The height and width of the spot, plus the width of the line under symmetrical conditions, gives us a measure of the resolution capabilities of the tube. Testing in this way is expensive and not satisfactory for production.

Having measured resolution, we must then test the performance of the fluorescent screen. Formerly, we required two light measurements; now we measure only once and calculate the other reading. The beam current is reduced to 50 microamperes, and the pattern size reduced to a standard area, usually 6 x 8 centimeters. The light output is measured from this pattern, and the screen efficiency calculated in terms of candlepower output per watt input to the screen. In the factory this reading is taken by means of a photronic cell whose output is measured on a light-beam galvanometer. The cell is first placed over the scanned pattern, and the meter reading noted. The cell is then placed over a ground-glass plate behind which is mounted a calibrated lamp on a movable socket. The distance from the lamp to the screen is varied until the cell shows the same reading on the meter as before. The light output is then determined from a chart showing light output vs distance of the lamp from the screen. This equipment requires frequent calibration, but in general is considerably more satisfactory than the illuminometer used by the laboratory. Some sort of light reading will probably have to be taken until that far-off day when we can depend upon the uniformity of fluorescent materials to a much greater extent than we can to-day.

Lecture 25

ELECTRON BEAMS AND THEIR APPLICATION IN RADIO TUBES

H. M. Wagner

BEAM TUBE DEFINED

The term "beam tube," as used here, covers a wide variety of devices. The term is sufficiently new so that it does not yet have a generally accepted meaning. Of the many different tube types manufactured by RCA, only the "Beam Power Tubes," such as the 6L6, 25L6, and 807, are catalogued with the "beam" designation. Other types, which are beam tubes, are the many cathode-ray tubes and the "Magic Eye" types 6E5, 6G5, 6N5, 6U5, and 6AF6-G.

A tentative definition[1] of "Electron Beam Tube" has been proposed as follows: "An electron beam tube is a vacuum tube providing a directed flow of charged particles substantially in the form of a beam or beams and in which the beam formation contributes materially to the performance characteristics."

It is not enough, therefore, to classify a tube as a beam tube because its space current is segregated into beams, except when some new or improved performance is achieved by means of the beam formation. Among the tube types already cited, the beam power tubes depend on the segregation of the electrons into beams between the grid wires without deflection of the beams. The cathode-ray and magic-eye tubes depend on the properties of focusing and deflection of electron beams.

METHODS OF STUDYING BEAM FORMATION AND PROPERTIES

Information on electron paths is an essential part of the design of beam tubes. These paths can be found and studied using the following methods.

A. Without Building Tube

1) Graphical Method Using Electric-Field Plot

In this method a plot is made of equipotential lines throughout a cross-section representing the actual tube. An electrolytic tank is commonly employed. Metal parts are shaped similar to the electrodes of the tube and are placed in a tank of slightly conducting electrolyte, such as drinking water, to form an enlarged section of the tube. The parts are connected to potentials proportional to the voltages to be used on the tube electrodes and a movable probe dipping into the electrolyte takes the potential at any point. By having a null voltage indicating device between the probe and an adjustable voltage source, the probe can be guided along lines of constant voltage and the equipotential lines plotted. The equipotential plots in Figs. 2 to 7 were made by the above method.

The path of an electron can then be plotted from the electric-field plot by various means involving subdivision of the path into a number of short straight lines approximating the true curved path.

This method is laborious and requires considerable time for plotting the many possible paths of electrons emerging from different parts of the cathode and having different initial velocities. Initial velocities can be disregarded in many cases.

2) Rubber-Membrane Potential Contour

Parts representing a cross-section of the tube are made out of wood or other material. These are mounted with their plane surfaces at depths below the cathode plane proportional to the corresponding tube-electrode voltages to cathode. An electrode at negative voltage, such as a biased grid, is higher than the cathode. A thin rubber sheet uniformly stretched by a frame is held over the parts so that it presses against all the surfaces. It can be shown that a thin flexible elastic membrane held under uniform tension and placed in this way over parts having relatively small differences in elevation between them has a topographical surface the same as that of the electric field in the tube and that the path of a ball allowed to roll from the cathode is the same as that taken by an electron in the tube. Initial velocities of electrons are simulated by starting velocities of the balls. The electron trajectories are determined by observation or photographing the paths of the balls.

3) Langmuir Automatic Electron-Path Plotter

The radius of curvature of an electron moving in an electric field can be shown to be equal to

$$\frac{2\,(V + V_o)}{\frac{dV}{dr}}$$

where V is the electric potential in space; V_o, the initial velocity with which the electron enters the electric field; and dV/dr, the electric gradient normal to its direction of motion. In Langmuir's apparatus,[2] a double-pointed probe

[1] Proposed by Subcommittee on Electron Beam Tubes of the I.R.E. Technical Committee on Electronics, November, 1937.

[2] D. B. Langmuir, "Nature," Vol. 139, p. 1066; June 19, 1937.

dips into an electrolytic tank holding electrodes representing the tube cross-section. The small difference of potential between the points of the probe, which are close together and insulated, gives the potential gradient, while, simultaneously, the voltage on them gives the electric potential. The probe is supported by a motor-driven carriage, the wheels of which are steered by a vacuum-tube-controlled electric motor that makes the carriage move in a path having a radius of curvature proportional to the ratio of the voltages on and between the points of the probe. The carriage and probe thus trace the electron path. A plot is made by a pencil attached to the carriage.

The three foregoing methods of plotting electron paths without building a tube are of considerable value in beam-tube design, although their application is limited. One limitation is the neglect of electron space charge, which is important in many instances. The class of tubes in which a magnetic field influences or controls the electron paths cannot be studied by this means. Because of the geometrical form of many tubes, the electron paths cannot be found or else they can be determined only for certain cross-sections through the tube. Specifically, the electron paths which can be determined are those which lie in a plane surface,[3] and the cross-section through the tube must be chosen so that this plane cuts no equipotential lines, i.e., the equipotential lines must be completely contained in the plane.

As an example showing the geometrical limitations of the methods, consider electron paths in a triode, such as the RCA-56, having a cylindrical cathode and anode and a helical-wound grid. The helical grid does not permit a plane to be passed which does not intersect lines of force. As an approximation, a ring-type grid can be assumed and substituted. Then there are three planes that can be passed which satisfy the requirements, i.e., one perpendicular to the cathode and halfway between grid rings, one parallel to and through the axis of the cathode and grid side-rods, and another through the cathode axis and at right angles to the plane through the grid side-rods. Electron focusing between grid wires occurs in the latter plane and can be examined by making a model of half of the tube and immersing it in the electrolytic tank so that the plane through the model is at the surface of the liquid. But, placing the rubber membrane over the straight edges of the cathode and anode pieces and the circular cross-section grid wires simulates and shows the grid-control action of a parallel-plane structure instead of the cylindrical tube. In general, the geometrical configurations suitable for the rubber membrane are more limited than for the electrolytic tank and must not only have a cross-sectional area containing a two-dimensional electric field but they are also limited to two-dimensional mechanical symmetry, i.e., they must have the same cross-section throughout the length.

B. By Visual Observations on Actual Tubes

1) Tubes Containing Small Quantity of Gas

The presence of a small quantity of gas makes visible the region of electron flow and shows the outline of the beam. The quantity of gas needed to make the beam distinctly visible and hence the change in electron paths from good vacuum conditions, depends among other things upon the depth of the beam cross-section looked through. The greater the depth of beam, the smaller the gas pressure required. In observations on various beam tubes having willemite-coated electrodes, H. C. Thompson found that gas can be introduced in the tube in quantity sufficient to give a glow without causing observable change in the luminescent traces on the willemite. It would appear, therefore, that the electron path can be only slightly different than in vacuum.

The gas can be introduced into the bulb by heating the tube parts, by not flashing the getter, or by the introduction of specific gases in controlled quantities on an exhaust system.

2) Tubes with Solid Metal or Transparent Metal Electrodes Coated with Willemite

While the glow in gas is throughout the beam, it is often sufficient to know the size and shape of the beam cross-section along the path, for example, where the beam strikes an output electrode or passes through an aperture. Willemite-coated parts at the desired places in the tube and operating at appropriate voltages show the luminescent beam traces. In some tubes certain parts have been made movable in order to observe the beam traces for different electrode positions. The best location of an electrode for sharp focus has been determined in this way.

A technique of utilizing willemite for observation and study of electron beams was developed and used extensively by H. C. Thompson in his research work on beam tubes.[4] It consists of coating the electrodes (the plate, screen grid, etc.) with a very thin coating of willemite like that used on the target of the magic-eye tubes, in contrast to the thick layer applied in cathode-ray tubes. It is important that the

[3] It is conceivable that the electrolytic-tank or Langmuir method with a submerging probe (three-pointed in the Langmuir method) could be adapted to plot paths in three dimensions, but the design and constructional difficulties involved might easily make this adaptation impractical.

[4] H. C. Thompson, "Electron Beams and Their Applications in Low Voltage Devices," Proc. I.R.E., Vol. 24, No.10, pp. 1276-1297; October, 1936.

willemite be coated thin, particularly on electrodes operating at low or medium voltages which are commonly used in radio tubes, because willemite is an insulator and unless it is in a thin film the voltage drop due to electrons striking and passing through the willemite can be appreciable, and hence the potential on the coated electrode is different from that on an uncoated one. In the early manufacture of the magic eye, some targets were too thickly coated with the result that in some tubes the target was completely dark while others were dark only in parts or spots where the resistance between the willemite and the target was sufficient to drop the potential to a voltage below the visible threshold. The electrodes in many of H. C. Thompson's tubes have been made of sheet metal on which parallel millimeter-spaced lines were marked, and thus have the appearance of graph paper when there are two sets of lines at right angles. Measurements of beam width and deflection are relatively easy with tubes having coordinate-marked, willemite-coated electrodes. The beam is measured by the coordinate lines and can be seen directly or with the aid of a microscope.

Transparent metal films can be coated with willemite and used as electrodes. Their advantage over solid metal electrodes is that the luminescent traces are visible from either side. The metal film, when it is thin enough to be transparent, has a moderately high electrical resistance so that there is a voltage drop along its surface due to the beam current. This potential difference can be small. Various tuning indicator tubes which I have made use targets in the form of a clear mica disc coated with a film of metal and willemite over the metal. Electrical and mechanical contact is made to the disc around its circumference. In this particular application of transparent films, the maximum voltage drop along the target for a few milliamperes beam current need not exceed ten volts.

3) Incandescence of Thin Metal Foil

A sheet of thin metal can be heated to incandescence in areas bombarded by an intense beam of electrons. Due to thermal conductivity along the metal sheet, the incandescence is not confined to the bombarded area but extends beyond it some distance depending on the thickness of the metal. I have tried tantalum foil as thin as one micron. Even foil as thin as this is inadequate where good beam resolution is needed, although it is good enough for some purposes.

4) Incandescence of Wire Mesh, Screen, or Helix

Surfaces other than thin foil can be used as incandescent beam-trace indicators. Wire mesh is one possibility. A row of closely spaced parallel wires[5] will show the spread or deflection of a beam from wire to wire. Each wire lights up as the beam passes over it. The heat conductivity between wires is substantially zero but there is some heat radiation. A fine wire helix, such as a lamp filament, will serve as an indicator. The twisting of the wire into the helical shape provides an increased length of path through which the heat must travel.

5) Incandescence of Soot on Electrodes

Soot deposited from a flame has a loose texture and is a poor heat conductor. This thermal insulating property makes it excellent as an incandescent indicator. Observations on an RCA-56 triode having a sooted anode give some quantitative idea of the efficacy of the surface. The parts in this tube are standard except that the carbonized nickel anode was held in a candle flame so as to coat its interior with a layer of soot. The soot adheres well to the carbonized metal base. During evacuation the anode was heated as hot as possible to degas the soot. I have observed the beam traces on the anode for various conditions including focus. At focus there are bright lines on the anode opposite the spaces between grid-wire turns, the lines and grid turns spaced 37.5 to the inch. The incandescent lines stand out sharp and distinct, their width being small (approximately 0.005") compared with the 0.027" spacing between them. The very abrupt change from light to dark at the boundary of the electron beam on soot is far superior to the beam traces on tantalum foil having a thickness of only one micron. The bombarded area of the soot may be intensely bright while the remaining surface and the exterior of the anode are dark.

Sooted and other incandescent surfaces have a sphere of usefulness in beam tubes somewhat different from that of willemite. Whereas the illumination from a willemite surface is approximately proportional to the beam density, the brightness of an incandescent screen varies much more steeply and shows high contrast between portions of the beam of slightly different intensity. Data from the tables by Jones and Langmuir[6] on the characteristics of tungsten are plotted in Fig. 1 and show this change. At a temperature of 1100°K, the tungsten receives and radiates 1.027 watts per sq cm and has a light output of 0.00107 candles. When the power input to the screen, and hence the electron-beam density, is doubled, there is a 12-fold increase of brightness. High contrast is advantageous for examin-

---

[5] M. Knoll, "Zeitschrift für Technische Physik," Vol. 15, No.12, pp. 584-591; 1934.

[6] Dr. Howard A. Jones and Dr. Irving Langmuir, "The Characteristics of Tungsten Filaments as Functions of Temperature," G. E. Review, Vol. 30, Nos.6, 7, and 8; 1927.

ing the distribution of electron density throughout a beam. Incandescent surfaces can be applied only to beams of high intensity since the surface is totally dark below a moderate degree of electron bombardment.

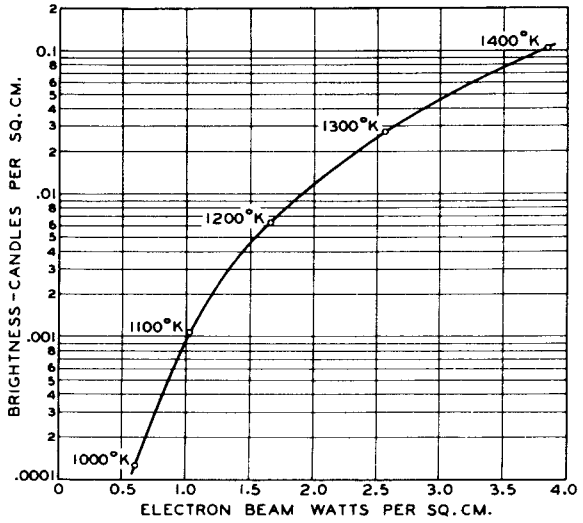


Fig. 1 - Indication of beam traces by incandescence of tungsten foil. (Data on tungsten filaments from Jones and Langmuir, reference 6)

An application of the sooted surface is its use in a type of tuning indicator tube having intersecting electron beams. Two identical sharp beams bombard the surface and are deflected so that their traces can be made to overlap. The current in the individual beams is adjusted to a point where a dull red glow on the soot is barely visible. When the beams are deflected, their overlap has double the electron density of a single beam and glows brilliantly.

