The Bell System Technical Journal

Vol. XXIX April, 1950 No. 2

Error Detecting and Error Correcting Codes

By R. W. HAMMING

1. Introduction

THE author was led to the study given in this paper from a consideration of large scale computing machines in which a large number of operations must be performed without a single error in the end result. This problem of "doing things right" on a large scale is not essentially new; in a telephone central office, for example, a very large number of operations are performed while the errors leading to wrong numbers are kept well under control, though they have not been completely eliminated. This has been achieved, in part, through the use of self-checking circuits. The occasional failure that escapes routine checking is still detected by the customer and will, if it persists, result in customer complaint, while if it is transient it will produce only occasional wrong numbers. At the same time the rest of the central office functions satisfactorily. In a digital computer, on the other hand, a single failure usually means the complete failure, in the sense that if it is detected no more computing can be done until the failure is located and corrected, while if it escapes detection then it invalidates all subsequent operations of the machine. Put in other words, in a telephone central office there are a number of parallel paths which are more or less independent of each other; in a digital machine there is usually a single long path which passes through the same piece of equipment many, many times before the answer is obtained.

In transmitting information from one place to another digital machines use codes which are simply sets of symbols to which meanings or values are attached. Examples of codes which were designed to detect isolated errors are numerous; among them are the highly developed 2 out of 5 codes used extensively in common control switching systems and in the Bell Relay

Computers,[1] the 3 out of 7 code used for radio telegraphy,[2] and the word count sent at the end of telegrams.

In some situations self checking is not enough. For example, in the Model 5 Relay Computers built by Bell Telephone Laboratories for the Aberdeen Proving Grounds,[1] observations in the early period indicated about two or three relay failures per day in the 8900 relays of the two computers, representing about one failure per two to three million relay operations. The self-checking feature meant that these failures did not introduce undetected errors. Since the machines were run on an unattended basis over nights and week-ends, however, the errors meant that frequently the computations came to a halt although often the machines took up new problems. The present trend is toward electronic speeds in digital computers where the basic elements are somewhat more reliable per operation than relays. However, the incidence of isolated failures, even when detected, may seriously interfere with the normal use of such machines. Thus it appears desirable to examine the next step beyond error detection, namely error correction.

We shall assume that the transmitting equipment handles information in the binary form of a sequence of 0's and 1's. This assumption is made both for mathematical convenience and because the binary system is the natural form for representing the open and closed relays, flip-flop circuits, dots and dashes, and perforated tapes that are used in many forms of communication. Thus each code symbol will be represented by a sequence of 0's and 1's.

The codes used in this paper are called *systematic* codes. Systematic codes may be defined[3] as codes in which each code symbol has exactly n binary digits, where m digits are associated with the information while the other $k = n - m$ digits are used for error detection and correction. This produces a *redundancy* R defined as the ratio of the number of binary digits used to the minimum number necessary to convey the same information, that is,

$$R = n/m.$$

This serves to measure the efficiency of the code as far as the transmission of information is concerned, and is the only aspect of the problem discussed in any detail here. The redundancy may be said to lower the effective channel capacity for sending information.

The need for error correction having assumed importance only recently, very little is known about the economics of the matter. It is clear that in

[1] Franz Alt, "A Bell Telephone Laboratories' Computing Machine"—I, II. Mathematical Tables and Other Aids to Computation, Vol. 3, pp. 1–13 and 60–84, Jan. and Apr. 1948.

[2] S. Sparks, and R. G. Kreer, "Tape Relay System for Radio Telegraph Operation," *R.C.A. Review*, Vol. 8, pp. 393–426, (especially p. 417), 1947.

[3] In Section 7 this is shown to be equivalent to a much weaker appearing definition.

using such codes there will be extra equipment for encoding and correcting errors as well as the lowered effective channel capacity referred to above. Because of these considerations applications of these codes may be expected to occur first only under extreme conditions. Some typical situations seem to be:

a. unattended operation over long periods of time with the minimum of standby equipment.
b. extremely large and tightly interrelated systems where a single failure incapacitates the entire installation.
c. signaling in the presence of noise where it is either impossible or uneconomical to reduce the effect of the noise on the signal.

These situations are occurring more and more often. The first two are particularly true of large scale digital computing machines, while the third occurs, among other places, in "jamming" situations.

The principles for designing error detecting and correcting codes in the cases most likely to be applied first are given in this paper. Circuits for implementing these principles may be designed by the application of well-known techniques, but the problem is not discussed here. Part I of the paper shows how to construct special minimum redundancy codes in the following cases:

a. single error detecting codes
b. single error correcting codes
c. single error correcting plus double error detecting codes.

Part II discusses the general theory of such codes and proves that under the assumptions made the codes of Part I are the "best" possible.

PART I

SPECIAL CODES

2. Single Error Detecting Codes

We may construct a single error detecting code having n binary digits in the following manner: In the first $n - 1$ positions we put $n - 1$ digits of information. In the n-th position we place either 0 or 1, so that the entire n positions have an even number of 1's. This is clearly a single error detecting code since any single error in transmission would leave an odd number of 1's in a code symbol.

The redundancy of these codes is, since $m = n - 1$,

$$R = \frac{n}{n-1} = 1 + \frac{1}{n-1}.$$

It might appear that to gain a low redundancy we should let n become very large. However, by increasing n, the probability of at least one error in a

symbol increases; and the risk of a double error, which would pass undetected, also increases. For example, if $p \ll 1$ is the probability of any error, then for n so large as $1/p$, the probability of a correct symbol is approximately $1/e = 0.3679\ldots$, while a double error has probability $1/2e = 0.1839\ldots$.

The type of check used above to determine whether or not the symbol has any single error will be used throughout the paper and will be called a *parity check*. The above was an *even* parity check; had we used an odd number of 1's to determine the setting of the check position it would have been an odd parity check. Furthermore, a parity check need not always involve all the positions of the symbol but may be a check over selected positions only.

3. Single Error Correcting Codes

To construct a single error correcting code we first assign m of the n available positions as information positions. We shall regard the m as fixed, but the specific positions are left to a later determination. We next assign the k remaining positions as check positions. The values in these k positions are to be determined in the encoding process by even parity checks over selected information positions.

Let us imagine for the moment that we have received a code symbol, with or without an error. Let us apply the k parity checks, in order, and for each time the parity check assigns the value observed in its check position we write a 0, while for each time the assigned and observed values disagree we write a 1. When written from right to left in a line this sequence of k 0's and 1's (to be distinguished from the values assigned by the parity checks) may be regarded as a binary number and will be called the *checking number*. We shall require that this checking number give the position of any single error, with the zero value meaning no error in the symbol. Thus the check number must describe $m + k + 1$ different things, so that

$$2^k \geq m + k + 1$$

is a condition on k. Writing $n = m + k$ we find

$$2^m \leq \frac{2^n}{n+1}.$$

Using this inequality we may calculate Table I, which gives the maximum m for a given n, or, what is the same thing, the minimum n for a given m.

We now determine the positions over which each of the various parity checks is to be applied. The checking number is obtained digit by digit, from right to left, by applying the parity checks in order and writing down the corresponding 0 or 1 as the case may be. Since the checking number is

TABLE I

n	m	Corresponding k
1	0	1
2	0	2
3	1	2
4	1	3
5	2	3
6	3	3
7	4	3
8	4	4
9	5	4
10	6	4
11	7	4
12	8	4
13	9	4
14	10	4
15	11	4
16	11	5
	Etc.	

to give the position of any error in a code symbol, any position which has a 1 on the right of its binary representation must cause the first check to fail. Examining the binary form of the various integers we find

$$\begin{aligned} 1 &= 1 \\ 3 &= 11 \\ 5 &= 101 \\ 7 &= 111 \\ 9 &= 1001 \end{aligned}$$

Etc.

have a 1 on the extreme right. Thus the first parity check must use positions

$$1, 3, 5, 7, 9, \cdots .$$

In an exactly similar fashion we find that the second parity check must use those positions which have 1's for the second digit from the right of their binary representation,

$$\begin{aligned} 2 &= 10 \\ 3 &= 11 \\ 6 &= 110 \\ 7 &= 111 \\ 10 &= 1010 \\ 11 &= 1011 \end{aligned}$$

Etc.,

the third parity check

$$
\begin{aligned}
4 &= 100 \\
5 &= 101 \\
6 &= 110 \\
7 &= 111 \\
12 &= 1100 \\
13 &= 1101 \\
14 &= 1110 \\
15 &= 1111 \\
20 &= 10100 \\
&\text{Etc.}
\end{aligned}
$$

It remains to decide for each parity check which positions are to contain information and which the check. The choice of the positions 1, 2, 4, 8, $\cdots$ for check positions, as given in the following table, has the advantage of making the setting of the check positions independent of each other. All other positions are information positions. Thus we obtain Table II.

TABLE II

Check Number	Check Positions	Positions Checked
1	1	1, 3, 5, 7, 9, 11, 13, 15, 17,$\cdots$
2	2	2, 3, 6, 7, 10, 11, 14, 15, 18,$\cdots$
3	4	4, 5, 6, 7, 12, 13, 14, 15, 20,$\cdots$
4	8	8, 9, 10, 11, 12, 13, 14, 15, 24,$\cdots$
.	.	.
.	.	.
.	.	.

As an illustration of the above theory we apply it to the case of a seven-position code. From Table I we find for $n = 7$, $m = 4$ and $k = 3$. From Table II we find that the first parity check involves positions 1, 3, 5, 7 and is used to determine the value in the first position; the second parity check, positions 2, 3, 6, 7, and determines the value in the second position; and the third parity check, positions 4, 5, 6, 7, and determines the value in position four. This leaves positions 3, 5, 6, 7 as information positions. The results of writing down all possible binary numbers using positions 3, 5, 6, 7, and then calculating the values in the check positions 1, 2, 4, are shown in Table III.