C. By Electrical Measurements on Actual Tubes

The electrical characteristics of a tube are often the end result of design rather than the means for beam study. However, these electrical measurements on final tube designs or special experimental types can give valuable quantitative information about the beam and moreover show what can be done electrically by means of the beam. They can show how much of a beam may be confined to a given cross-section or caught by certain electrodes; or, a movable electrode or deflected beam can explore the current density throughout the beam. Unless care is taken the presence of secondary emission or stray electrons may lead to an erroneous interpretation of the results.

EXAMPLES OF BEAM FORMATION

Contrary to popular belief, good beam formation can be had in relatively simple structures. Certain work on beams has been undertaken with the object of simplifying tube construction and of obtaining novel electrical characteristics. An instance of this is O. H. Schade's investigation[7] of tubes with flat-apertured, punched plates replacing wire-wound grids.

The three types of beam-forming structures which follow have been selected because of their mechanical simplicity and practical significance.

A. Cathode Between Parallel Planes

Fig. 2 shows a cylindrical filament or cathode between two parallel plates. The cathode is at space potential, i.e., the potential that would exist at the axis of the cathode if the cathode were not there. As shown, there is a positive accelerating field for electrons on the side of the cathode facing the more positive electrode and a retarding field on the other side. The cathode equipotential line divides the cathode so that only half of it emits. The electrons from the emitting half of the cathode's circumference tend to move in the direction of maximum potential gradient, i.e., normal to the equipotential lines of the field plot or down hill on the rubber membrane model, so that when they arrive at the positive plate the beam trace subtends an angle at the cathode that may be much less than 180°.

The beam trace is wide for the cathode at space potential. It can be narrowed by raising the cathode potential, or, by what is equivalent, lowering the potential on either or both plates. The narrowed beam trace is shown in Fig. 3 in which most of the cathode surface lies in a retarding field below the zero equipotential line, the emission coming from only a narrow strip on the cathode. The electrostatic lines are sharply curved in this case whereas the cathode when at its space potential causes hardly any disturbance of the parallel-plane equipotentials.

Tests on a tube of this type show sharp-edged boundaries of the beam. One plate of this tube is operated at 250 volts and the other plate is operated at a variable negative potential. When the negative voltage is increased, the beam trace narrows and the beam current decreases. For a tube with a 0.050"-diameter cathode spaced one centimeter away from the positive plate, the beam current per centimeter length of cathode (depth of beam) is approximately one milliampere for a beam-trace width of one millimeter.

When the bias on the negative plate exceeds a certain voltage, the cathode is surrounded by a retarding field and the electron emission is wholly cut off. The equipotential lines are then the same as those shown in Fig. 7 for a grid at a voltage above space potential, if we substitute

[7] Ref. 25-7.

in the diagram a row of cathodes for the grid wires.

B. Helical Grid

The presence of beams in conventional tubes having grids is not new. There have been beams due to grids ever since DeForest put a grid inside of a vacuum tube. But the knowledge of their properties, their use and application, as in beam power tubes,[8] is of more recent date. The following information on beams formed by grids comes mostly from the work of H. C. Thompson. Figs. 4 to 9 are from his paper.[4] The figures refer to a parallel-plane structure but the beam behavior is similar to that for the helical grid.

1) Beams Between Grid Wires

When a grid is at its space potential (Fig. 4), its presence in the space has relatively little effect on the equipotential lines. Hence, the electron paths are substantially the same as they would be without the grid except for those

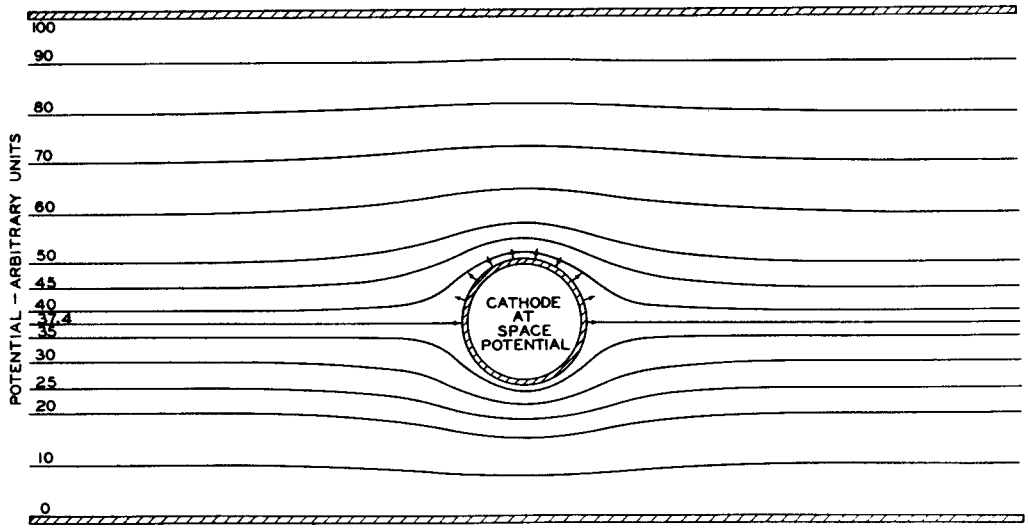


Fig. 2 - Equipotential map of region around cylindrical cathode between two parallel plates, with cathode at space potential. Arrows on portion of cathode circumference indicate an accelerating field and electron emission. The remainder of cathode circumference is in retarding field with no electron emission.

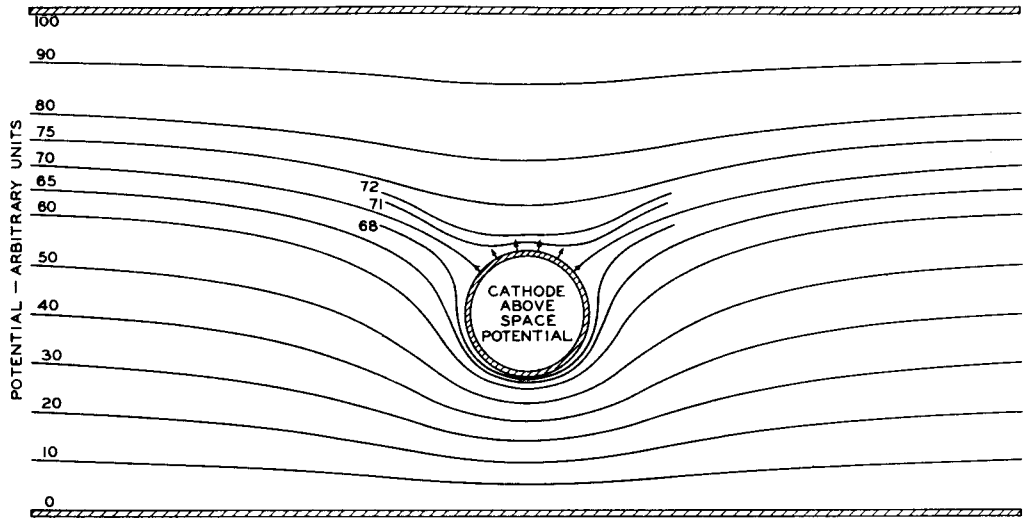


Fig. 3 - Same as Fig. 2 except that cathode is above space potential.

[8] O. H. Schade, "Beam Power Tubes," Proc. I.R.E., Vol. 26, No.2, pp. 137-181; February, 1938.

[4] Loc. cit.

electrons which are caught by the grid wires and leave gaps in the electron flow to the anode. An electron shadow of the grid appears on the anode, its size being a geometrical projection of the grid wires. When the equipotential lines are parallel, as between a plane parallel cathode and anode, the electrons move in parallel paths leaving shadows on the anode of the same width as the grid-wire diameter.[9] When the electron paths are diverging or converging, the shadow size is magnified or reduced.

The equipotential plot when the grid is above its space potential (Fig. 5) shows a diverging field in which the beam widens on passage

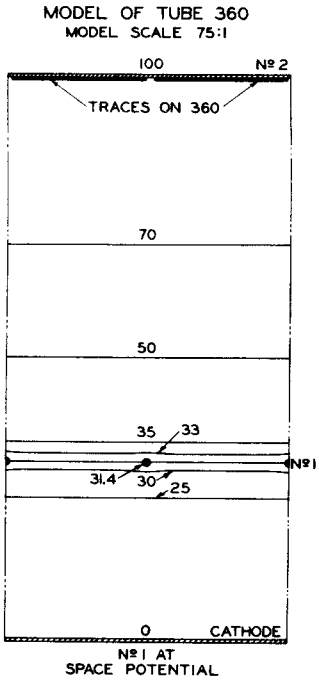


Fig. 4 - Equipotential map on scale model, with grid No.1 at space potential. (Courtesy of The Institute of Radio Engineers)

between the grid turns. As the grid potential is increased above space potential, adjacent beams widen and fill the gaps left behind the grid wires. The beam traces widen and cover the entire anode; then they overlap. On a fluorescent anode, we can see approximately double brightness at the overlapping beams opposite grid wires. A photograph of beam traces on a willemite-coated anode in a parallel-plane electrode tube is shown in Fig. 8 for various grid voltages and a space potential of 76 volts. Fig. 9 is an approximate outline of the beam.

[9] Actually the shadow size is smaller since the electron space charge crowds the electrons into the space-current gap left by the shadow of each grid wire.

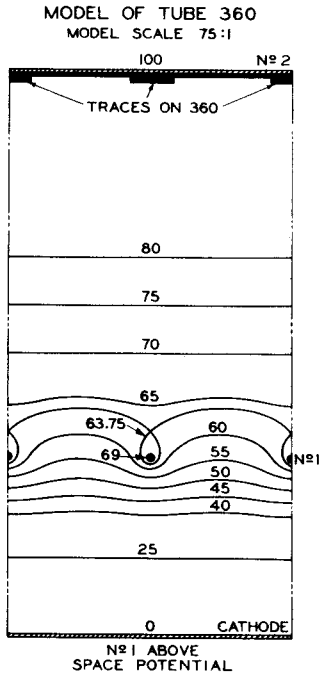


Fig. 5 - Equipotential map on scale model, with grid No.1 above space potential. (Courtesy of The Institute of Radio Engineers)

Below space potential, the field of the grid wires is converging, as shown in Figs. 6 and 7. At a certain grid voltage below space potential, the beam converges to a focus on the anode (Fig. 6). At a lower voltage (Fig. 7), the elec-

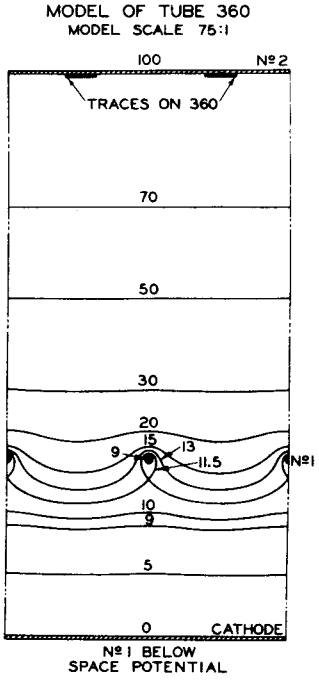


Fig. 6 - Equipotential map on scale model, with grid No.1 below space potential. (Courtesy of The Institute of Radio Engineers)

trostatic lines are more sharply curved and bring the beam to a focus short of the anode where there is a cross-over of electrons so that the beam widens in going the rest of the way to the anode. As the grid voltage is lowered further, the focus recedes from the anode and the beam

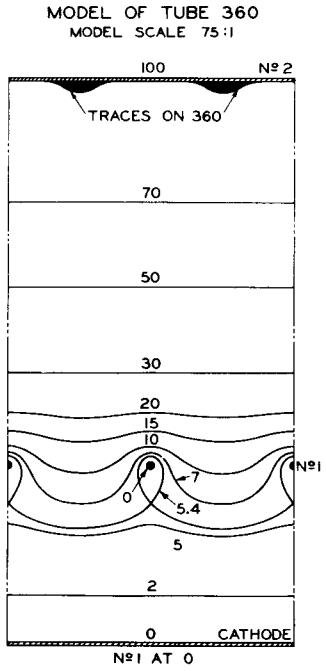


Fig. 7 - Equipotential map on scale model, with grid No.1 at zero potential.
(Courtesy of The Institute of Radio Engineers)

trace widens. However, when the grid becomes negative near the cut-off voltage, the emission is first cut off from the portions of the cathode directly beneath the grid wires and last from halfway between wires. Hence, as the grid becomes more negative and the focus recedes from the anode, the emitting area on the cathode narrows so that the beam trace becomes narrow at

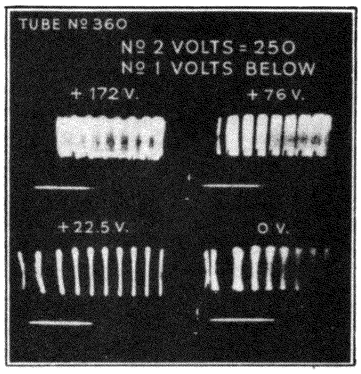


Fig. 8 - Beam traces on an anode.
(Courtesy of The Institute of Radio Engineers)

close to cut-off voltage on the grid. This narrow trace is not due to focusing action by the grid but results from the restricted emitting area at grid cut-off.