Thus a seven-position single error correcting code admits of 16 code symbols. There are, of course, $2^7 - 16 = 112$ meaningless symbols. In some applications it may be desirable to drop the first symbol from the code to avoid the all zero combination as either a code symbol or a code symbol plus a single error, since this might be confused with no message. This would still leave 15 useful code symbols.

TABLE III

Position							Decimal Value of Symbol
1	2	3	4	5	6	7	
0	0	0	0	0	0	0	0
1	1	0	1	0	0	1	1
0	1	0	1	0	1	0	2
1	0	0	0	0	1	1	3
1	0	0	1	1	0	0	4
0	1	0	0	1	0	1	5
1	1	0	0	1	1	0	6
0	0	0	1	1	1	1	7
1	1	1	0	0	0	0	8
0	0	1	1	0	0	1	9
1	0	1	1	0	1	0	10
0	1	1	0	0	1	1	11
0	1	1	1	1	0	0	12
1	0	1	0	1	0	1	13
0	0	1	0	1	1	0	14
1	1	1	1	1	1	1	15

As an illustration of how this code "works" let us take the symbol 0 1 1 1 1 0 0 corresponding to the decimal value 12 and change the 1 in the fifth position to a 0. We now examine the new symbol

0 1 1 1 0 0 0

by the methods of this section to see how the error is located. From Table II the first parity check is over positions 1, 3, 5, 7 and predicts a 1 for the first position while we find a 0 there; hence we write a

1 .

The second parity check is over positions 2, 3, 6, 7, and predicts the second position correctly; hence we write a 0 to the left of the 1, obtaining

0 1 .

The third parity check is over positions 4, 5, 6, 7 and predicts wrongly; hence we write a 1 to the left of the 0 1, obtaining

1 0 1 .

This sequence of 0's and 1's regarded as a binary number is the number 5; hence the error is in the fifth position. The correct symbol is therefore obtained by changing the 0 in the fifth position to a 1.

4. Single Error Correcting Plus Double Error Detecting Codes

To construct a single error correcting plus double error detecting code we begin with a single error correcting code. To this code we add one more posi-

tion for checking all the previous positions, using an even parity check. To see the operation of this code we have to examine a number of cases:

1. No errors. All parity checks, including the last, are satisfied.
2. Single error. The last parity check fails in all such situations whether the error be in the information, the original check positions, or the last check position. The original checking number gives the position of the error, where now the zero value means the last check position.
3. Two errors. In all such situations the last parity check is satisfied, and the checking number indicates some kind of error.

As an illustration let us construct an eight-position code from the previous seven-position code. To do this we add an eighth position which is chosen so that there are an even number of 1's in the eight positions. Thus we add an eighth column to Table III which has:

TABLE IV

0
0
1
1

1
1
0
0

1
1
0
0

0
0
1
1

PART II

GENERAL THEORY

5. A GEOMETRICAL MODEL

When examining various problems connected with error detecting and correcting codes it is often convenient to introduce a geometric model. The model used here consists in identifying the various sequences of 0's and 1's which are the symbols of a code with vertices of a unit n-dimensional cube. The code points, labelled x, y, z, $\cdots$, form a subset of the set of all vertices of the cube.

Into this space of 2^n points we introduce a *distance*, or, as it is usually called, a *metric*, $D(x, y)$. The definition of the metric is based on the observation that a single error in a code point changes one coordinate, two errors, two coordinates, and in general d errors produce a difference in d coordinates.

Thus we define the distance $D(x, y)$ between two points x and y as the number of coordinates for which x and y are different. This is the same as the least number of edges which must be traversed in going from x to y. This distance function satisfies the usual three conditions for a metric, namely,

$$D(x, y) = 0 \quad \text{if and only if } x = y$$

$$D(x, y) = D(y, x) > 0 \quad \text{if } x \neq y$$

$$D(z, y) + D(y, z) \geq D(x, z) \text{ (triangle inequality).}$$

As an example we note that each of the following code points in the three-dimensional cube is two units away from the others,

0 0 1
0 1 0
1 0 0
1 1 1 .

To continue the geometric language, a sphere of radius r about a point x is defined as all points which are at a distance r from the point x. Thus, in the above example, the first three code points are on a sphere of radius 2 about the point (1, 1, 1). In fact, in this example any one code point may be chosen as the center and the other three will lie on the surface of a sphere of radius 2.

If all the code points are at a distance of at least 2 from each other, then it follows that any single error will carry a code point over to a point that is *not* a code point, and hence is a meaningless symbol. This in turn means that any single error is detectable. If the minimum distance between code points is at least three units then any single error will leave the point nearer to the correct code point than to any other code point, and this means that any single error will be correctable. This type of information is summarized in the following table:

TABLE V

Minimum Distance	Meaning
1	uniqueness
2	single error detection
3	single error correction
4	single error correction plus double error detection
5	double error correction
	Etc.

Conversely, it is evident that, if we are to effect the detection and correction listed, then all the distances between code points must equal or exceed the minimum distance listed. Thus the problem of finding suitable codes is

the same as that of finding subsets of points in the space which maintain at least the minimum distance condition. The special codes in sections 2, 3, and 4 were merely descriptions of how to choose a particular subset of points for minimum distances 2, 3, and 4 respectively.

It should perhaps be noted that, at a given minimum distance, some of the correctability may be exchanged for more detectability. For example, a subset with minimum distance 5 may be used for:

a. double error correction, (with, of course, double error detection).
b. single error correction plus triple error detection.
c. quadruple error detection.

Returning for the moment to the particular codes constructed in Part I we note that any interchanges of positions in a code do not change the code in any essential way. Neither does interchanging the 0's and 1's in any position, a process usually called complementing. This idea is made more precise in the following definition:

Definition. Two codes are said to be *equivalent* to each other if, by a finite number of the following operations, one can be transformed into the other:

1. The interchange of any two positions in the code symbols.
2. The complementing of the values in any position in the code symbols.

This is a formal equivalence relation ($\sim$) since $A \sim A$; $A \sim B$ implies $B \sim A$; and $A \sim B$, $B \sim C$ implies $A \sim C$. Thus we can reduce the study of a class of codes to the study of typical members of each equivalence class.

In terms of the geometric model, equivalence transformations amount to rotations and reflections of the unit cube.

6. Single Error Detecting Codes

The problem studied in this section is that of packing the maximum number of points in a unit n-dimensional cube such that no two points are closer than 2 units from each other. We shall show that, as in section 2, 2^{n-1} points can be so packed, and, further, that any such optimal packing is equivalent to that used in section 2.

To prove these statements we first observe that the vertices of the n-dimensional cube are composed of those of two $(n-1)$-dimensional cubes. Let A be the maximum number of points packed in the original cube. Then one of the two $(n-1)$-dimensional cubes has at least $A/2$ points. This cube being again decomposed into two lower dimensional cubes, we find that one of them has at least $A/2^2$ points. Continuing in this way we come to a two-dimensional cube having $A/2^{n-2}$ points. We now observe that a square can have at most two points separated by at least two units; hence the original n-dimensional cube had at most 2^{n-1} points not less than two units apart.

To prove the equivalence of any two optimal packings we note that, if the packing is optimal, then each of the two sub-cubes has half the points. Calling this the first coordinate we see that half the points have a 0 and half have a 1. The next subdivision will again divide these into two equal groups having 0's and 1's respectively. After $(n - 1)$ such stages we have, upon reordering the assigned values if there be any, exactly the first $n - 1$ positions of the code devised in section 2. To each sequence of the first $n - 1$ coordinates there exist $n - 1$ other sequences which differ from it by one coordinate. Once we fix the n-th coordinate of some one point, say the origin which has all 0's, then to maintain the known minimum distance of two units between code points the n-th coordinate is uniquely determined for all other code points. Thus the last coordinate is determined within a complementation so that any optimal code is equivalent to that given in section 2.

It is interesting to note that in these two proofs we have used only the assumption that the code symbols are all of length n.

7. Single Error Correcting Codes

It has probably been noted by the reader that, in the particular codes of Part I, a distinction was made between information and check positions, while, in the geometric model, there is no real distinction between the various coordinates. To bring the two treatments more in line with each other we redefine a *systematic* code as a code whose symbol lengths are all equal and

1. The positions checked are independent of the information contained in the symbol.
2. The checks are independent of each other.
3. We use parity checks.

This is equivalent to the earlier definition. To show this we form a matrix whose i-th row has 1's in the positions of the i-th parity check and 0's elsewhere. By assumption 1 the matrix is fixed and does not change from code symbol to code symbol. From 2 the rank of the matrix is k. This in turn means that the system can be solved for k of the positions expressed in terms of the other $n - k$ positions. Assumption 3 indicates that in this solving we use the arithmetic in which $1 + 1 = 0$.

There exist non-systematic codes, but so far none have been found which for a given n and minimum distance d have more code symbols than a systematic code. Section 9 gives an example of a non-systematic code.

Turning to the main problem of this section we find from Table V that a single error correcting code has code points at least three units from each other. Thus each point may be surrounded by a sphere of radius 1 with no two spheres having a point in common. Each sphere has a center point and

n points on its surface, a total of $n + 1$ points. Thus the space of 2^n points can have at most:

$$\frac{2^n}{n+1}$$

spheres. This is exactly the bound we found before in section 3.