2) Effect of Anode Potential on Beams

When all voltages on the electrodes of a tube with reference to cathode are varied in the same proportion, the electron paths remain the same.[10] This statement is true both with space charge and with cathode-temperature-limited con-

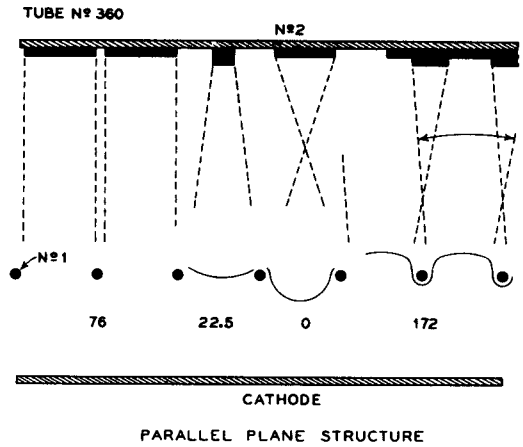


Fig. 9 - Beam formation by grid openings.
No.2 volts = 250

No.1 Volts	Beam Focus
0	Before No.2
22.5	On No.2
22.5 - 76	Beyond No.2
76	At infinity
172	Virtual

(Courtesy of The Institute of Radio Engineers)

ditions over a range where space charge is negligible, that is, where the paths do not vary with voltage for either unlimited cathode emission or with negligible cathode emission, but if the emission becomes limited as the voltages are raised, the paths change in changing from space charge to a limited-emission state. Thus, in a triode, if the voltages are varied so as to maintain some fixed ratio between grid and anode potentials, the electron paths do not change. Since it is only the ratio of voltages that determines the beam path, the effect of a variation in anode potential can be considered in terms of a change in grid voltage in the opposite direction. An increase of anode potential corresponds to a decrease in the magnitude of the grid voltage. If the grid voltage is negative, a decrease in its

[10] This statement neglects initial velocity of electrons from the cathode. The small velocities that exist become important at low voltages.

magnitude is in the direction of raising its potential and hence decreasing the convergence of the electron lens. If the grid voltage is positive, the increased anode potential is equivalent to decreased grid voltage and greater convergence of the lens. Thus, an increase in anode potential causes either an increase or decrease in convergence of the electron beam depending on whether the grid is positive or negative. The direction of change in beam-trace width with anode potential depends not only on the polarity of the grid voltage but also on whether or not there is a cross-over of electrons or focus in the space in front of the anode.

The tube of Fig. 10, devised by H. C. Thompson, shows a possible way of utilizing the variation of the beam trace with anode potential for a two-terminal oscillator in which the anode negative-resistance characteristic is due to beam action instead of secondary emission, as in the

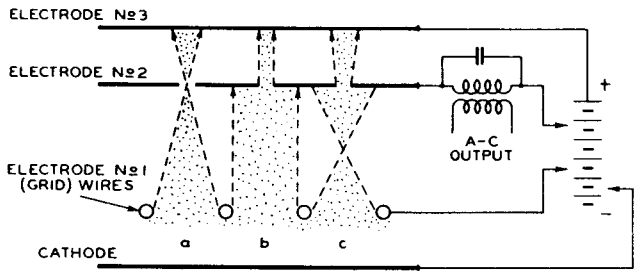


Fig. 10 - Tube utilizing variation of beam-trace width with anode potential for a two-terminal oscillator in which the anode negative-resistance characteristic is due to beam action instead of secondary emission.

dynatron. For certain initial voltages on the tube electrodes, the beam focuses on electrode No.2 so that all the electrons pass through the narrow slits in this electrode to be received by electrode No.3. Change in voltage on a single electrode defocuses the beam. Hence, a rise or fall in potential on electrode No.2 defocuses the beam and permits part of the beam to be caught by electrode No.2. The rising current to electrode No.2 for decreasing voltage provides a negative-resistance element for exciting the tuned circuit between electrode No.2 and the positive supply voltage.

3) Effect of Grid Side-Rods

The grid support-rods split the emission from the cathode into two sector-shaped beams passing between the rods, each rod casting a sector-shaped electron shadow. The size of shadow (or narrowness of beam) increases as the grid voltage is made increasingly negative. In a tube having a cylindrical anode and cathode and a regular two-side-rod helical grid, the pattern on the willemite-coated anode appears as two rows of arcs, the length of arc or angular spread of beam on the cylindrical anode being governed by the side-rods, while the width of each arcuate line is due to the beam formation of the grid turns. In triodes where the grid is negatively operated and hence below space potential, the grid-rod shadow angle may be appreciable so that part of the anode circumference is at no time bombarded by electrons and may operate considerably cooler than the active portion of the anode.

When the grid is positive and somewhat above space potential, the entire anode surface receives electrons. Just as beams between adjacent grid turns widen and overlap as the grid voltage is raised above space potential, the two beams from the side-rods spread out to occupy a greater angle of anode circumference. When each beam angle exceeds 180°, the traces overlap on the anode opposite the side-rods.

When the grid is far negative, near cut-off, the angle of beam spread on the anode may be small with certain grids. Side-rods alone, without grid winding, provide narrow beams at negative voltages near cut-off. An equipotential plot[4] shows that the emission comes only from two areas on opposite sides of the cathode which are midway between the rods. In tubes with grid windings of elliptical cross-section, such as in an RCA-56, the combined field of the grid winding and side-rods provides sufficient uniformity of field around the cathode circumference so that the beam angle is relatively wide near cut-off voltage.

C. Orbital Beams

By the term "orbital beams" is meant any of various, curved, beam formations. The beam as a whole may be curved or bent around so that it arrives at an electrode hidden from the electron source. An application of orbital beams is in secondary-electron multipliers. Certain secondary-emissive surfaces have been found to lose their sensitivity upon exposure to an oxide-coated, thermionic cathode. This desensitization is avoided in both magnetic and electrostatic multipliers by curving the beam so that the secondary cathode receives electrons but not evaporated material from the thermionic cathode.

The type of orbital beam to be discussed here is that due to a radial electrostatic field.

1) Concentric Cylindrical Electrostatic Focusing System

The focusing properties of a radial electrostatic field have been examined by Hughes,

[4] Loc. cit., Fig. 5.

Rojansky, and McMillen,[11] and are illustrated in Fig. 11. There are two concentric cylinders with the inner one at a higher potential than the outer one. This arrangement provides an inward-gradient, radial, electrostatic field. An electron of a certain velocity can be introduced tangentially into the space so as to travel around in a circular path. This condition exists when the centrifugal force of the electron moving in a circle concentric with the cylinders is exactly counterbalanced by the electrostatic attraction

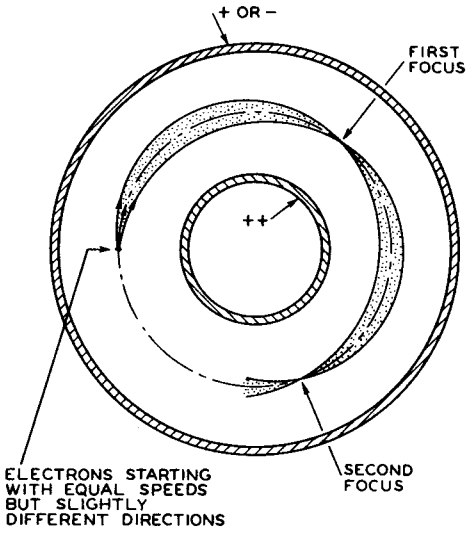


Fig. 11 - Focusing with the electrostatic field of concentric cylinders.

toward the center. If another electron enters the space at the same place and same speed as the tangential electron, but in a slightly different direction, it will move in a curved path which crosses the circular path. Hughes, Rojansky, and McMillen show that a group of electrons starting nearly tangentially with equal speeds cross each other and come to a good focus at an angle of $\pi/\sqrt{2}$ (= 127°-17').

H. C. Thompson several years ago examined the properties of the radial electrostatic field with a view to practical applications. The adaptation of the radial focusing field to radio tubes having beam currents of several milliamperes at low voltages introduces problems not considered in the above analysis. The electrons come from a relatively broad straight cathode instead of from a line source and may enter the field as a broad beam. The introduction of control grids, screen grids, and output electrodes in the space leaves only an approximation of the radial field.

[11] Hughes, Rojansky, and McMillen, "On the Analysis of Electronic Velocities by Electrostatic Means," "Refocussing of Electron Paths in a Radial Electrostatic Field," Phys. Review, Vol.34, No.2, pp. 284-295; July 15, 1929.

Thompson has found from observations on tubes with willemite-coated electrodes that a focus occurs at some angle, and that the beam can be deflected appreciable distances, if the voltage on either cylinder is changed, and yet remain in focus. Hence, it is possible to move a narrow beam trace back and forth over a radial vane located at the correct angle. He has found a second focus at approximately twice the angle of the first focus. In my work I have designed and tested a variety of structures to serve as a basis of tube design. The quality of focus changes appreciably with angle.

In brief, the concentric cylindrical structure has inherent beam-focusing and deflecting properties. Use has been made of it, as will be shown later, in secondary-electron multipliers, beam-deflection amplifiers, and converter tubes.

SPACE-CHARGE AND TEMPERATURE-LIMITED CONDITIONS

Various things have been said about the effect of space charge on the electron paths. In order to illustrate this effect, I have taken data, plotted in Fig. 12, which indicate the grid-voltage variation necessary to maintain the best focus on the willemite-coated anode of a triode as the space current is varied by changing the cathode temperature. The triode has a cylindrical cathode of 0.050" diameter, cylindrical anode of 0.790" diameter, and helical grid of 0.160" diameter wound 10 turns to the inch with 0.005" wire.

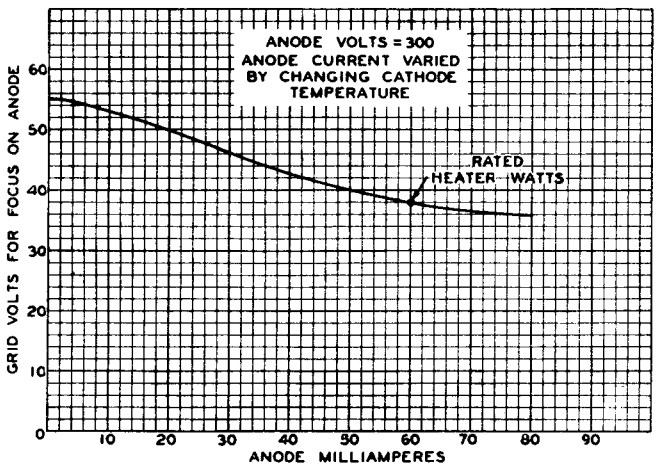


Fig. 12 - Variation of grid focusing voltage with space charge in triode.

The tests were made with a constant potential of 300 volts on the anode and the grid voltage was varied to give focus with different cathode temperatures. At normal heater input, the anode current was 60 milliamperes with 38 volts on the grid for focusing. When the heater input was reduced until the anode current was limited to a

small fraction of the space-charge current (observations made for anode currents as low as 0.1 milliampere), the potential on the grid had to be raised to 55 volts for best focus.

The need for raising the grid voltage to restore focus when the cathode temperature and space current were reduced can be considered in terms of the changing convergence of the grid lens. Reduction of space current causes a rise in potential throughout the space between the electrodes. The grid voltage, if unchanged, then becomes relatively lower with respect to the space and hence the lens formed by the grid wires becomes more convergent, and brings the electrons, which were in focus on the anode, to a focus short of the anode. The grid voltage must be raised to refocus on the anode and show the minimum beam-trace width.

SOME APPLICATIONS OF BEAMS IN COMMERCIAL TUBES

The following table gives instances of the utilization of beams in commercial tubes.

Application	Tube Type	Date
A. Reversed grid windings	47	1929
B. Elliptical grid	56, 57, 58	1931
C. Magic eye	6E5	1935
D. Beam power tube	6L6	1936
E. Low-noise, low-screen-current tube	EF-9 (Philips)	1937
F. Anodes of small dimensions	6K8, 1851	1938

A. Reversed Grid Windings

In the early development of power output tubes using screen grids, it was noticed that some of the screen-grid wires in sample tubes operated red hot while others were cool. This effect, explained by W. R. Ferris in U.S. Patent 2,047,019 was due to the impinging of electron beams formed between the control-grid wires which happened to be positioned approximately halfway between the control-grid turns. At that time the control grid and screen grid were wound in the same direction, so that complete screen-grid turns could be exposed to the electron beams from the control grid. This arrangement was changed in the above patent by winding one of the grids right-handed and the other left-handed so as to permit bombardment of only portions of the screen-grid turns, instead of complete turns. In this way, the maximum screen-grid temperature was lowered because of heat conduction from the short bombarded lengths or arcs of the screen-grid wire.

Tubes using reversed grid windings have been referred to as anti-beam tubes because of the chopping up of the control-grid beams by the screen grid. However, the Ferris patent shows diagrammatically the presence of the beams and covers an invention based upon their knowledge.

B. Elliptical Grid

One effect of a grid side-rod, as has already been mentioned, is to leave a gap in the space current to the anode. There is also an increase in amplification factor around the cathode circumference in the neighborhood of the side-rods of such a magnitude that with a circular cross-section grid the region on the cathode beneath the side-rods may be practically non-emitting throughout most of the negative grid-voltage range. In order to get a more uniform cut-off and an increased active emitting arc of cathode (B.Salzberg, U.S. Patents 2,073,946 and 2,100,723) the spacing is increased between grid side-rods to make an almost elliptical grid. The sector-shaped beams due to the wide-spaced side-rods of the elliptical grid are relatively wide even at high grid bias.

C. Magic Eye

The magic eye (H. M. Wagner, U.S. Patents 2,051,189 and 2,122,268) uses a vane or narrow strip parallel to a cylindrical cathode for controlling the angular portion of illuminated area on a conical target. The action is similar to that of a grid side-rod in shadowing a certain arc of anode circumference from the space current. The vane splits the beam of electrons and leaves sharp-edged traces on the target; the more negative (less positive) the vane, the wider the beam is split, and the larger the sector of dark area on the target. A short helix surrounds and is welded to the cathode of the eye. Its function is to limit the target current. The wire of the helix causes some disturbance in the electrostatic field and introduces in the beam velocity components of varying amounts parallel to the vane. The presence of the helix results in some loss of beam sharpness, but the loss is too slight to notice without close examination. The helix, which is at cathode potential, gives a diffused spiral pattern on the target. The spiral-beam trace is broad enough so that adjacent turns overlap and merge to approximate the appearance of uniform illumination.