While we have shown that the special single error correcting code constructed in section 3 is of minimum redundancy, we cannot show that all optimal codes are equivalent, since the following trivial example shows that this is not so. For $n = 4$ we find from Table I that $m = 1$ and $k = 3$. Thus there are at most two code symbols in a four-position code. The following two optimal codes are clearly not equivalent:

$$\begin{matrix} 0\,0\,0\,0 \\ 1\,1\,1\,1 \end{matrix} \quad \text{and} \quad \begin{matrix} 0\,0\,0\,0 \\ 0\,1\,1\,1 \end{matrix}.$$

8. Single Error Correcting Plus Double Error Detecting Codes

In this section we shall prove that the codes constructed in section 4 are of minimum redundancy. We have already shown in section 4 how, for a minimum redundancy code of $n - 1$ dimensions with a minimum distance of 3, we can construct an n dimensional code having the same number of code symbols but with a minimum distance of 4. If this were not of minimum redundancy there would exist a code having more code symbols but with the same n and the same minimum distance 4 between them. Taking this code we remove the last coordinate. This reduces the dimension from n to $n - 1$ and the minimum distance between code symbols by, at most, one unit, while leaving the number of code symbols the same. This contradicts the assumption that the code we began our construction with was of minimum reduncancy. Thus the codes of section 4 are of minimum redundancy.

This is a special case of the following general theorem: To any minimum redundancy code of N points in $n - 1$ dimensions and having a minimum distance of $2k - 1$ there corresponds a minimum redundancy code of N points in n dimensions having a minimum distance of $2k$, and conversely. To construct the n dimensional code from the $n - 1$ dimensional code we simply add a single n-th coordinate which is fixed by an even parity check over the n positions. This also increases the minimum distance by 1 for the following reason: Any two points which, in the $n - 1$ dimensional code, were at a distance $2k - 1$ from each other had an odd number of differences between their coordinates. Thus the parity check was set oppositely for the two points, increasing the distance between them to $2k$. The additional coordinate could not decrease any distances, so that all points in the code are now at a minimum distance of $2k$. To go in the reverse direction we simply

drop one coordinate from the n dimensional code. This reduces the minimum distance of $2k$ to $2k - 1$ while leaving N the same. It is clear that if one code is of minimum redundancy then the other is, too.

9. Miscellaneous Observations

For the next case, minimum distance of five units, one can surround each code point by a sphere of radius 2. Each sphere will contain

$$1 + C(n, 1) + C(n, 2)$$

points, where $C(n, k)$ is the binomial coefficient, so that an upper bound on the number of code points in a systematic code is

$$\frac{2^n}{1 + C(n, 1) + C(n, 2)} = \frac{2^{n+1}}{n^2 + n + 2} \geq 2^m.$$

This bound is too high. For example, in the case of $n = 7$, we find that $m = 2$ so that there should be a code with four code points. The maximum possible, as can be easily found by trial and error, is two.

In a similar fashion a bound on the number of code points may be found whenever the minimum distance between code points is an odd number. A bound on the even cases can then be found by use of the general theorem of the preceding section. These bounds are, in general, too high, as the above example shows.

If we write the bound on the number of code points in a unit cube of dimension n and with minimum distance d between them as $B(n, d)$, then the information of this type in the present paper may be summarized as follows:

$$B(n, 1) = 2^n$$

$$B(n, 2) = 2^{n-1}$$

$$B(n, 3) = 2^m \leq \frac{2^n}{n + 1}$$

$$B(n, 4) = 2^m \leq \frac{2^{n-1}}{n}$$

$$B(n - 1, 2k - 1) = B(n, 2k)$$

$$B(n, 2k - 1) = 2^m \leq \frac{2^n}{1 + C(n, 1) + \cdots + C(n, k - 1)}.$$

While these bounds have been attained for certain cases, no general methods have yet been found for contructing optimal codes when the minimum distance between code points exceeds four units, nor is it known whether the bound is or is not attainable by systematic codes.

We have dealt mainly with systematic codes. The existence of non-systematic codes is proved by the following example of a single error correcting code with $n = 6$.

0 0 0 0 0 0
0 1 0 1 0 1
1 0 0 1 1 0
1 1 1 0 0 0
0 0 1 0 1 1
1 1 1 1 1 1 .

The all 0 symbol indicates that any parity check must be an even one. The all 1 symbol indicates that each parity check must involve an even number of positions. A direct comparison indicates that since no two columns are the same the even parity checks must involve four or six positions. An examination of the second symbol, which has three 1's in it, indicates that no six-position parity check can exist. Trying now the four-position parity checks we find that

1 2 5 6
2 3 4 5

are two independent parity checks and that no third one is independent of these two. Two parity checks can at most locate four positions, and, since there are six positions in the code, these two parity checks are not enough to locate any single error. The code is, however, single error correcting since it satisfies the minimum distance condition of three units.

The only previous work in the field of error correction that has appeared in print, so far as the author is aware, is that of M. J. E. Golay.[4]

[4] M. J. E. Golay, Correspondence, Notes on Digital Coding, *Proceedings of the I.R.E.*, Vol. 37, p. 657, June 1949.

Optical Properties and the Electro-optic and Photoelastic Effects in Crystals Expressed in Tensor Form

By W. P. MASON

I. Introduction

THE electro-optic and photoelastic effects in crystals were first investigated by Pöckels,[1] who developed a phenomenological theory for these effects and measured the constants for a number of crystals. Since then not much work has been done on the subject till the very large electro-optic effects were discovered in two tetragonal crystals ammonium dihydrogen phosphate (ADP) and potassium dihydrogen phosphate (KDP). With these crystals light modulators can be obtained which work on voltages of 2000 volts or less. Their use has been suggested[2] in such equipment as light valves for sound on film recording and in television systems. Furthermore, since the electro-optic effect depends on a change in the dielectric constant with voltage, and the dielectric constant is known to follow the field up to 10^{10} cycles, it is obvious that this effect can be used to produce very short light pulses which may be of interest for physical investigations and for stroboscopic instruments of very high resolution. Hence these crystals renew an interest in the electro-optic effect.

In looking over the literature on the electro-optic effect and photoelastic effect in crystals, there do not seem to be any derivations that give them in terms of thermodynamic potentials, which allow one to investigate the condition under which equalities occur between the various electro-optic and photoelastic constants. Hence it is the purpose of this paper to give such a derivation. Another object is to give a derivation of Maxwell's equations in tensor form, and to apply them to the derivation of the Fresnel ellipsoid.

The first sections deal with the optics of crystals, and derive the Fresnel ellipsoid from Maxwell's equations. Other sections give a derivation of the two effects, discuss methods for measuring them by determining the birefrigence in various directions and give the constants for the two effects in terms of crystal symmetries. The final section discusses the application of the photoelastic effect for measuring strains in isotropic media.

[1] F. Pöckels, Lehrbuck Der Kristalloptic, B. Teubner, Leipzig, 1906.

[2] See *Patent* 2,467,325 issued to the writer; "Light Modulation by P type Crystals," George D. Gotschall, *Jour. Soc. Motion Picture Engineers*, July, 1948, pp. 13–20; B. H. Billings, *Jour. Opt. Soc. Am.*, 39, 797, 802 (1949).

II. Solution of Maxwell's Equations In Tensor Form

In tensor notation, Maxwell's equations for a nonmagnetic medium with no free charges take the form

$$\frac{1}{V}\frac{\partial D_i}{\partial t} = \epsilon_{ijk}\frac{\partial H_j}{\partial x_k}; \quad \frac{1}{V}\frac{\partial H_j}{\partial t} = -\epsilon_{jki}\frac{\partial E_k}{\partial x_i}; \quad \frac{\partial D_i}{\partial x_i} = 0; \quad \frac{\partial H_j}{\partial x_j} = 0 \tag{1}$$

where D_i is the electric displacement, H_j the magnetic field, E_k the electric field, V the velocity of light in vacuo and ϵ_{ijk} a tensor equal to zero when $i = j$ or k or $j = k$, but equal to 1 or -1 when all three numbers are different. If the numbers are in rotation, i.e. 1, 2, 3; 2, 3, 1; 3, 1, 2 the value is $+1$ while, if they are out of rotation, the value is -1.

We assume the electric vector to be representable by a plane wave whose planes of equal phase are taken normal to the unit vector n_i. Then

$$E_k = E_{0_k} e^{j\omega(t - x_i n_i/v)} \tag{2}$$

where E_{0_k} are constants representing the maximum values of the field along the three rectangular coordinates and $j = \sqrt{-1}$. Substituting (2) in the second of equations (1), noting that E_{0_k} are not functions of the space coordinates, we have

$$\frac{1}{V}\frac{\partial H_j}{\partial t} = \frac{j\omega}{v}[\epsilon_{jki} E_{0_k} n_i] e^{j\omega[t - x_i n_i/v]}. \tag{3}$$

Integrating with respect to the time

$$H_j = \frac{V}{v}[\epsilon_{jki} E_{0_k} n_i] e^{j\omega[t - x_i n_i/v]} = H_{0_j} e^{j\omega(t - x_i n_i/v)}. \tag{4}$$

Hence,

$$H_{0_j} = \frac{V}{v}[\epsilon_{jki} E_{0_k} n_i] \tag{5}$$

and therefore the magnetic vector is normal to the plane determined by E_{0_k} and n_i.