D. Beam Power Tube

The beam power tube,[8] in contrast to other power tubes having reversed control-grid and screen-grid windings, uses grids of equal pitch wound in the same sense and assembled so that the turns on the two grids are in alignment. The beams formed by the control-grid wires pass almost entirely between the screen-grid turns. The screen current is materially reduced and hence

[8] Loc. cit.

the voltage on the screen may be raised above that of a conventional tube of similar dimensions without exceeding the safe screen-grid operating temperature. The cathode current is increased by the higher screen voltage and the anode current is increased not only by the higher screen voltage but also because less of the cathode current goes to the screen. As a result, the tube operates at higher efficiency and power output.

Suppression of secondary electrons from anode to screen grid, when the anode voltage is below that of the screen, is accomplished by space charge instead of by the conventional suppressor grid. The high electron density and moderately large spacing between the screen and anode lower the potential in this region sufficiently to prevent all but high-velocity secondary electrons from reaching the grid. The design of tube is complicated by the presence of the grid side-rods which give a restricted angle or sector-shaped beam. Small plates at the sides connected to the cathode, and the dimensioning of the tube electrodes as a whole, are the means employed in approximating uniformity of space potential throughout the angle of the sector.

E. Low-Noise, Low-Screen-Current Tube

Experimental results by S. W. Seeley and W. S. Barden and a theoretical analysis by B. J. Thompson and D. O. North led to the discovery that as the screen current in a tube is decreased the signal-to-noise ratio is improved, not alone as a result of increased transconductance but in consequence of a decrease in the noise itself. The beam power tube is an example of reduction of current to a screen grid by means of grid alignment. The screen grid in the Philips EF-9 actually consists of two grids with their turns in alignment, the outer one at a positive potential and having current to it reduced by an inner grid connected to cathode. The extra "screen-shadowing" grid lessens the need for alignment with the control grid and permits it to be a variable-pitch type or to have a different pitch from the screen grid.

F. Anodes of Small Dimensions

Tube types RCA-6K8 and RCA-1851 use flat (rectangular section) cathodes and flat control grids. The cathode of either type if positioned by itself along the axis of a cylindrical anode to form a diode, will emit electrons at all angles and reach the full 360° of anode circumference. In the above types, the anode consists of two small flat plates, one opposite each wide cathode face. That the beam can be confined to these small plates is due mainly to the moderately large side-rods of the flat control grid. These rods are more effective in narrowing the beams from a flat cathode than from a cylindrical type. Other electrodes in the tube at cathode potential, including the metal shell, tend to narrow the beam.

OTHER APPLICATIONS OF BEAMS

The preceding section deals with existing applications of beams. This section is concerned with applications not at present commercial. There are many possible ways of using beams which have been proposed and tried, but I am limiting the applications in this final section to tubes which I have made and tested. The data are on actual tubes, which, while indicating what can be done, may be far from all that can be realized in the types described.

A. Power Tubes with Beam-Forming Cathodes

The control grid of power tubes must be driven positive in many instances in order to deliver the high anode current at low anode voltage needed for efficient operation and large power output. The positive grid receives part of the cathode current which causes a power loss in the grid-input circuit and grid heating. Most of the grid current in certain structures comes from an area of the cathode directly under the grid wires. H. C. Thompson, in recognition of this fact, devised a cathode with alternate electron-emitting and non-emitting areas. The grid wires were directly over the non-emitting areas, and the emitting areas were between the grid wires. By this means beams of electrons emitted from the cathode passed between the grid wires without bombarding the positive grid. The grid current was reduced to a small value. An actual triode with these features is shown in the paper by H. C. Thompson.[4] It has a cylindrical anode, helical grid, and a cylindrical cathode with oxide coating in the form of a helical band. The cathode resembles a barber pole. The grid side-rods are shielded from the beam current by small rods which lie along the cathode.

A different arrangement for providing a beam-forming cathode is shown in Fig. 13a. The cathode is a fluted column with eight grooves.[12] It was made by milling grooves lengthwise in a thick-walled, 0.100"-outside-diameter, nickel tube. The cathode was completely oxide-sprayed and afterwards scraped so as to leave oxide only in the grooves. In effect, an inlaid oxide-coating was obtained. The grid is of the longitudinal type, and consists of eight straight wires equally spaced about the cathode. Each wire is opposite an uncoated stripe on the cathode. The anode is a cylinder inside of which there is a stack of spaced washers for secondary-electron suppression from the anode. Suppression is necessary if the grid at any time operates at a more positive voltage than the anode. The following data show an extreme condition of operation: anode at 10 volts, grid at 60 volts, anode current of 114 milliam-

---

[4] Loc. cit.

[12] Ref. 25-12.

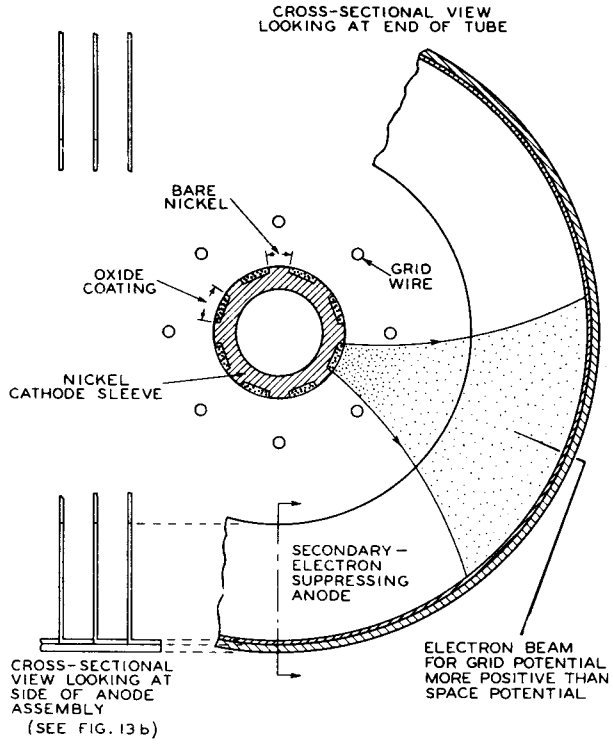


Fig. 13a - "Currentless grid" beam triode with columnar cathode having oxide-coated flutes.

peres, and grid current of 9 milliamperes. The grid receives only 8 per cent of the anode current when its voltage is six times as positive as the anode. When the grid and anode are both at 60 volts, their currents are 3.1 and 181 milliamperes, respectively. It will be noted that the grid current is less than 2 per cent of the anode current.[13]

Although the data just presented are for a tube of small power, it is of interest to consider the ultimate size of tube of this general type that may be built. The oxide-coated cathode has received a bad reputation for use in power tubes of large size and at high voltages. This reputation is to a large extent due to unsuccessful attempts to design large tubes without the application of beam principles. I am convinced that by proper use of beams it will be possible to extend the voltage and power rating of tubes with oxide-coated cathodes far in excess of present limits. The cathode-ray tube is an example of the successful application of the oxide-coated cathode at high voltage and high electron-emission density and offers encouragement in the power-tube field. Grid emission due to grid heating is one of the important limitations in power-tube design. The absence of grid heating by electron bombardment in beam tubes, is an important step in avoiding grid emission.

[13] Ref. 25-13.

Fig. 13b shows a photograph of the cathode-grid assembly and a photograph of the anode assembly for a triode of moderately large size designed by me three years ago. The cathode has a diameter of 0.800" and 30 oxide-coated stripes 2-1/4" long. This tube, tested with 150 volts

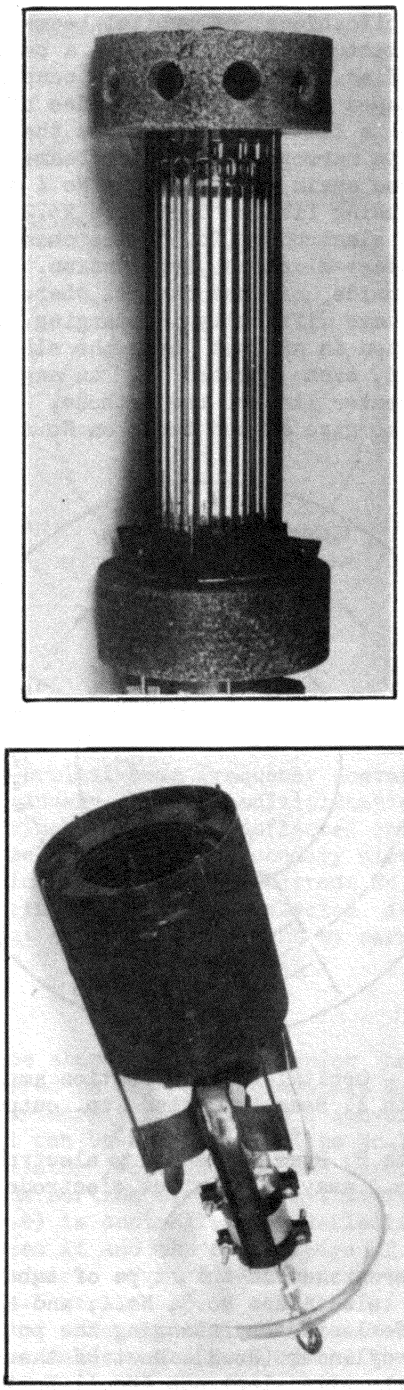


Fig. 13b - Top: Grid-cathode assembly of large beam power tube with oxide-coated flutes in cathode sleeve. Bottom: Anode assembly of same tube.

on the grid and anode, had currents of approximately 0.05 ampere and 2.4 amperes to the grid and anode, respectively. Lack of facilities prevented proper exhausting or complete testing of this tube.

B. Beam-Deflection Amplifiers

Some applications of orbital beams will conclude the lecture. Fig. 14 shows a beam-deflection amplifier. Electrode No.1, consisting of channel-shaped pieces at the sides of and connected to a flat cathode, causes the electrons to emerge in narrow beams. These beams widen but are narrowed again at electrode No.4 due to the radial focusing field of cylinders No.2 and No.3. The output electrodes (No.5) are channel-shaped for secondary-electron suppression. The beam strikes inside of the channel where secondary electrons have difficulty in emerging. The beams are deflected in and out over the slotted electrodes No.4, each located at an angle of 150° from the center line of the cathode. This angle was found to give a good focus on No.4.

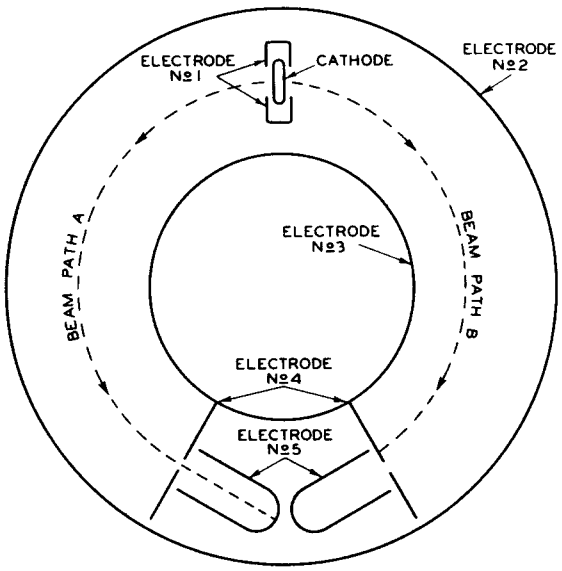


Fig. 14 - Orbital-beam-deflection amplifier.
Path A: Beam deflected to output electrode No.5.
Path B: Beam deflected to electrode No.4 away from output electrode.

Data were taken on this type of tube with 250 volts on electrodes No.3, No.4, and No.5. The beam was deflected by changing the potential on the outer cylinder (No.2). Most of the beam goes through the slit electrode (No.4) to the output electrode (No.5) when there is 60 volts on cylinder No.2. The output electrode (No.5) is made in two pieces so as to permit separate measurement of the beams in both halves of the tube. This arrangement was necessary because of dissymmetry of the beams, which was to be expected, since no special care was taken to make the tube accurate mechanically. Data show that with 4 milliamperes per beam, with the beam about half over one edge of the slit in electrode No.4 (2 milliamperes to No.4 and 2 milliamperes output current to No.5), the transconductance is approximately 1000 micromhos per beam and is reasonably constant for a total output-current variation of 1.5 milliamperes. The output electrode is moderately well-shielded from the beam and has pentode characteristics with a "knee" at about 10 volts. The transconductance is only slightly different with the output electrode at 10 volts from what it is when the output electrode is at 250 volts.

C. Negative Resistance

The device of Fig. 14 can be used as a negative resistance element. In this application, the outer cylinder (No.2) is at a positive potential sufficient to deflect the beam away from the slit in electrode No.4 and onto the outer part of No.4. Electrodes No.3 and No.4 are connected together, or can be made as one electrode. When their voltage is raised, the beams are deflected inwardly. Their current also decreases as the beams are deflected over the edge of the slit in No.4 and reach No.5, which may be at a fixed voltage. The resistance is negative over this part of the characteristic, a value of the order of 7500 ohms having been observed for a cathode current of 4 milliamperes.

It is of interest to compare the above device with the orbital-beam tube shown in Fig. 18 of H. C. Thompson's paper,[4] in which an increase of potential on the central electrode No.3 causes the beam to be deflected away from it, instead of toward it, as in Fig. 14 shown here.

D. Secondary-Electron Multipliers

Fig. 15 shows a two-stage, electrostatic-type, electron multiplier in which the input voltage is applied to grid No.1 and the output is taken from electrode No.6, which is in the form of a mesh or grid. Electrodes No.4 and No.5 are sensitized as secondary emitters. The beam first strikes No.4, which is at 100 volts. Secondary electrons from it are attracted toward No.6, at 300 volts, and No.5, at 200 volts. Most of them go between the openings of the mesh of No.6 and bombard the upper part of No.5. Secondaries from it are attracted to No.6, which is at the highest potential in the tube, and receives most of them. Since the output electrode is a grid, the electrons circulate back and forth between the wires before reaching them. This circulation of the electrons is undesirable at high frequencies and the structure would have to be modified for high-frequency use. The orbital-beam tube lends itself readily to a single-stage multiplier design, an example

[4] Loc. cit.

of which is shown in Fig. 16.

Tests were made on the tube of Fig. 15 as a double- and single-stage multiplier. For double-stage use, the secondary surfaces are at 100 and 200 volts. For single-stage use, electrodes No.5 and No.6 are connected together as an output electrode and the secondary cathode is at 200 volts. Actually, the tube is not designed for single-stage use, but the data are approximately the same as for the single-stage design of Fig. 16. The tube gives high transconductance because of secondary multiplication without resorting to especially close spacing between cathode and control grid, as in the RCA-1851. Data for a cathode current of 2.5 milliamperes in all cases are shown in the following table.

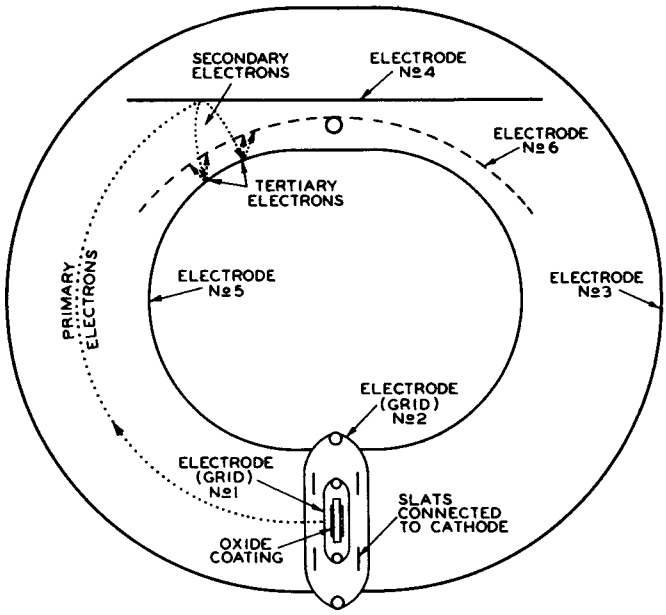


Fig. 15 - Two-stage, orbital-beam, electrostatic-type, electron multiplier.

Stages of Multiplication	Output Current Milliamperes	Transconductance Micromhos
None	2.5	1750
1	10.6	7800
2	19.5	16000

E. Converters

In conventional types of converter or mixer tubes, the signal voltage is applied to one grid and oscillator voltage to another grid. One grid is close to the cathode and controls the current from it. The other grid, which is surrounded by screen grids, varies the portion of this current that reaches the anode. The part that does not go through to the anode is reflected back toward the cathode, where it is a source of trouble, and causes electron coupling between the oscillator and signal circuits.

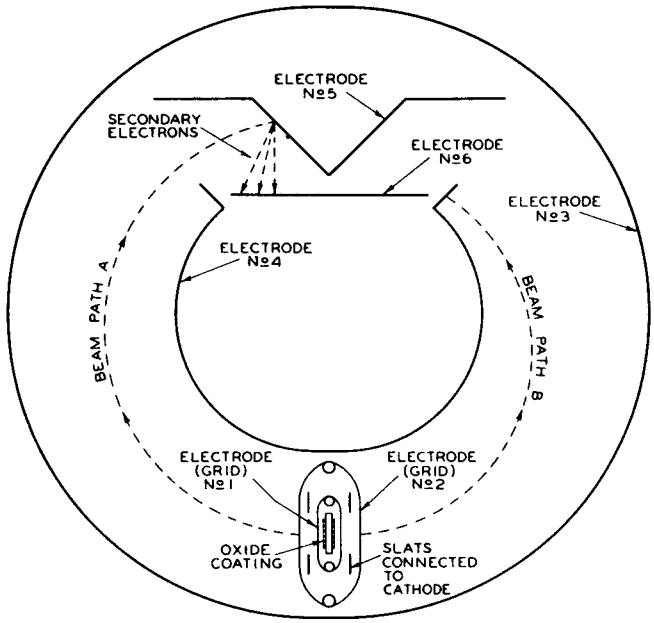


Fig. 16 - Orbital-beam frequency converter with secondary-electron multiplication.
Path A: Primary electrons deflected into multiplier section causing secondary electrons which reach output electrode No.6.
Path B: Primary electrons deflected away from output section and caught by ears of electrode No.4.

The tube shown in Fig. 16, which has already been referred to as a single-stage multiplier, was made primarily for use as a converter tube. The signal can be applied to the No.1 grid and the oscillator voltage to the outer cylinder (No.3). The portion of the beam that reaches the anode (No.6) is controlled by deflection of the beam between it and the folded edge of electrode No.4. The part of the beam that goes over the edge of No.4 strikes No.5, which is a secondary cathode, and the multiplied electrons from it reach the anode. The rest of the beam strikes the edge of No.4 and cannot return to the cathode region.

The aim of this tube design is not only to avoid electrons reflected to the region between cathode and control grid, but also to increase the output by means of secondary multiplication.

The secondary cathode (No.5) has a secondary-emission ratio of 3.3 at 150 volts. The data are approximately as follows:

Conversion Conductance	1500 micromhos
Maximum Voltage on any electrode	250 volts
Output-Electrode Voltage	250 volts
Output-Electrode Current	3.5 milliamperes
Total Cathode Current	3.5 milliamperes
Peak Oscillator Voltage (applied to beam-deflection electrode)	20 volts
Average Current to beam-deflection electrode	<10 microamperes

Lecture 26

THE DESIGN AND PERFORMANCE OF RECTIFIERS

A. P. Kauzmann

INTRODUCTION

This lecture will deal with rectifying tubes as used in ordinary receivers. The performance of the rectifying tube is to a large extent affected by the type of filter circuit following it. Filter circuits may be classified into two groups, i.e., choke-input filters and condenser-input filters. The former are comparatively simple to analyze quantitatively, but the latter are greatly complicated mainly by the fact that the diode is not a linear resistance, that is, the current through the diode is not proportional to the voltage across it but to the voltage raised to the three-halves power. Further complications are due to the short periods of current flow through the diode and the resulting transient phenomena which occur not only on starting but also in the steady state. Exact mathematical treatment of diode performance is too tedious for practical application. O. H. Schade has explained in Lecture 19 how to obtain by empirical methods a practical solution for the output voltage, peak current, and dissipation of a diode operating into a condenser filter by replacing the diode with a fictitious ideal diode (one which has no voltage drop for any magnitude of current passing through it) and an equivalent series resistance which has different values for a given d-c output current depending on whether we are observing the d-c output, the peak current, or the heating of the diode. Application of Schade's method to a typical rectifier circuit will be discussed later in the lecture.

A condenser-input filter usually has a large condenser immediately following the rectifier tube, and then a network of chokes and condensers to reduce the ripple voltage across the load, as shown in Fig. 1a. In so far as the tube performance is concerned, this network simplifies to a condenser (C_1) and load resistor (R_L), as shown in Fig. 1b. R_L is equal to the sum of the ohmic resistances of the series chokes and the output resistor R_o. The value of C_1 is the value of the first capacitance only.

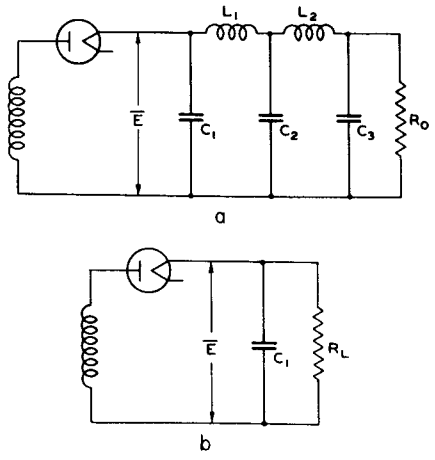


Fig. 1

Similarly, the choke-input circuit of Fig. 2a can be converted to the equivalent circuit of Fig. 2b.

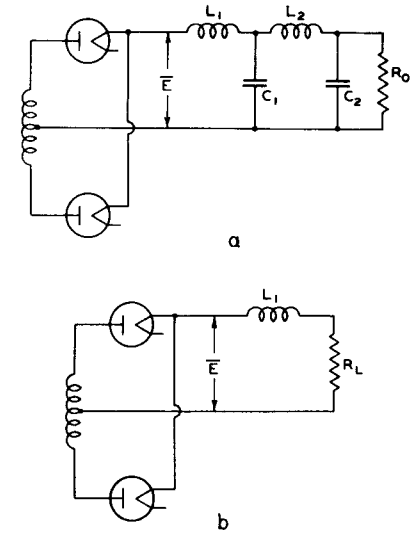


Fig. 2

Here, R_L is again equal to the sum of the ohmic resistances of the chokes, L_1, L_2, and the output load R_o.

The voltage-doubling circuit as discussed in this lecture is really two half-wave rectifiers in series. This fact is readily apparent from Fig. 3 if a connection were made, as shown by the dotted line, between points 1 and 2 which are both approximately at ground potential. The main difference is that the ripple voltage across R_L has a frequency of twice the supply frequency, whereas for the half-wave circuit the ripple is equal to the fundamental supply frequency. The advantage of the voltage-doubling circuit is that it gives twice the output voltage of a half-wave circuit. Its disadvantage is that it does not have both cathodes at the same potential and as a consequence, better heater-cathode insulation must be provided for at least one cathode. Compared with the full-wave rectifier, the voltage-doubling circuit has the economic disadvantage of requiring twice the filter capacitance to give the same ripple voltage.

THE DIODE CHARACTERISTIC

We can predict the performance of a diode be-

cause we know how to predict and extrapolate the diode characteristic provided emission saturation does not occur. In all diodes with properly designed oxide-coated cathodes, the latter requirement is fulfilled. In the case of tungsten or thoriated-tungsten filaments, there may or may not be saturation.

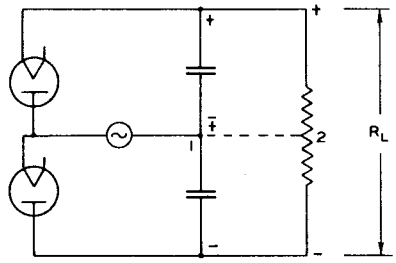


Fig. 3

For a plane-parallel cathode and anode, the current is given as

$$I = \frac{2.334\ A}{X^2}\ e_d^{\frac{3}{2}} \quad \text{microamperes}$$

where

A = area of the cathode in sq cm
X = distance from cathode to anode in cm
e_d = voltage between cathode and anode in volts (corrected for contact potential)

For cylindrical electrodes, the cathode being inside and concentric with the anode,

$$I = \frac{14.66\ L}{r\beta^2}\ e_d^{\frac{3}{2}} \quad \text{microamperes}$$

where

L = length of cathode in cm
r = inside radius of anode in cm
β^2 = $f(r/r_o)$ where r_o = radius of cathode
e_d = voltage between cathode and anode in volts (corrected for contact potential)

The values of β^2 and β have been determined by I. Langmuir.[1] However, from the curves of Fig. 4 the value of the constant $(14.66\ L)/r\beta^2$ is determinable as a function of the anode and cathode diameters, the length L being assumed as 1 millimeter. The ordinates of this figure labelled "perveance in microamperes per volt$^{3/2}$ per millimeter" must be multiplied by the sprayed length of the cathode in millimeters to evaluate the above constant.

Both equations can be expressed for a given structure as

$$I = \kappa\ e_d^{\frac{3}{2}}$$

where,

κ = perveance of the diode

$$= \frac{2.344\ A}{X^2} \quad \text{(for parallel-plane case)}$$

$$= \frac{14.66\ L}{r\beta^2} \quad \text{(for cylindrical case)}$$

If we express this equation in terms of logarithms, we have

$$\log I = \log \kappa + \frac{3}{2} \log e_d$$

This is a linear equation in log I and log e_d and will, therefore, plot as a straight line on log-log paper (see Fig. 5) and will have a slope of 3/2 (with e_d as abscissa). Note also that the perveance may be read directly from this log-log plot where it is numerically equal to the plate current at e_d = 1 volt. For some of the types, it will be necessary to extend the curves in Fig. 5 in order to read the perveance.

The above equations assume no initial velocities and zero contact potentials. The sum of both effects in terms of voltage is never greater than about ± 1 volt. This is small enough to be negligible in most cases where the applied voltages are 50 volts rms or more, but care must be taken if the characteristic of a tube is plotted from observed data. A practical method is to place a line having the slope of 3/2 on log-log graph paper where the ordinate represents plate current and the abscissa represents diode plate voltage. This line is drawn to go through the highest observed plate current—diode plate voltage point. The voltage should be at least 30 volts in order to make negligible the effect of contact potential. If the observed values are plotted point by point on the above log-log paper, a curved line usually results; but by trial a constant voltage, representing the contact potential, E_o, can be found which when added to or subtracted from the observed curve will give a straight line having the 3/2-power relationship. In Fig. 5 are shown such curves for some of the more common RCA rectifiers. Here e_d = voltage between anode and cathode corrected for contact potential; E_o = contact potential (assumed to be + 1 volt); e_p = applied external voltage between anode and cathode.

In filament-type tubes, such as the 80, the perveance may be computed by a method used by Yuziro Kusunose.[2] The effective area, A, of each

[1] I. Langmuir, "Electrical Discharges in Gases," Part II — Fundamental Phenomena in Electrical Discharges, Rev. of Mod. Phys., Vol. 3, No.2, pp. 247-248, Figs. 43 and 44; April, 1931.

[2] Yuziro Kusunose, "Calculation of Characteristics and Design of Triodes," Proc. I.R.E., Vol. 17, No.10, pp. 1706-1749; October, 1929.