Next, using the first of equations (1),

$$\frac{\partial D_i}{\partial t} = V\epsilon_{ijk}\frac{\partial H_j}{\partial x_k} = V\epsilon_{ijk} H_{0_j}\frac{\partial e^{j\omega[t - x_k n_k/v]}}{\partial x_k} = -\frac{j\omega V}{v}[\epsilon_{ijk} H_{0_j} n_k] e^{j\omega[t - x_k n_k/v]}. \tag{6}$$

Integrating with respect to time,

$$D_i = -\frac{V}{v}[\epsilon_{ijk} H_{0_j} n_k] e^{j\omega[t - x_k n_k/v]}. \tag{7}$$

Inserting the value of H_{0_j} from (5), this equation takes the form

$$D_i = -\frac{V^2}{v^2}[\epsilon_{ijk}(\epsilon_{jki} E_{0_k} n_i)n_k]e^{j\omega[t-x_i n_i/v]}$$

and, in general,

$$D_i = -\frac{V^2}{v^2}[\epsilon_{ijk}(\epsilon_{jki} E_k n_i)n_k]. \tag{9}$$

Expanding the inner parenthesis, we have the components

$$(E_2 n_3 - E_3 n_2)_1; \qquad (E_3 n_1 - E_1 n_3)_2; \qquad (E_1 n_2 - E_2 n_1)_3. \tag{10}$$

Then

$\epsilon_{ijk}[(E_2 n_3 - E_5 n_2); (E_3 n_1 - E_1 n_3); (E_1 n_2 - E_2 n_1)]n_k$ gives

$$\begin{aligned}
D_1 &= -\frac{V^2}{v^2}[(E_3 n_1 - E_1 n_3)n_3 - (E_1 n_2 - E_2 n_1)n_2] \\
&= [(E_3 n_3 + E_2 n_2 + E_1 n_1)n_1 - E_1(n_1^2 + n_2^2 + n_3^2)] \\
D_2 &= -\frac{V^2}{v^2}[(E_1 n_2 - E_2 n_1)n_1 - (E_2 n_3 - E_3 n_2)n_3] \\
&= [(E_3 n_3 + E_2 n_2 + E_1 n_1)n_2 - E_2(n_1^2 + n_2^2 + n_3^2)] \\
D_3 &= -\frac{V^2}{v^2}[(E_2 n_3 - E_3 n_2)n_2 - (E_3 n_1 - E_1 n_3)n_1] \\
&= [(E_3 n_3 + E_2 n_2 + E_1 n_1)n_3 - E_3(n_1^2 + n_2^2 + n_3^2)].
\end{aligned} \tag{11}$$

Now, since $n_1^2 + n_2^2 + n_3^2 = 1$ because n is a unit vector, we have

$$D_i = \frac{V^2}{v^2}[E_i - (E_j n_j)n_i] \quad \text{or} \quad \frac{v^2}{V^2} D_i - E_i - (E_j n_j)n_i = 0. \tag{12}$$

This equation states that D_i, E_i and n_i are in the same plane, H_j being normal to the plane as shown by Fig. 1. The energy flow vector

$$S_i = \frac{V^2}{4\pi} \epsilon_{ijk} E_j H_k \tag{13}$$

also lies in the plane since it is perpendicular to E and H. It is at the same angle θ with n that E is with D. The velocity of energy flow is $v/\cos\theta$. The energy velocity is called the ray velocity and the energy path the ray path.

Next, from the relation for a material medium, that

$$D_i = K_{ij} E_j \text{ or conversely } E_j = \beta_{ji} D_i \tag{14}$$

where K_{ij} are the dielectric constants measured at optical frequencies and β_{ji} are the impermeability constants determined from the relations

$$\beta_{ji} = \Delta^{ji}/\Delta^{K} \tag{15}$$

where

$$\Delta^{K} = \begin{vmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{vmatrix}$$

and Δ^{ji} the determinant obtained by suppressing the j^{th} row and i^{th} column, we can eliminate E_i from equation (12) and obtain

$$\begin{aligned} \frac{v^2}{V^2} D_1 &= \beta_{11} D_1 + \beta_{12} D_2 + \beta_{13} D_3 - (E_j n_j) n_1 \\ \frac{v^2}{V^2} D_2 &= \beta_{12} D_1 + \beta_{22} D_2 + \beta_{23} D_3 - (E_j n_j) n_2 \\ \frac{v^2}{V^2} D_3 &= \beta_{13} D_1 + \beta_{23} D_2 + \beta_{33} D_3 - (E_j n_j) n_3. \end{aligned} \tag{16}$$

This can be put in the form

$$\begin{aligned} (E_j n_j) n_1 &= D_1[\beta_{11} - v^2/V^2] + \beta_{12} D_2 + \beta_{13} D_3 \\ (E_j n_j) n_2 &= \beta_{12} D_1 + (\beta_{22} - v^2/V^2) D_2 + \beta_{23} D_3 \\ (E_j n_j) n_3 &= \beta_{13} D_1 + \beta_{23} D_2 + (\beta_{33} - v^2/V^2) D_3. \end{aligned} \tag{17}$$

Solving for D_1, D_2 and D_3

$$\begin{aligned} D_1 &= [(\beta_{22} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{23}^2][E_j n_j] n_1 \\ D_2 &= [(\beta_{11} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{13}^2][E_j n_j] n_2 \\ D_3 &= [(\beta_{11} - v^2/V^2)(\beta_{22} - v^2/V^2) - \beta_{12}^2][E_j n_j] n_3. \end{aligned} \tag{18}$$

Now, since D and n are at right angles,

$$D_1 n_1 + D_2 n_2 + D_3 n_3 = 0. \tag{19}$$

Hence,

$$\begin{aligned} 0 = {} & [(\beta_{22} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{23}^2] n_1^2 \\ & + [(\beta_{11} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{13}^2] n_2^2 \\ & + [(\beta_{11} - v^2/V^2)(\beta_{22} - v^2/V^2) - \beta_{12}^2] n_3^2. \end{aligned} \tag{20}$$

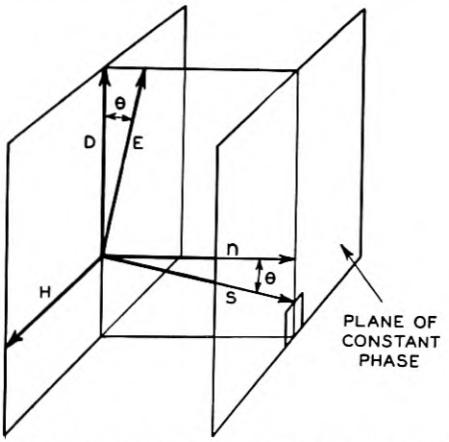


Fig. 1—Position of electric, magnetic and normal vectors for an electromagnetic plane wave in a crystal.

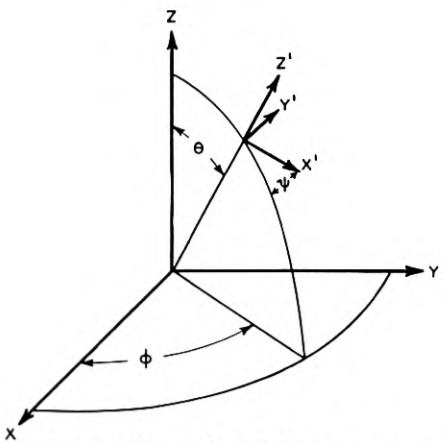


Fig. 2—Rotated axes and angles for relating them to unrotated axes.

By choosing the original x, y, z axes so that $\beta_{12} = \beta_{13} = \beta_{23} = 0$ and using the values $\beta_{11} = \beta_1$, $\beta_{22} = \beta_2$, $\beta_{33} = \beta_3$ this gives the equation

$$\frac{n_1^2}{\beta_1 - \frac{v^2}{V^2}} + \frac{n_2^2}{\beta_2 - \frac{v^2}{V^2}} + \frac{n_3^2}{\beta_3 - \frac{v^2}{V^2}} = 0. \tag{21}$$

For transmission along the X axis $n_1 = 1$, $n_2 = n_3 = 0$ and the two velocities are given by

$$v^2 = \beta_2 V^2 = b^2, \qquad v^2 = \beta_3 V^2 = c^2. \tag{22}$$

Similarly the third velocity $v^2 = \beta_1 V^2 = a^2$ can also be used and equation (21) reduces to

$$\frac{n_1^2}{a^2 - v^2} + \frac{n_2^2}{b^2 - v^2} + \frac{n_3^2}{c^2 - v^2} = 0. \tag{23}$$

This is a quadratic equation for the velocities v in terms of the principal velocities a, b and c which are usually taken so that $a > b > c$.

Solving for the velocities, we obtain the quadratic equation

$$v^4 - v^2[n_1^2(b^2 + c^2) + n_2^2(a^2 + c^2) + n_3^2(a^2 + b^2)] + n_1^2b^2c^2 + n_2^2a^2c^2 + n_3^2a^2b^2 = 0. \tag{24}$$

Letting $L = n_1^2(b^2 - c^2)$, $M = n_2^2(c^2 - a^2)$, $N = n_3^2(a^2 - b^2)$ the solutions for the velocities become

$$2v^2 = n_1^2(b^2 + c^2) + n_2^2(c^2 + a^2) + n_3^2(a^2 + b^2) \pm \sqrt{L^2 + M^2 + N^2 - 2LM - 2LN - 2MN}. \tag{25}$$

This equation can be put into a simpler form if we change to the coordinate system shown by Fig. 2. Here the rotated system is related to the original system by three angles θ, φ, ψ. θ is the angle between the Z' axis and the Z axis, φ is the angle the plane containing Z and Z' makes with the X axis while ψ represents a rotation of the primed coordinate systems about the Z' axis. The direction cosines for the primed system with respect to the normal system are designated by the matrix