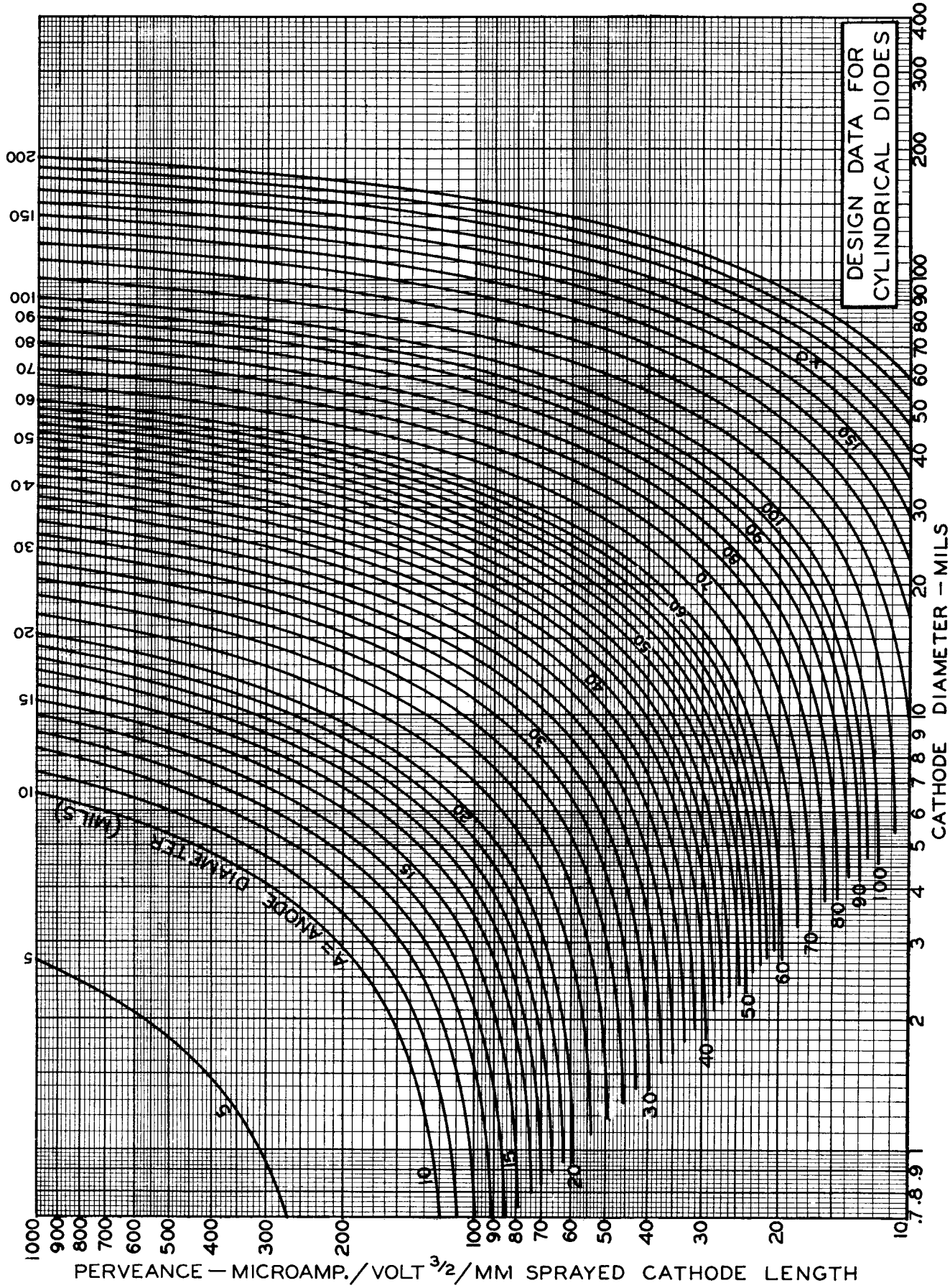

DESIGN DATA FOR CYLINDRICAL DIODES
CATHODE DIAMETER — MILS
PERVEANCE — MICROAMP./VOLT 3/2/MM SPRAYED CATHODE LENGTH
ANODE DIAMETERS (MILS)

Fig. 4

filament is assumed to be the area of a strip running around the filament and having the width 2X, where X is the distance from filament to anode (Fig. 6). If the anode is on both sides of the filament (as is usually the case), this area must be doubled.

METHOD OF OBTAINING GENERALIZED CURVES FOR CONDENSER-INPUT OPERATION OF RECTIFIERS

Diodes with various perveances placed in a condenser-input rectifying circuit, have effects on the output voltage, peak current, and plate dissipation which do not lend themselves to a practical mathematical solution. The difficulty is overcome by use of empirical curve data which have been generalized so that all circuit constants, the input and output voltages, and the peak, average, and rms currents appear as dimensionless parameters. Typical parameters are $\% \bar{E}/\hat{E}_{max}$, $\omega C R_L$, $\hat{I}_p/\bar{I}_p$, and $\% R_S/R_L$.

We will take as an example a type 80 tube in

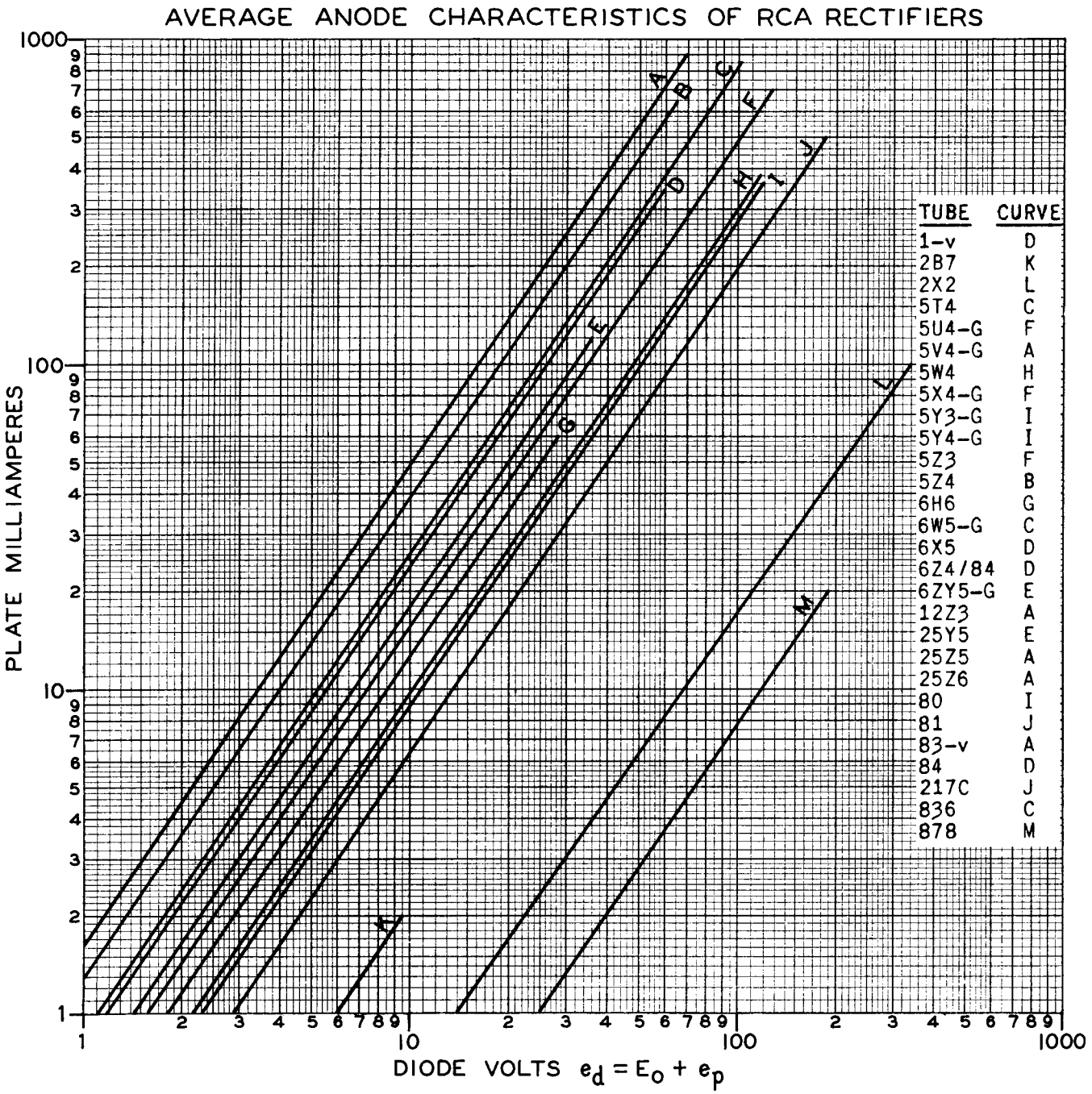


Fig. 5

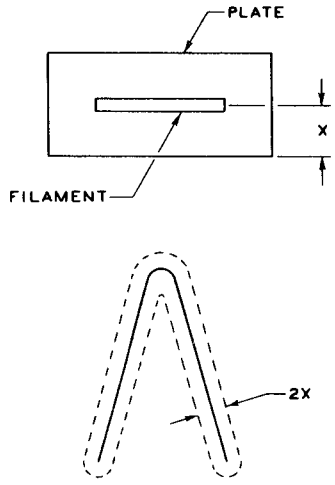


Fig. 6

a typical full-wave circuit (Fig. 7). Added to this circuit, as shown in Fig. 8, in one of the plate leads is an oscillograph for measuring the peak current, $\hat{I}_p$, an a-c meter for reading the effective heating current, $|I_p|$, a d-c meter for reading the direct or average current, $\bar{I}_p$, and a decade box for determining the equivalent resistances of the diode, $\hat{r}_d$, $|r_d|$, and $\bar{r}_d$. Across a single plate of the type 80 are placed some twenty type 83-v diode plates with a switch in each plate lead so that any number may be introduced in the circuit.

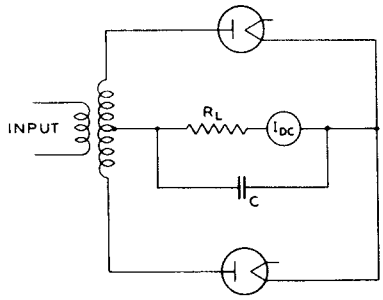


Fig. 7

Starting with zero resistance in the decade box R, and with all the switches in the plate leads of the 83-v's open, the circuit will perform in its normal manner. If we watch only the Idc meter as we close one switch after another of the 83-v's, we shall observe that the $\bar{I}_p$ readings increase due to more efficient rectification. After a certain number of switches have been closed, no further increase is observable no matter how many more diodes are switched in. This observation means that we have approached the condition of having an ideal diode in the circuit, i.e., one which has zero resistance in the conducting part of its cycle and infinite resistance in the non-conducting part of its cycle. If, now, the decade-box resistance R is adjusted to some value, $\bar{r}$, the I_{dc} meter will again read its original value. In other words, the type 80 diode has been replaced by a perfect diode and linear resistance of value $\bar{r}$. However, the I_{rms} meter and the oscillograph will not read the same values as originally. The same procedure is repeated for $\hat{I}_p$ and $|I_p|$, and values of resistance $\hat{r}$ and $|r|$ noted.

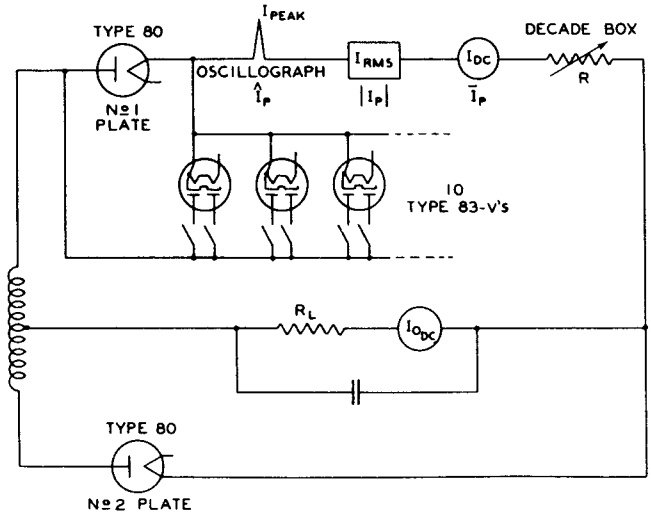


Fig. 8

Hence, it is evident that so long as we are interested in only one of the three dependent currents, we can replace the diode by an equivalent resistance (and of course an ideal diode). Fortunately, from the data taken, the peak diode resistance $\hat{r}_d$ always can be expressed in terms of the peak current $\hat{i}_d$ and peak tube drop $\hat{e}_d$ by the equation

$$\hat{r}_d = \frac{\hat{e}_d}{\hat{i}_d}$$

where the peak tube drop (diode peak voltage) is corrected for contact potential, E_o, as shown in Lecture 19, Fig. 17. This equation is very important since it is the only connecting link between the graphs (where only ratios and percentages enter as parameters in both dependent and independent functions) and the plate voltage vs plate-current characteristics of the tube. The relationships,

$$\frac{\hat{e}_d}{\hat{i}_p} = \hat{r}_d = 0.88\ \bar{r}_d = 0.935\ |r_d|$$

in Fig. 17 of Lecture 19 are justified not only because calculations show them to be reasonably accurate for current surges having sinusoidal,

triangular, and semicircular wave shapes. For example, in the equation $\hat{r}_d = C\,\bar{r}_d$, determination of the constant C for the different wave shapes gave values between 0.84 and 0.89. The value of 0.88 shown in the relationships was derived for a sine wave and is applicable in most cases to wave shapes usually encountered.

Obviously, just replacing the actual diode by an ideal diode and an equivalent series resistance does not take into account the effect of external resistance present in the supply line and transformers, or of the series resistance purposely inserted in some close-spaced rectifier circuits to prevent sputter on switching. This external resistance is purely ohmic, and is added to the equivalent diode resistance directly in each case (see Fig. 17, Lecture 19). For a transformer, the external resistance in a half-wave circuit is

$$R_S = a^2 R_{prim} + R_{sec}$$

where a = transformation ratio

$$= \frac{E_{sec}}{E_{prim}}, \text{ read at no load.}$$

In the case of a full-wave rectifying transformer, R_{sec} and E_{sec} are for only half of full secondary windings.

PLATE DISSIPATION

The total heat dissipated by the plate of a rectifier consists of that radiated by the cathode (which for practical purposes may be considered as entirely surrounded by the plate) and that generated at the plate as a result of the plate current through the diode. The heat radiated by the cathode is easily determined by the equation

$$W_{fil} = E_f \times I_f$$

where

W_{fil} = filament or heater power in watts
E_f = filament or heater volts
I_f = filament or heater amperes

The plate-current dissipation can be experimentally determined by observing the envelope temperature under the desired operating condition and duplicating this temperature with d-c voltage applied to the plate. This method is inconvenient and does not allow of prediction of the dissipation of a tube before it is made.