	X	Y	Z
X'	ℓ_1	m_1	n_1
Y'	ℓ_2	m_2	n_2
Z'	ℓ_3	m_3	n_3

(26)

where, in terms of θ, φ and ψ, these direction cosines are,

$$\ell_1 = \cos\theta\cos\varphi\cos\psi - \sin\varphi\sin\psi,$$

$$m_1 = \cos\theta\sin\varphi\cos\psi + \cos\varphi\sin\psi, \qquad n_1 = -\sin\theta\cos\psi$$

$$\ell_2 = -\cos\theta\cos\varphi\sin\psi - \sin\varphi\cos\psi,$$

$$m_2 = \cos\varphi\cos\psi - \sin\varphi\sin\psi\cos\theta, \qquad n_2 = \sin\theta\sin\psi$$

$$\ell_3 = \cos\varphi\sin\theta, \qquad m_3 = \sin\varphi\sin\theta, \qquad n_3 = \cos\theta. \tag{27}$$

If we take Z' as the direction of the wave normal, then in equation (25)

$$n_1 = \ell_3, \qquad n_2 = m_3, \qquad n_3 = n_3$$

and the equation for the velocities becomes

$$2v^2 = a^2(\sin^2\varphi \sin^2\theta + \cos^2\theta) + b^2(\cos^2\varphi \sin^2\theta + \cos^2\theta) + c^2 \sin^2\theta \pm \sqrt{(a^2 - b^2)^2(\cos^2\theta \cos^2\varphi + \sin^2\varphi)^2 + 2(a^2 - b^2)(c^2 - b^2) \sin^2\theta(\cos^2\theta \cos^2\varphi - \sin^2\varphi) + (c^2 - b^2)^2 \sin^4\theta} \quad (28)$$

A very elegant construction for the wave-velocities and the directions of vibration is the Fresnel index ellipsoid. Consider the ellipsoid

$$a^2x^2 + b^2y^2 + c^2z^2 = 1 \quad (29)$$

Then Fresnel[3] showed that, for any diametral plane perpendicular to the wave normal, the two principal axes of the ellipse were the directions of the two permitted vibrations, while the wave velocities were the reciprocals of the principal semi-axes.

We wish to show now that the maximum and minimum values of the impermeability constants in a plane perpendicular to the direction of the wave normal determine the directions of vibration and the values of the two velocities. To show this we make use of the fact that β_{ij} is a second rank tensor and transforms according to the tensor transformation formula

$$\beta'_{ij} = \frac{\partial x'_i}{\partial x_k} \frac{\partial x'_j}{\partial x_\ell} \beta_{k\ell} \quad (30)$$

where the partial derivatives are the direction cosines

$$\frac{\partial x'_1}{\partial x_1} = \ell_1, \quad \frac{\partial x'_1}{\partial x_2} = m_1, \quad \frac{\partial x'_1}{\partial x_3} = n_1$$

$$\frac{\partial x'_2}{\partial x_1} = \ell_2, \quad \frac{\partial x'_2}{\partial x_2} = m_2, \quad \frac{\partial x'_2}{\partial x_3} = n_2$$

$$\frac{\partial x'_3}{\partial x_1} = \ell_3, \quad \frac{\partial x'_3}{\partial x_2} = m_3, \quad \frac{\partial x'_3}{\partial x_3} = n_3.$$

Expanding equation (30) the six transformation equations become

$$\beta'_{11} = \ell_1^2\beta_{11} + 2\ell_1m_1\beta_{12} + 2\ell_1n_1\beta_{13} + m_1^2\beta_{22} + 2m_1n_1\beta_{23} + n_1^2\beta_{33}$$

$$\beta'_{12} = \ell_1\ell_2\beta_{11} + (\ell_1m_2 + m_1\ell_2)\beta_{12} + (\ell_1n_2 + n_1\ell_2)\beta_{13} + m_1m_2\beta_{22} + (m_1n_2 + n_1m_2)\beta_{23} + n_1n_2\beta_{33}$$

$$\beta'_{13} = \ell_1\ell_3\beta_{11} + (\ell_1m_3 + m_1\ell_3)\beta_{12} + (\ell_1n_3 + n_1\ell_3)\beta_{13} + m_1m_3\beta_{22} + (n_1m_3 + m_1n_3)\beta_{23} + n_1n_3\beta_{33} \quad (31)$$

[3] See for example "Photoelasticity," Coker and Filon, Cambridge University Press, pages 17 and 18.

$$\beta'_{22} = \ell_2^2\beta_{11} + 2\ell_2 m_2\beta_{12} + 2\ell_2 n_2\beta_{13} + m_2^2\beta_{22} + 2m_2 n_2\beta_{23} + n_2^2\beta_{33}$$

$$\beta'_{23} = \ell_2\ell_3\beta_{11} + (\ell_2 m_3 + m_2\ell_3)\beta_{12} + (\ell_2 n_3 + n_2\ell_3)\beta_{13} + m_2 m_3\beta_{22} + (m_2 n_3 + n_2 m_3)\beta_{23} + n_2 n_3\beta_{33}$$

$$\beta'_{33} = \ell_3^2\beta_{11} + 2\ell_3 m_3\beta_{12} + 2\ell_3 n_3\beta_{13} + m_3^2\beta_{22} + 2m_3 n_3\beta_{23} + n_3^2\beta_{33}.$$

Now, if the axes refer to the axes of a Fresnel ellipsoid, $\beta_{12} = \beta_{13} = \beta_{23} = 0$ and one of the impermeability constants for any direction, say β'_{33}, can be expressed in the form

$$\beta'_{33} = \ell_3^2\beta_1 + m_3^2\beta_2 + n_3^2\beta_3 \tag{32}$$

If r, which lies along Z' of Fig. 2, is the radius vector of the Fresnel ellipsoid, then the direction cosines ℓ_3, m_3 and n_3 are

$$\ell_3 = \frac{x}{r}, \qquad m_3 = \frac{y}{r}, \qquad n_3 = \frac{z}{r}.$$

From equation (24) $\beta_1 = a^2/V^2$, $\beta_2 = b^2/V^2$, $\beta_3 = c^2/V^2$ and equation (32) becomes

$$r^2V^2\beta'_{33} = a^2x^2 + b^2y^2 + c^2z^2 = 1.$$

Hence the square of the radius vector of the Fresnel ellipsoid is $1/V^2\beta'_{33}$ and the radius vector of the impermeability ellipsoid agrees with that of the Fresnel ellipsoid. Hence, the directions of vibration can be determined from the principal axes of the impermeability ellipsoid for any diametral plane.

When light transmission occurs along Z', the direction for maximum and minimum impermeability can be obtained by evaluating β'_{11} and determining the angle ψ for which it has an extreme value. Inserting the direction cosines ℓ_1, m_1 and n_1 from equation (27), we find

$$\begin{aligned}\beta'_{11} = {} & \beta_1\left[\cos^2\theta\cos^2\varphi\cos^2\psi - \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \sin^2\varphi\sin^2\psi\right] \\ & + \beta_2\left[\cos^2\theta\sin^2\varphi\cos^2\psi + \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \cos^2\varphi\sin^2\psi\right] \\ & + \beta_3\sin^2\theta\cos^2\psi.\end{aligned} \tag{33}$$

Differentiating with respect to ψ and setting the resultant derivative equal to zero, the value of ψ that will satisfy the equation is given by

$$\begin{aligned}\tan 2\psi &= \frac{(\beta_2 - \beta_1)\sin 2\varphi\cos\theta}{(\beta_1 - \beta_2)(\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (\beta_3 - \beta_2)\sin^2\theta} \\ &= \frac{(b^2 - a^2)\sin 2\varphi\cos\theta}{(a^2 - b^2)(\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (c^2 - b^2)\sin^2\theta}.\end{aligned} \tag{34}$$

For a given value on the right-hand side there are two values of ψ, 90° apart, that will satisfy the equation and hence we have two directions of vibration at right angles to each other. Inserting (34) in (33) the values of β'_{11} and β''_{11} for these two directions are

$$2\beta'_{11} = \beta_1(\sin^2\varphi \sin^2\theta + \cos^2\theta) + \beta_2(\cos^2\varphi \sin^2\theta + \cos^2\theta) + \beta_3 \sin^2\theta \pm \sqrt{(\beta_1 - \beta_2)^2 (\cos^2\theta \cos^2\varphi + \sin^2\varphi)^2 + 2(\beta_1 - \beta_2)(\beta_3 - \beta_2) \cdot \sin^2\theta (\cos^2\theta \cos^2\varphi - \sin^2\varphi) + (\beta_3 - \beta_2)^2 \sin^4\theta}.$$

Since β_1 corresponds to a^2, etc., this equation agrees with the two velocities given in equation (28) and shows that the directions of vibration correspond with the maximum and minimum values of β'_{11}.

It can also be shown that the two directions of electric displacement coincide with the two values of ψ given by equation (34). Transforming the electrical displacements to the X', Y', Z' set of axes we have

$$\begin{aligned} D'_1 &= \frac{\partial x'_1}{\partial x_1} D_1 + \frac{\partial x'_1}{\partial x_2} D_2 + \frac{\partial x'_1}{\partial x_3} D_3 = \ell_1 D_1 + m_1 D_2 + n_1 D_3 \\ D'_2 &= \frac{\partial x'_2}{\partial x_1} D_1 + \frac{\partial x'_2}{\partial x_2} D_2 + \frac{\partial x'_2}{\partial x_2} D_3 = \ell_2 D_1 + m_2 D_2 + n_2 D_3 \\ D'_3 &= \frac{\partial x'_3}{\partial x_1} D_1 + \frac{\partial x'_3}{\partial x_2} D_2 + \frac{\partial x'_3}{\partial x_3} D_3 = \ell_3 D_1 + m_3 D_2 + n_3 D_3. \end{aligned} \tag{35}$$

Hence, inserting the values of D_1, D_2, D_3 from equation (18), we find

$$\begin{aligned} D'_1 &= \ell_1\ell_3(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_1m_3(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_1n_3(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}) \\ D'_2 &= \ell_2\ell_3(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_2m_3(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_2n_3(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}) \\ D'_3 &= \ell_3^2(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_3^2(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_3^2(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}). \end{aligned} \tag{36}$$

From equation (20) with $\beta_{12} = \beta_{13} = \beta_{23} = 0$, it is evident that the D_3 component vanishes and hence the two values of electric displacement lie in a plane perpendicular to Z'. By inserting the values of β'_{11} and the value of ψ found from equation (34) we find that $D_2 = 0$ and hence the electric displacement lies along the directions of the greatest value of β'_{11}. Similarly, from the second value of β'_{11}, D_1 vanishes and hence the second wave is perpendicular to the first and in the direction of the smallest value of β'_{11}.

III. Location of Optic Axes in a Crystal

When the expression in the radical of equation (28) vanishes the two velocities are equal and an optic axis exists. Since the expression inside the radical can be written

$$[(a^2 - b^2)(\cos^2\theta\cos^2\varphi + \sin^2\varphi) - (b^2 - c^2)\sin^2\theta]^2 - 4(a^2 - b^2)(c^2 - b^2)\sin^2\theta\sin^2\varphi = 0 \tag{37}$$

then, since the square is always positive and since $(a^2 - b^2) > 0$ and $(b^2 - c^2) > 0$, the equation can vanish only if $\varphi = 0$. But $\varphi = 0$ indicates that the two optic axes always lie in a plane perpendicular to the intermediate velocity b. With $\varphi = 0$ then the square vanishes when

$$\tan^2\theta = \frac{(a^2 - b^2)}{(b^2 - c^2)} \quad \text{or} \quad \tan\theta = \pm\sqrt{\frac{a^2 - b^2}{b^2 - c^2}}. \tag{38}$$

If $(a^2 - b^2) < (b^2 - c^2)$ the value of the $\tan\theta$ is less than unity and the crystal is called a positive crystal. For this case the two axes approach more closely the Z axis having the velocity c than they do the X axis. If $(a^2 - b^2) > (b^2 - c^2)$ the crystal is negative.

If $a = b$ or $b = c$ the crystal has a single optic axis and is respectively a positive or negative uniaxial crystal. For the first case the two velocities are given by

$$v_1 = a = b, \qquad v_2 = \sqrt{a^2\cos^2\theta + c^2\sin^2\theta}. \tag{39}$$

The first velocity is that of the ordinary ray while that of the second is that of the extraordinary ray. Since $a > c$, the ordinary ray will have a velocity greater than the extraordinary ray except along the optic axis where they are equal. Since $c < a$, the maximum axis for any ellipse, formed by intersecting the Fresnel ellipsoid at an angle to the optic axis, will lie in the plane formed by the normal and the c axis and hence the direction of polarization of the extraordinary ray will lie in the c, n plane. The polarization of the ordinary ray will be perpendicular to this plane.

If $b = c$ the a axis is the optic axis and the velocities of the two rays are again

$$v_1 = c \text{ and } v_2^2 = a^2(1\text{-}\sin^2\theta\cos^2\varphi) + c^2(\sin^2\theta\cos^2\varphi) \tag{40}$$

Hence, when $\theta = 90°$, $\varphi = 0°$, the two velocities are equal and a is the optic axis. In this case the velocity of the extraordinary ray is greater than that of the ordinary ray except along the a axis, and the crystal is a negative uniaxial crystal. The polarization of the extraordinary ray lies again in the

plane of the normal and the optic axis while the ordinary ray is perpendicular to it.

IV. Derivation of the Electro-optic and Photoelastic Effects

In a previous paper[4] and in the book "Piezoelectric Crystals and Their Application to Ultrasonics", D. Van Nostrand, 1950, it was shown that the electro-optic and photoelastic effects can be expressed as third derivatives of one of the thermodynamic potentials. Probably the most fundamental way of developing these properties is to express them in terms of the strains, electric displacements and the entropy. For viscoelastic substances it has been shown that the photoelastic effects are directly related to the strains. In terms of the electric displacements, the electro-optic constants do not vary much with temperature whereas, if they are expressed in terms of the fields, the constants of a ferroelectric type of crystal such as KDP increase many fold near the Curie temperature. The entropy is chosen as the fundamental heat variable, since most measurements are carried out so rapidly that the entropy does not vary.

The thermodynamic potential which has the strains, electric displacements and entropy as the independent variables is the internal energy U, given by

$$dU = T_{ij}\, dS_{ij} + E_m \frac{dD_m}{4\pi} + \Theta\, d\sigma \tag{41}$$

where S_{ij} are the strains, T_{ij} the stresses, E_m the fields, D_m the electric displacements, Θ the temperature and σ the entropy. In this equation the strains S_{ij} are defined in the tensor form

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{42}$$

where the u's are the displacements along the three axis. In the case of a shearing strain occurring when $i \neq j$, the strain is only half that usually used in engineering practice. In order to avoid writing the factor $1/4\pi$, we use the variable $\delta_m = D_m/4\pi$. Then, from (41),

$$T_{ij} = \frac{\partial U}{\partial S_{ij}}, \qquad E_m = \frac{\partial U}{\partial \delta_n}, \qquad \theta = \frac{\partial U}{\partial \sigma}. \tag{43}$$

Since, for most conditions of interest, adiabatic conditions prevail, we can set $d\sigma$ equal to zero and can develop the dependent variables, the fields and

[4] "First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form," W. P. Mason, *B.S.T.J.*, Vol. 26, pp. 80–138, Jan., 1947.

the stresses in terms of the independent variables, the strains and the electric displacements. Up to the second derivatives, these are

$$E_m = \frac{\partial E_m}{\partial S_{ij}} S_{ij} + \frac{\partial E_m}{\partial \delta_n} \delta_n + \frac{1}{2!}\left[\frac{\partial^2 E_m}{\partial S_{ij}\partial S_{qr}} S_{ij}S_{qr} + \frac{2\partial^2 E_m}{\partial S_{ij}\partial \delta_n} S_{ij}\delta_n + \frac{\partial^2 E_m}{\partial \delta_n \partial \delta_o}\delta_n\delta_o\right] + \cdots$$

$$T_{k\ell} = \frac{\partial T_{k\ell}}{\partial S_{ij}} S_{ij} + \frac{\partial T_{k\ell}}{\partial \delta_n} \delta_n + \frac{1}{2!}\left[\frac{\partial^2 T_{k\ell}}{\partial S_{ij}\partial S_{qr}} S_{ij}S_{qr} + \frac{2\partial^2 T_{k\ell}}{\partial S_{ij}\partial \delta_n} S_{ij}\delta_n + \frac{\partial^2 T_{k\ell}}{\partial \delta_n \partial \delta_o}\delta_n\delta_o\right] + \cdots . \tag{44}$$

For the electro-optic and photoelastic cases, the two tensors of interest are

$$\frac{\partial^2 T_{k\ell}}{\partial \delta_n \partial \delta_o} = \frac{\partial^3 U}{\partial S_{k\ell}\partial \delta_n \partial \delta_o} = \frac{\partial^2 E_n}{\partial S_{k\ell}\partial \delta_o} = 4\pi m_{k\ell no}$$
$$\frac{\partial^2 E_m}{\partial \delta_n \partial \delta_o} = \frac{\partial^3 U}{\partial \delta_m \partial \delta_n \partial \delta_o} = (4\pi) r_{mno} . \tag{45}$$

For the first partial derivatives, we have the values

$$\frac{\partial T_{k\ell}}{\partial S_{ij}} = c^D_{ijk\ell} ; \qquad \frac{\partial T_{k\ell}}{\partial \delta_n} = \frac{\partial^2 U}{\partial S_{k\ell}\partial \delta_n} = \frac{\partial E_n}{\partial S_{k\ell}} = -h_{nk\ell}$$
$$\frac{\partial E_m}{\partial \delta_n} = 4\pi \beta^S_{mn} \tag{46}$$

where $c^D_{ijk\ell}$ are the elastic stiffnesses measured at constant electric displacement, $h_{nk\ell}$ are the piezoelectric constants that relate the open circuit voltages to the strains, and β^S_{mn} are the impermeability constants measured for constant strain.

With these substitutions and neglecting the other second partial derivatives, we have, from (44),

$$E_m = -h_{mij}S_{ij} + D_n\left[\beta^S_{mn} + m_{ijmn}S_{ij} + \frac{r^S_{mno}}{2} D_o\right] + \cdots$$
$$T_{k\ell} = c^D_{ijk\ell}S_{ij} + D_o\left[-\frac{h_{ok\ell}}{4\pi} + \frac{m_{k\ell on}D_n}{2}\right]. \tag{47}$$

This equation shows that there is a relation between the change in the impermeability constant due to stress in the first equation, and the electrostrictive constant in the second equation through the tensor m_{ijmn}. These

effects, however, have to be measured at the same frequency before equality exists.