Another method, which is extremely tedious, is to take two simultaneous oscillographs of the actual tube voltage drop and the conducting current during the conduction period. The simultaneous ordinates of voltage and current are multiplied together to give instantaneous watts output vs time. The area under the resultant curve is then integrated over a full period to obtain the average dissipation. This method was successfully used to check within 5 per cent the results of the following two methods.

By use of the upper chart of Fig. 21 in Lecture 19, we can compute the dissipation. A typical example will be given later in the lecture. Considerable error is possible unless great care is used in interpolating the value for % $|R_S|$ of R_L. By this method, the plate dissipation is

$$\text{Watts}_{\text{per plate}} = |I_p|^2 \times |r_d|$$

The last method, and the one most convenient to use, is the result of assuming that the current pulse is sinusoidal in shape. Mathematically this is not true but practically the approximation is close, the error being probably less than 5 per cent based on comparison with the oscillograph method. Under this assumption, the plate dissipation is

$$\text{Watts}_{\text{per plate}} = 0.84\,\hat{e}_d\,\bar{I}_{p\,\text{per plate}}$$

where

$\hat{e}_d$ = diode peak voltage drop during conducting period.

The value of $\hat{e}_d$ is obtained from the lower chart of Fig. 21 in Lecture 19, or where applicable from Fig. 9 by first obtaining the peak current, and then from the proper diode characteristic in Fig. 5 obtaining the corresponding $\hat{e}_d$.

SIMPLIFICATIONS WHEN INPUT CONDENSER IS LARGE

1. Peak Currents

If we inspect Figs. 18, 19, and 20 in Lecture 19, which show the relationship between output voltage and ωCR_L, we notice that as ωCR_L becomes large all the curves of $\bar{R}_S$ (representing various diodes) flatten. In other words, if we increase the value of the input condenser C, the output voltage becomes constant, and since R_L is fixed, the output current, $\bar{I}_p = \bar{E}/R_L$ becomes constant. Similarly, from the lower chart of Fig. 21 (Lecture 19), we notice that the peak-current-to-average-current ratios also become constant as we increase the value of C. Stated in other words, if we choose a filter condenser large enough, the peak current, output voltage, and output current no longer are a function of the condenser size. The data of these curves have been replotted in Fig. 9 where the current ratio $\hat{I}_p/\bar{I}_p$ per plate is shown plotted against the per cent voltage ratio $\bar{E}/\bar{E}_{max}$ per plate. For most receiving-set rectifiers, a suitable value of C is 16 μf in a full-wave circuit and 32 μf in a half-wave circuit. These values are for maximum rated output of the rectifier tube.

2. Dissipation

If we continue along the above line of rea-

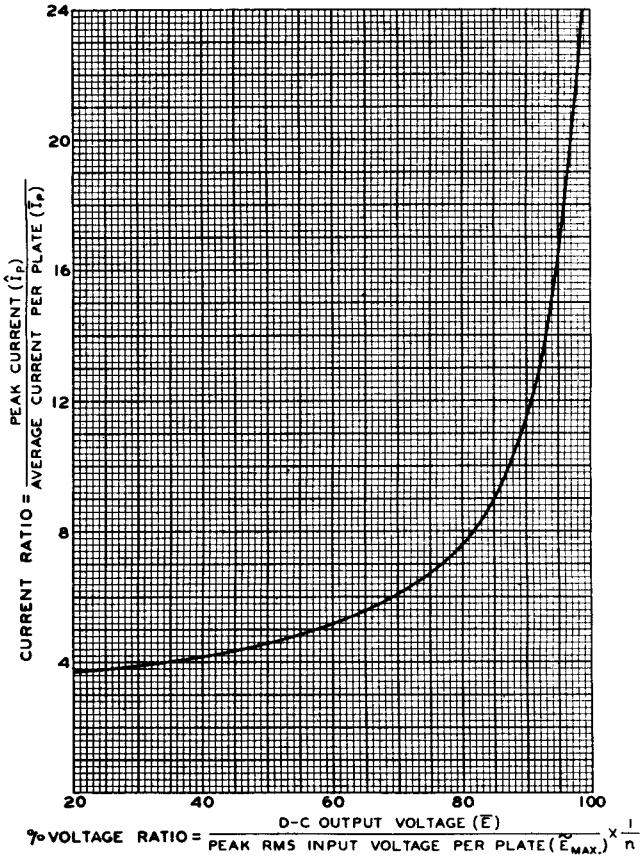


Fig. 9 - This curve is for full-wave, half-wave, and voltage-doubler rectifiers with condenser-input filters. It applies for any size condenser as long as the condenser is large enough to give maximum output voltage for the given output current and rms voltage input. For half-wave and voltage-doubler rectifiers, average load current $\bar{I}_p = \bar{I}$; for full-wave rectifiers, $\bar{I}_p = 0.5\ \bar{I}$.
n = 1 for full-wave and half-wave rectifiers.
n = 2 for voltage-doubler rectifiers.

soning and place an infinitely large condenser after the rectifier, there will be no ripple in the output, and we shall have a pure d-c output voltage $\bar{E}$. As a result, the maximum tube drop $\hat{e}_d$ (Fig. 10) must be equal to

$$\hat{e}_d = \tilde{E}_{max} - \bar{E}$$

where $\tilde{E}_{max}$ = peak input voltage per plate. We have already indicated that the plate dissipation can be expressed as

$$\text{Watts}_{\text{per plate}} = 0.84\ \hat{e}_d\ \bar{I}_p \text{ per plate}$$

Substituting for $\hat{e}_d$ the value just considered, it is evident that

$$\text{Watts}_{\text{per plate}} = 0.84\ (\tilde{E}_{max} - \bar{E})\ \bar{I}_p \text{ per plate}$$

However, from the lower chart in Fig. 21 of Lecture 19, it is apparent that $\hat{I}_p$ reaches its maximum with values of condenser not necessarily infinitely large; in fact, the $\hat{I}_p/\bar{I}_p$ ratio has reached a constant value before $\bar{E}$ reaches a constant value. The above equation is, therefore, justified so long as a sufficiently large con-

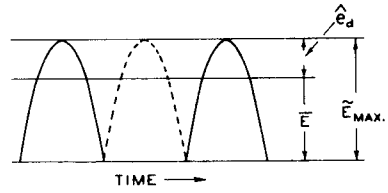


Fig. 10

denser is used. The condenser should be large enough so that any further increase in its capacitance will cause no further appreciable increase in $\bar{E}$.

CALCULATION OF PEAK CURRENT AND DISSIPATION

Assume that a type 5W4 tube is working as a full-wave rectifier with the following voltages, currents, and circuit constants:

Transformer —

Primary: Voltage = 120 volts (rms)
Frequency (f) = 60 cycles
Resistance = 1.05 ohms

Secondary: Voltage = 400 volts (rms) per plate
Resistance = 22 ohms per ½ secondary
Load Resistance (R_L) = 3500 ohms
Output Voltage ($\bar{E}$) = 402 volts
Output Current ($\bar{I}$) = 114 milliamperes
Input Condenser (C) = 4 μf

The notations used will be those indicated in Fig. 17 of Lecture 19. The external series resistance per plate, R_S, is all due to the transformer. Therefore, as we have already seen,

$$R_S = a^2 R_{prim} + R_{1/2\ sec} = \left(\frac{400}{120}\right)^2 1.05 + 22 = 10.65 + 22 = 33 \text{ ohms}$$

Other working constants needed are:

$$\% \frac{\bar{E}}{\tilde{E}_{max}} = \frac{402}{400\sqrt{2}} \times 100 = 71\%$$

(See Fig. 19 in Lecture 19)

$$\omega C R_L = (2\pi \times 60)(4 \times 10^{-6})(3500) = 5.28$$

(See Figs. 18, 19, 20 in Lecture 19)

$$n\omega C R_L = 2 \times 5.28 = 10.5$$

(See Fig. 21 in Lecture 19)

1. Peak Current Determination

With entries of $\omega CR_L = 5.28$ and $\% (\bar{E}/\tilde{E}_{max}) = 71\%$, we locate point A on Fig. 19 (Lecture 19) and obtain

$$\% \frac{\bar{R}_S}{R_I} = 11\%$$

whence

$$\bar{R}_S = 11\% \times 3500 = 385 \text{ ohms},$$

Since

$$\bar{R}_S = R_S + \bar{r}_d$$

where R_S is the external series resistance of the transformer and equals 33 ohms, we have from Fig. 17 (Lecture 19)

$$\bar{r}_d = 385 - 33 = 352 \text{ ohms}$$

$$\hat{r}_d = 0.85\, \bar{r}_d = 0.88 \times 352 = 310 \text{ ohms}$$

$$|r_d| = \hat{r}_d/0.935 = 332 \text{ ohms}$$

$$\hat{R}_S = R_S + \hat{r}_d = 33 + 310 = 343 \text{ ohms}$$

$$\% \frac{\hat{R}_S}{R_L} = \frac{343}{3500} \times 100 = 9.8\%$$

$$\% \frac{\hat{R}_S}{nR_L} = \frac{9.8\ \%}{2} = 4.9\%$$

Now we can obtain the ratio of peak current to d-c plate current from the lower chart of Fig. 21 (Lecture 19). Using as entries $n\omega CR_L = 10.5$ and $\% (\hat{R}_S/nR_L) = 4.9\ \%$, we locate point B, and obtain

$$\frac{\hat{I}_p}{\bar{I}_{p \text{ per plate}}} = 6.2$$

Therefore,

$$\hat{I}_p = 6.2 \times \frac{114}{2} = 354 \text{ milliamperes}$$

2. Total Dissipation

a) By equation:

$$\text{Watts}_{\text{per plate}} = 0.84\, \hat{e}_d\, \bar{I}_{p \text{ per plate}}$$

Since $\hat{I}_p = 354$ milliamperes, and from curve H of Fig. 5 showing the diode characteristic of the 5W4, we obtain

$$\hat{e}_d = 109 \text{ volts}$$

Therefore,

$$\text{Watts}_{\text{per plate}} = 0.84 \times 109 \times (0.114/2) = 5.2$$

and the total dissipation is the sum of the plate watts (W_p) and the filament watts (W_f) or

$$W_p + W_f = 2(5.2) + (5 \times 2) = 20.4 \text{ watts } *$$

b) By equation: $\text{Watts}_{\text{per plate}} = |I_p|^2 \times |r_d|$

Since we have found the value of $|r_d|$ to be 332 ohms, and if we add the transformer resistance $R_S = 33$ ohms, then from the equation

$$|R_S| = R_S + |r_d|$$

we have

$$|R_S| = 33 + 332 = 365 \text{ ohms}$$

and

$$\% \frac{|R_S|}{nR_L} = \frac{365}{2 \times 3500} \times 100 = 5.2\%$$

Using as entries

$$n\omega CR_L = 10.5 \text{ and } \% (|R_S|)/ nR_L = 5.2\%$$

we locate point C on the upper chart of Fig. 21 (Lecture 19) and obtain

$$\frac{|I_p|}{\bar{I}_{p \text{ per plate}}} = 2.2$$

Therefore,

$$|I_p| = 2.2 \times \frac{114}{2}$$

$$= 125.5 \text{ rms milliamperes equivalent}$$

Hence,

$$\text{Watts}_{\text{per plate}} = (0.1255)^2 \times 332 = 5.2$$

and the total dissipation is

$$W_p + W_f = 2(5.2) + 10 = 20.4 \text{ watts } *$$

DESIGN OF A RECTIFIER FOR A GIVEN OUTPUT CURRENT

The problem is to design a full-wave rectifier to operate at 400 volts rms per plate, and to supply an output current of 200 milliamperes at an output voltage of not less than 400 volts d.c. The filter is to be of the condenser-input type, and it will be assumed that the size of the condenser will be large, i.e., 30 μf or greater. This assumption greatly simplifies the work.

* This dissipation is not safe for the 5W4 because its MT8G envelope has a maximum dissipation limited to 19.7 watts.

The problem has definite steps as follows: (1) Design of cathode and determination of heater power; (2) determination of spacing between anode and cathode; and (3) choice of proper envelope based mainly on dissipation requirements. In connection with the second step, it should be noted that improper spacing may give rise to impractical manufacturing clearances, high peak currents, excessive dissipation, and too low a d-c voltage output.

1. The Cathode Design

The cathode will be assumed to be cylindrical and is indirectly heated. From the tabulation of design constants in Table I, we find that for each 19 to 25 milliamperes of d-c output current required, there must be a power input of one watt to the cathode. Choosing 20 milliamperes per watt, and remembering that we want an output current of 200 milliamperes, we find that the total power for both cathodes will be about 10 watts. If we assume a nickel cathode with an outside diameter of 0.065" and a radiation loss of 3.5 watts per square centimeter (see Table I), the length of the coated area of the cathode is determined. Allowing an uncoated length of 3.5 millimeters at each end, and assuming that the radiation from shiny nickel is one-half that of oxide-coated nickel, we have by the formula

$$\text{Effective Watts}_{\text{per sq cm}} = \frac{\text{Watts Input}_{\text{per cathode}}}{\pi d \left(L_{\text{coated}} + \frac{L_{\text{uncoated}}}{2} \right)}$$

where d and L are in centimeters,

$$3.5 = \frac{5}{\pi\ (0.065" \times 2.54)(L_{\text{coated}} + 0.35)}$$

whence

$$L_{\text{coated}} = 2.4 \text{ cm}$$

and

$$\text{Coated Area} = 1.25 \text{ sq cm}$$

The sketch in Fig. 11 shows the cathode design.

2. Anode Design and Its Effects on Peak Current, Voltage Output, and Dissipation

In Fig. 12 are shown the effects on peak current, voltage output, and total dissipation as the anode diameter is varied. From these curves the final anode design may be established. Only the determination of one point on each of the curves will be carried through to show the methods used.