To obtain the changes in the optical properties caused by the strain and the electric displacement we have to determine the fields and displacements occurring at the high frequencies of optics. Even for piezoelectric vibrations occurring at as high frequencies as they can be driven by the piezoelectric effect, these frequencies are small compared to the optic frequencies f and can be considered to be static displacements or strains. Hence, writing

$$E_m = E_m^0 + E_m e^{j\omega t}, \qquad D_n = D_n^0 + D_n e^{j\omega t},$$
$$D_o = D_o^0 + D_o e^{j\omega t}, \qquad S_{ij} = S_{ij}^0$$

where $\omega = 2\pi f$, the first of equation (47) can be written in the form

$$E_m^0 = -h_{mij} S_{ij} + D_n^0 \left[\beta_{mn}^S + m_{ijmn} S_{ij} + \frac{r_{mno}}{2} D_o^0 \right]$$
$$E_m e^{j\omega t} = D_n e^{j\omega t} \left[\beta_{mn}^S + m_{ijmn} S_{ij} + \frac{r_{mno}^S}{2} D_o^0 \right] + \frac{r_{mno}}{2} D_n^0 D_o e^{j\omega t}. \tag{48}$$

If we develop one of the fields, say E_1, this can be written in the form

$$\begin{aligned} E_1 e^{j\omega t} &= [\beta_{11} + m_{ij11} S_{ij} + r_{111} D_1^0 + r_{112} D_2^0 + r_{113} D_3^0] D_1 e^{j\omega t} \\ &+ [\beta_{12} + M_{ij12} S_{ij} + r_{121} D_1^0 + r_{122} D_2^0 + r_{123} D_3^0] D_2 e^{j\omega t} \\ &+ [\beta_{13} + M_{ij13} S_{ij} + r_{131} D_1^0 + r_{132} D_2^0 + r_{133} D_3^0] D_3 e^{i\omega t} \end{aligned} \tag{49}$$

where the first number of r refers to the field, the second to the optical value of D and the third to the static value of D. Hence, for the general case,

$$E_m e^{j\omega t} = D_n e^{j\omega t} [\beta_{mn} + m_{ijmn} S_{ij} + r_{mno} D_o^0]. \tag{50}$$

From the definition of the two tensors m_{ijno} and r_{mno} given by equation (45), we can show that there are relations between the various components of the tensors. For the first tensor m_{ijno}, since $S_{ij} = S_{ji}$ is a symmetrical tensor, then

$$m_{ijno} = m_{jino} \tag{51}$$

From the definition of the tensor m_{ijno} in the form

$$4\pi m_{ijno} = \frac{\partial}{\partial S_{ij}} \left(\frac{\partial^2 U}{\partial \delta_n \, \partial \delta_o} \right) \tag{45}$$

it is obvious that we can interchange the order of δ_n and δ_o so that

$$m_{ijno} = m_{ijon}$$

Since ij and no are reversible, it has been customary to abbreviate the tensor by writing one number in place of the two in the following form:

$$11 = 1; 22 = 2; 33 = 3; 12 = 21 = 6; 13 = 31 = 5; 23 = 32 = 4 \quad (52)$$

Since the reduced tensor is associated with the engineering strains, it is necessary to investigate the numerical relationships between the four index symbols and the two index symbols. From equation (48), when $m \neq n$, the change in the impermeability constant β_{mn} is given by

$$m_{ijmn} S_{ij} + m_{jimn} S_{ji} = m_{rs} S_r \quad (53)$$

Since $S_r = 2S_{ij} = 2S_{ji}$ we have the relation that

$$m_{ijmn} = m_{rs} (i, j, m, n = 1 \text{ to } 3, r, s, = 1 \text{ to } 6) \quad (54)$$

In equation (45) we cannot in general interchange the order of ij and no since U does not contain product terms of strains and electric displacements and hence in general

$$m_{rs} \neq m_{sr} . \quad (55)$$

Hence in the most general case there are 36 photoelastic constants. Crystal symmetrics cut down the number of constants as shown in a later section.

The tensor r_{mno} defined in equation (45) as

$$(4\pi)^2 r_{mno} = \frac{\partial^3 U}{\partial \delta_m \partial \delta_n \partial \delta_o} \quad (56)$$

shows that we can interchange the order of m and n since U contains product terms of δ_m and δ_n . Hence

$$r_{mno} = r_{nmo} \quad (57)$$

and this is usually replaced by the two index symbols

$$r_{qo} = r_{mno} (m, n, o = 1 \text{ to } 3; q = 1 \text{ to } 6).$$

The so called "true" electro-optic constants are measured at constant strain and for this case the modifications in the impermeability constants are given by the equation

$$E_m = D_n [\beta^S_{mn} + r^S_{mno} D_o] . \quad (58)$$

Since m and n are interchangeable, the third rank tensor is usually replaced by the two index symbols

$$r^S_{mno} = r_{qo} (m, n, o = 1 \text{ to } 3; q = 1 \text{ to } 6). \quad (59)$$

As discussed in the next sections, these constants can be determined by applying an electric field of a frequency high enough so that the principal resonances and their harmonics cannot be excited by the applied field, and measuring the resulting birefringence along definite directions in the crystal. On the other hand if we apply a static field to the crystal, an additional effect occurs because the crystal is strained by the piezoelectric effect and this causes a photoelastic effect in addition to the "true" electro-optic effect. A

better designation for these effects is the electro-optic effect at constant strain and stress.

This latter effect can be calculated from equation (47) by setting the stresses $T_{k\ell}$ equal to zero and eliminating the S_{ij} strains. After neglecting second order corrections,

$$E_m = D_n e^{j\omega t}\left[\beta^S_{mn} + \left(r^S_{mno} + \frac{m_{ijmn} h_{ok\ell}}{4\pi c^D_{ijk\ell}}\right) D^0_o\right]. \tag{60}$$

Since $h_{ok\ell}/c^D_{ijk\ell} = g_{oij}$, the other piezoelectric constant relating the open circuit voltage to the stress, the electro-optic effect at constant stress can be written in the form

$$r^T_{mno} = r^S_{mno} + \frac{m_{ijmn}\, g_{oij}}{4\pi}. \tag{61}$$

In terms of the two index symbols

$$r^T_{qo} = r^S_{qo} + \frac{m_{pq}\, g_{op}}{4\pi} \tag{62}$$

since it has been shown[4] that $g_{oij} = g_{op}/2$ when $i \neq j$, and the tensor in (61) has ij as common symbols which involves the summations of two terms.

The electro-optic effect is usually measured in terms of an applied field. The change in the impermeability constant β^S_{mn} for this case can be determined from the first equations (47), setting $T_{k\ell}$ equal to zero and neglecting second order terms. Multiplying through by the tensor K^T_{op} of the dielectric constants

$$D^0_p = E^0_0 K^T_{op} \tag{63}$$

since the product $K^T_{op}\beta^T_{op} = 1$. Introducing this equation into (58) we have

$$E_m = D_n[\beta^S_{mn} + r^S_{mnp} K^T_{op} E^0_o] = D_n[\beta^S_{mn} + z^S_{mno} E^0_o]. \tag{64}$$

where the new tensor z_{mno} is equal to

$$z^S_{mno} = r^S_{mnp} K^T_{op}. \tag{65}$$

In terms of the two index symbols

$$z^S_{qo} = r^S_{qp} K^T_{op}. \tag{66}$$

in which the repeated index indicates a summation. The difference between the electro-optical constant at constant stress expressed in terms of the field and the electro-optical constant at constant strain is

$$z^T_{mno} = z^S_{mno} + \frac{m_{ijmn}\, g_{oij}}{4\pi} K^T_{op} = z^S_{mno} + m_{ijmn}\, d_{pij} \tag{67}$$

since the piezoelectric constants d_{pij} are related to the g constants by the equation

$$d_{pij} = \frac{g_{oij} K^T_{op}}{4\pi}. \tag{68}$$

In terms of two index symbols

$$z_{qo}^{T} = z_{qo}^{S} + m_{pq}d_{op}(p, q = 1 \text{ to } 6; o = 1 \text{ to } 3) \tag{69}$$

where a repeated index means a summation with respect to this index.

Finally the photoelastic effect is sometimes expressed in terms of the stresses rather than the strains. As can be seen from equation (47), the new set of constants is

$$\pi_{pq} = m_{pr}s_{rq}^{D} \tag{70}$$

where the s_{rq}^{D} are the elastic compliances measured at constant electric displacement.

V. Birefringence Along Any Direction In the Crystal and Determination of the Electro-optic and Photoelastic Constants

If we take axes along the Fresnel ellipsoid when no stress or field is applied to the crystal, the result of the electro-optic and photoelastic effects is to change the impermeability constants by the values

$$\begin{aligned} &\beta_{11} = \beta_1 + \Delta_1\,; \quad \beta_{22} = \beta_2 + \Delta_2\,; \quad \beta_{33} = \beta_3 + \Delta_3 \\ &\beta_{23} = \Delta_4\,; \quad \beta_{13} = \Delta_5\,; \quad \beta_{12} = \Delta_6 \end{aligned} \tag{71}$$

where

$$\begin{aligned} \Delta_1 &= z_{11}E_1 + z_{12}E_2 + z_{13}E_3 + m_{11}S_1 + m_{12}S_2 + m_{13}S_3 + m_{14}S_4 \\ &\quad + m_{15}S_5 + m_{16}S_6 \\ \Delta_2 &= z_{21}E_1 + z_{22}E_2 + z_{23}E_3 + m_{21}S_1 + m_{22}S_2 + m_{23}S_3 + m_{24}S_4 \\ &\quad + m_{25}S_5 + m_{26}S_6 \\ \Delta_3 &= z_{31}E_1 + z_{32}E_2 + z_{33}E_3 + m_{31}S_1 + m_{32}S_2 + m_{33}S_3 + m_{34}S_4 \\ &\quad + m_{35}S_5 + m_{36}S_6 \\ \Delta_4 &= z_{41}E_1 + z_{42}E_2 + z_{43}E_3 + m_{41}S_1 + m_{42}S_2 + m_{43}S_3 + m_{44}S_4 \\ &\quad + m_{45}S_5 + m_{46}S_6 \\ \Delta_5 &= z_{51}E_1 + z_{52}E_2 + z_{53}E_3 + m_{51}S_1 + m_{52}S_2 + m_{53}S_3 + m_{54}S_4 \\ &\quad + m_{55}S_5 + m_{56}S_6 \\ \Delta_6 &= z_{61}E_1 + z_{62}E_2 + z_{63}E_3 + m_{61}S_1 + m_{62}S_2 + m_{63}S_3 + m_{64}S_4 \\ &\quad + m_{65}S_5 + m_{66}S_6\,. \end{aligned} \tag{72}$$