Since a large input condenser for the filter was assumed in order to give maximum output voltage, we may use the simplifications previously described for our calculations. The voltage output and peak tube drop are computed from

$$\bar{E} = \tilde{E}_{max} - \hat{e}_d$$

and the peak currents are established from the curve of Fig. 9 showing ratio $\hat{I}_p / \bar{I}_p$ per plate versus the % voltage ratio $\bar{E} / \tilde{E}_{max}$ per plate.

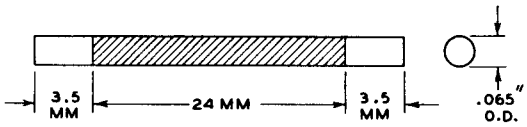


Fig. 11

Let us arbitrarily assume that $\bar{E} = 0.75\ \tilde{E}_{max}$. The steps necessary to obtain points for Fig. 12 are to determine: (1) the output voltage $\bar{E}$; (2) the peak current $\hat{I}_p$; and (3) the total dissipation W_{total}. From the equation above, $\bar{E}$ and the peak voltage drop are immediately obtainable.

$$\bar{E} = 0.75 \times 400\sqrt{2} = 424 \text{ volts d.c.}$$

$$\hat{e}_d = 400\sqrt{2} - 424 = 141 \text{ volts d.c.}$$

From Fig. 9, for $\bar{E} / \tilde{E}_{max} = 0.75$, we obtain

$$\hat{I}_p / \bar{I}_{p \text{ per plate}} = 6.65$$

and since the $\bar{I}_p$ for one plate is half the total output, the peak current is

$$\hat{I}_p = \frac{200}{2} \times 6.65 = 665 \text{ milliamperes}$$

Peak currents, for convenience in comparing with other tubes, are expressed in Fig. 12 in terms of milliamperes per square centimeter of cathode surface. Since the coated area of the cathode is 1.25 square centimeters,

$$\hat{I}_p/\text{sq cm} = \frac{665}{1.25} = 585 \text{ milliamperes}$$

The plate dissipation can now be computed. By the equation

$$\text{Watts}_{\text{per plate}} = 0.84\ \hat{e}_d\ \bar{I}_{p \text{ per plate}}$$

we have

$$\text{Watts}_{\text{per plate}} = 0.84 \times 141 \times 0.100 = 11.84$$

The total watts dissipation for two plates and two cathodes is

$$\text{Watts}_{\text{total}} = 2(11.84) + 10 = 33.7$$

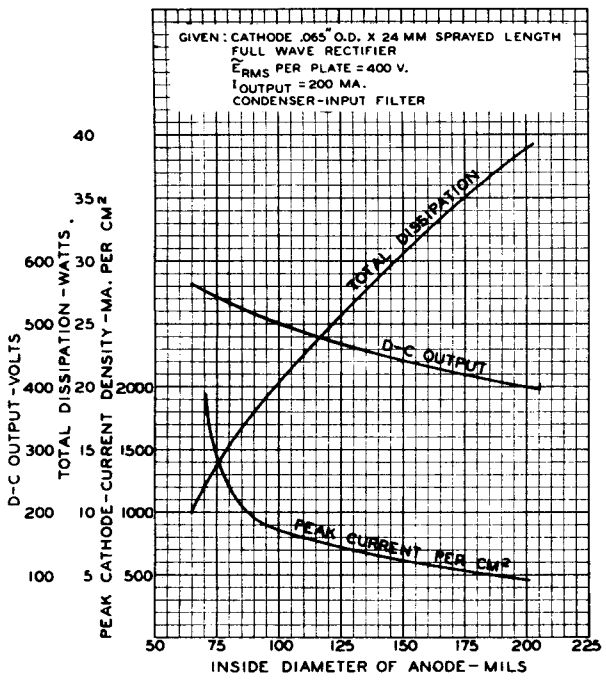


Fig. 12 - Effect of varying anode diameter in rectifier design.

Note that all data have been obtained except the diameter of the anode. This can now be determined to satisfy the assumption that $\bar{E} = 0.75\ \tilde{E}_{max}$, in view of the fact that we now know the peak voltage drop $\hat{e}_d$ and the corresponding peak current $\hat{I}_p$. As the first step in determining the anode diameter, we must establish the perveance, κ, of the diode by means of the equation

$$\hat{I}_p = \kappa\, \hat{e}_d^{\frac{3}{2}}$$

It is a simple slide-rule calculation to determine κ if it is remembered that $\hat{e}_d^{3/2} = \hat{e}_d \sqrt{\hat{e}_d}$. Therefore,

$$\kappa = \frac{665}{141\sqrt{141}} = 0.397\ \frac{\text{ma.}}{\text{volts}^{\frac{3}{2}}} =$$

$$397\ \frac{\text{microamp.}}{\text{volts}^{\frac{3}{2}}}$$

In Fig. 4 the perveance is given in microamperes per volt$^{3/2}$ for a cathode 1 millimeter long. Therefore, the perveance per millimeter length for our problem diode with its cathode having a coated length of 24 millimeters is

$$\kappa_{\text{per mm}} = \frac{397}{24} = 16.5\ \text{microamp./volt}^{\frac{3}{2}}\text{/mm}$$

With entries of cathode diameter (65 mils) and perveance per millimeter (16.5 microamp./volt$^{3/2}$/mm), we locate point D in Fig. 4 from which is obtained the inside diameter of 167 mils for the anode.

We have now obtained data for the d-c output voltage, peak current density, and total dissipation corresponding to our anode diameter of 167 mils. Each of these values corresponds to one point for each of the three curves of Fig. 12. The actual curves for Fig. 12 were obtained by repeating the above computations starting with various assumed values of output voltage $\bar{E}$, as the independent variable.

Now from the data presented in Fig. 12, we are able to weigh one factor against another in determining the final inside diameter of the anode. These factors as previously explained are d-c output voltage, peak plate current, manufacturing clearances, and total dissipation. Firstly, it is apparent from the output voltage curve that in order to get a d-c output voltage of 400 volts d.c., the anode inside diameter must not exceed 190 mils. Secondly, from the curve of peak current per square centimeter, the minimum anode diameter can be set by the following considerations. It might appear on first thought that this peak current per square centimeter should be limited solely by the emission saturation current which for a cathode dissipation of 3.5 watts per square centimeter will be of the order of 12 amperes per square centimeter. However, from experience we know that such heavy current densities actually produce hot spots on the cathode with sufficient heat to vaporize the cathode and cause arc-back. For a practical value, therefore, past experience is called upon, and using the averages of tubes made in the past (see Table I) a value of 800 milliamperes per square centimeter must be chosen as a maximum. From Fig. 12, this fixes the minimum anode inside diameter at 110 mils. Thirdly, the clearance must be kept to not less than 20 mils for maintaining reasonable manufacturing tolerances. On this basis, the minimum inside diameter would be 105 mils (65 + 20 + 20), and so we see that the 110-mil value is satisfactory for clearance from a manufacturing standpoint. Fourthly, from the dissipation curve, the total dissipation for the smallest permissible anode diameter (which was determined by the peak current per sq cm) is 22.5 watts for two plates and two cathodes.

The final design, therefore, will have the following specifications:

Cathode diameter	= 65 mils
Anode diameter	= 110 mils I.D.
Peak plate current	= 800 ma./cm^2
	= 800 x 1.25 cm^2 = 1000 ma.
Total dissipation	= 22.5 watts
Max. possible voltage output	= 490 volts

3. Choice of Envelope

Having determined the total dissipation of the rectifier tube, we are ready to consider the size of the envelope. For a total dissipation of 22.5 watts, we find from Table I that the envelope in metal must be an MT10A, or in glass it must be an ST14.

CHOKE-INPUT FILTERS

No serious difficulties are involved in predicting diode performance in choke-input filter circuits. In Fig. 13 is shown a typical full-wave, choke-input circuit and the resulting output voltages and currents.

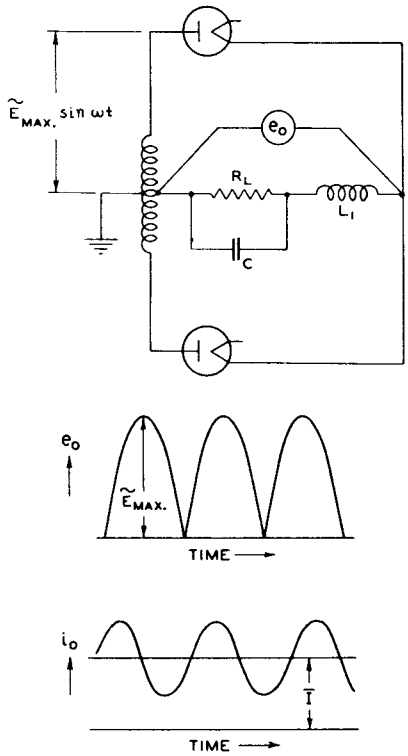


Fig. 13 - Typical full-wave, choke-input circuit.

The output voltage e_o is for all practical purposes a series of half-sine waves which may be expressed by a Fourier series as

$$e_o = \tilde{e}_{max} [0.637 + 0.425 \cos 2\omega t - 0.085 \cos 4\omega t + 0.364 \cos 6\omega t - .....]$$

From this immediately can be determined the d-c output voltage,

$$\bar{E} = 0.637\, \tilde{e}_{max} = 0.637 \sqrt{2}\, |\tilde{E}| = 0.90\, |\tilde{E}|$$

These relations for $\bar{E}$ represent the maximum output voltage available at very small output currents.

The output current consists of a d-c component, $\bar{I} = \bar{E}/R_o$, where R_o is the sum of the load resistance R_L and the ohmic resistance of the choke L. To this component are added the higher harmonic components of which the term in cos 2ωt is by far the largest. The average drop through the diode is, therefore, equal to the tube drop corresponding to the average current $\bar{I}$ flowing through it. The output voltage for any output current is thereby closely approximated by subtracting from the maximum possible $\bar{E}$ the corresponding value on the voltage-current characteristic of the diode as shown in Fig. 14.

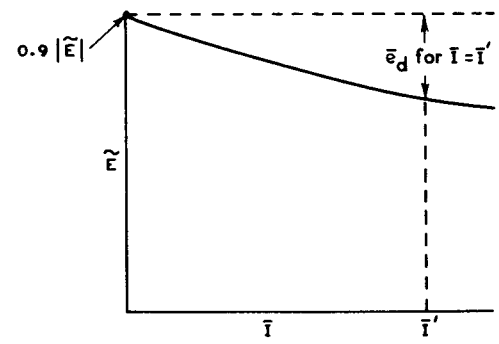


Fig. 14

The peak current $\hat{i}_d$ is closely approximated, if we neglect the higher harmonics, by determining the peak of the 2nd harmonic ripple current and adding it to the average output current $\bar{I}$.

$$\hat{i}_d = \bar{I} + \hat{I}_{2f}$$

$$\hat{I}_{2f} = \frac{\tilde{e}_{max}\ (0.425)}{Z_{2f}}$$

where

$$Z_{2f} = 2\omega L_1 - \frac{1}{2\omega C} \qquad \text{(See Fig. 13)}$$

This equation assumes that the impedance of the capacitance C is small compared to the load resistance R_L shunting it.

The plate dissipation for full-wave, choke-input filter circuits is also readily computed if it is assumed that the ripple current is small compared with $\bar{I}$. Then the total plate dissipation due to plate current flowing is simply the product of the d-c output current $\bar{I}$ by the tube drop $\bar{e}_d$ for that output current.

$$\text{Total Watts Dissipated} = \text{Watts}_{fil} + \bar{I}\, \bar{e}_d$$

RIPPLE VOLTAGE

The ripple voltage with which we are concerned is that appearing directly after the tube.

For condenser-input filters, the frequency of this ripple is the same as that of the supply voltage for half-wave circuits, and is twice the supply frequency for full-wave and voltage-doubler circuits. Fig. 22 in Lecture 19 evaluates the ripple voltage in terms of parameters containing equivalent tube resistance, condenser filter, and load resistance.

For full-wave, choke-input circuits, the ripple-voltage frequency is twice the supply frequency, and is approximately equal to the 2nd term of the Fourier series shown in the preceding section on Choke-Input Filters.

$$e_{2f} = \tilde{E}_{max} (0.425 \cos 2\omega t)$$

This may be further corrected by subtracting the tube drop corresponding to a plate current equal to the output current $\bar{I}$.

$$e_{2f} = (\tilde{E}_{max} - \bar{e}_d)(0.425 \cos 2\omega t)$$

Table I

EMPIRICAL DESIGN DATA FOR RECTIFIERS

This table has been prepared from data on 16 types of RCA rectifiers, mostly for receiving-set applications. In the case of watts per square centimeter radiated from the cathode, conduction losses were lumped with the radiation losses by assuming that the radiation per square centimeter from the uncoated portions was one half that for the coated portions of the cathode.

ITEM	FILAMENT TYPE *	CLOSE-SPACED INDIRECTLY HEATED CATHODE TYPE	MERCURY-VAPOR COATED-FILAMENT TYPE **	UNITS
Cathode Radiation per sq cm	3.5 - 4.0	3.5	4.6 - 4.7	Watts
D-C Output Current per watt input to filament or heater	17 - 25	19 - 25	17	Milliamperes
Peak Current per watt input to filament or heater	100 - 175	170 - 200	150	Milliamperes
D-C Output Current per sq cm of cathode surface	70 - 90	70 - 100	80 - 90	Milliamperes
Peak Current per sq cm of cathode surface	400 - 600	600 - 800	700	Milliamperes

* With No.16 alloy filament ** 2500 volts (rms) per plate max. applied voltage

ITEM	METAL SHELLS °				GLASS BULBS °°			UNITS
Envelope	MT8K	MT8G	MT8B	MT10A	ST12	ST14	ST16 #	-
Dissipation	7.8	14.5	19.7	35.0	16.5	23.5	41	Maximum Watts

° Values are on basis of 2.5 watts per sq. in. Above this value, the paint on the shell burns.
°° Values are on basis that getter flash is small and confined to bottom of bulb.
\# With No.814KW glass stem. When the ST16 bulb is used for full-wave, mercury-vapor tubes with 550 volts rms per plate, the effective dissipation is decreased to about 17.5 watts depending on the ambient temperature. This limitation is necessary to prevent arc-back under conditions where only natural cooling is used.