If we transmit light along the z' axis which, as shown by Fig. 2, makes an angle of θ degrees with the z axis in a plane making an angle φ with the xz plane, the birefringence can be calculated as follows: Keeping z' fixed and rotating the other two axes about z' by varying the angle ψ, one light vector

will occur when β'_{11} is a maximum and the other when β'_{11} is a minimum. Using the transformation equations (31) and the direction cosines of (27), we find that β'_{11} is given by the equations

$$\begin{aligned}\beta'_{11} = {} & \beta_{11}\left[\cos^2\theta\cos^2\varphi\cos^2\psi - \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \sin^2\varphi\sin^2\psi\right] \\ & + \beta_{12}[\sin 2\varphi\cos 2\psi - \sin^2\theta\sin 2\varphi\cos^2\psi + \cos\theta\sin 2\psi\cos 2\varphi] \\ & + \beta_{13}[-\sin 2\theta\cos\varphi\cos^2\psi + \sin\varphi\sin\theta\sin 2\psi] \\ & + \beta_{22}\left[\cos^2\theta\sin^2\varphi\cos^2\psi + \frac{\cos\theta\sin 2\varphi\sin 2\psi}{2} + \cos^2\varphi\sin^2\psi\right] \\ & + \beta_{23}[-\sin 2\theta\sin\varphi\cos^2\psi - \sin\theta\cos\varphi\sin 2\psi] + \beta_{33}\sin^2\theta\cos^2\psi\end{aligned} \tag{73}$$

Differentiating with respect to ψ and setting $\frac{\partial\beta'_{11}}{\partial\psi} = 0$, we find an expression for $\tan 2\psi$ in the form

$$\tan 2\psi = \frac{\begin{array}{r}-\beta_{11}\sin 2\varphi\cos\theta + 2\beta_{12}\cos\theta\cos 2\varphi \\ + 2\beta_{13}\sin\varphi\sin\theta + \beta_{22}\cos\theta\sin 2\varphi - 2\beta_{23}\sin\theta\cos\varphi\end{array}}{\begin{array}{r}\beta_{11}[\cos^2\theta\cos^2\varphi - \sin^2\varphi] + \beta_{12}[(1+\cos^2\theta)\sin 2\varphi] \\ - \beta_{13}\sin^2\theta\cos\varphi + \beta_{22}(\cos^2\theta\sin^2\varphi - \cos^2\varphi) \\ - \beta_{23}\sin 2\theta\sin\varphi + \beta_{33}\sin^2\theta\end{array}} \tag{74}$$

Inserting this value back in equation (73) we find that the two extreme values of β'_{11} are given by the equation

$$\begin{aligned}2\beta'_{11} = {} & 2\beta_{22} + (\beta_{11} - \beta_{22})(\cos^2\theta\cos^2\varphi + \sin^2\varphi) + (\beta_{33} - \beta_{22})\sin^2\theta \\ & - \beta_{12}\sin^2\theta\sin 2\varphi - \beta_{13}\sin 2\theta\cos\varphi - \beta_{23}\sin 2\theta\sin\varphi \\ & \pm \sqrt{\begin{array}{l}(\beta_{11} - \beta_{22})^2(\cos^2\theta\cos^2\varphi + \sin^2\varphi)^2 + 2(\beta_{11} - \beta_{22})(\beta_{33} - \beta_{22})\sin^2\theta\times \\ (\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (\beta_{33} - \beta_{22})^2\sin^4\theta - 2(\beta_{11} - \beta_{22})\times \\ [\beta_{12}(\sin 2\varphi\sin^2\theta(\cos^2\theta\cos^2\varphi + \sin^2\varphi) + \beta_{13}\sin 2\theta\cos\varphi\times \\ (\cos^2\theta\cos^2\varphi + \sin^2\varphi) - \beta_{23}\sin 2\theta\sin\varphi(1 + \cos^2\varphi\sin^2\theta)] \\ + 2(\beta_{33} - \beta_{22})\sin^2\theta[\beta_{12}\sin 2\varphi(1 + \cos^2\theta) - \beta_{13}\sin 2\theta\cos\varphi \\ - \beta_{23}\sin 2\theta\sin\varphi] + (2\beta_{12})^2[\sin^4\theta\sin^2\varphi\cos^2\varphi + \cos^2\theta] \\ - 4\beta_{12}\beta_{13}\sin^2\theta\sin\varphi[\cos^2\theta\cos^2\varphi + \sin^2\varphi] - 4(\beta_{12}\beta_{23}) \\ [\sin 2\theta\cos\varphi(\sin^2\varphi\cos^2\theta + \cos^2\varphi)] + (2\beta_{13})^2\sin^2\theta\times \\ (\cos^2\theta\cos^2\varphi + \sin^2\varphi) - 4\beta_{13}\beta_{23}\sin 2\varphi\sin^4\theta \\ + (2\beta_{23})^2\sin^2\theta(\cos^2\theta\sin^2\varphi + \cos^2\varphi)\end{array}}\end{aligned} \tag{75}$$

The birefringence in any direction can be calculated from equation (75); since $\beta'_{11} = v_1^2/V^2$, it equals $1/\mu_1^2$ where μ_1 is the index of refraction corresponding to a light wave with its electric displacement in the β'_{11} direction. Similarly, for the second solution at right angle to the first,

$$\beta''_{11} = \frac{v_2^2}{V^2} = \frac{1}{\mu_2^2} \tag{76}$$

Hence if we designate the expression under the radical by K_2 and half the expression on the right outside the radical by K_1, we have

$$\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} = K_1; \qquad \frac{1}{\mu_1^2} - \frac{1}{\mu_2^2} = \sqrt{K_2}. \tag{77}$$

Since μ_1 and μ_2 are very nearly equal even in the most birefringent crystal, we have nearly

$$\mu_2 - \mu_1 = B = \frac{\mu^3}{2}\sqrt{K_2}. \tag{78}$$

For special directions in the crystal, the expression for K_2 simplifies very considerably. Along the x, y and z axes, the values are

$$\begin{aligned} X,\ (\varphi = 0°, \theta = 90°); \qquad & B_x = \frac{\mu^3}{2}\sqrt{(\beta_{33} - \beta_{22})^2 + (2\beta_{23})^2} \\ Y,\ (\varphi = 90°, \theta = 90°); \qquad & B_y = \frac{\mu^3}{2}\sqrt{(\beta_{11} - \beta_{33})^2 + (2\beta_{13})^2} \\ Z,\ (\varphi = 0°, \theta = 0°); \qquad & B_z = \frac{\mu^3}{2}\sqrt{(\beta_{11} - \beta_{22})^2 + (2\beta_{12})^2}. \end{aligned} \tag{79}$$

If any natural birefringence exists along these axes, $(2\beta_{23})^2$ will be very small compared to this and

$$\begin{aligned} B_x &= \frac{\mu^3}{2}(\beta_3 - \beta_2 + \Delta_3 - \Delta_2) = \frac{\mu^3}{2}\left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right) \\ B_y &= \frac{\mu^3}{2}(\beta_1 - \beta_3 + \Delta_1 - \Delta_3) = \frac{\mu^3}{2}\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_c^2} + \Delta_1 - \Delta_3\right) \\ B_z &= \frac{\mu^3}{2}(\beta_1 - \beta_2 + \Delta_1 - \Delta_2) = \frac{\mu^3}{2}\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right). \end{aligned} \tag{80}$$

Hence, for this case, measurements along the three axes will tell the difference between the three effects Δ_1, Δ_2 and Δ_3. To get absolute values requires a direct measurement of the index of refraction along one of the axes and its change with fields or stresses. This is a considerably more difficult meas-

urement than a birefringence measurement and requires the use of an accurate interferometer.

If, however, the Z axis is an optic axis as it is in ADP, for example, and $\Delta_1 = \Delta_2 = 0$, a birefringence occurs due to the term β_{12}. As shown in the next section, the electro-optic constants for ADP (tetragonal $\bar{4}2m$) are z_{41} and z_{63}. z_{63} occurs in the expression for $\beta_{12} = \Delta_6$, as can be seen from equations (72), and hence the birefringence along the Z axis is

$$B_z = \frac{\mu_a^3}{2} x 2\beta_{12} = \mu_a^3 z_{63} E_3. \tag{81}$$

The constants z_{63} and z_{41} have been measured independently by W. L. Bond, Robert O'B. Carpenter, and Hans Jaffe. Probably the most accurate measurements, and the only one published, are those of Carpenter,[5] who finds that the indices of refraction and the z_{63} and z_{41} constants for ADP and KDP are in cgs units

	μ_a	μ_c	$r_{63} \times 10^7$	$r_{41} \times 10^7$
ADP	1.5254	1.4798	2.54 ± 0.05	6.25 ± 0.1
KDP	1.5100	1.4684	3.15 ± 0.07	2.58 ± 0.05

An even larger constant has been found for heavy hydrogen KDP by Zwicker and Scherrer.[6] They find at 20°C that $r_{63} = 6 \times 10^{-7}$. Using this constant, a half wave retardation for a $\lambda = 5461$ A° mercury line occurs for a voltage of 4000 volts.

For tetragonal crystals of these types the only photoeleastic constant for the z axis is m_{66}, and the birefringence for this case is given by

$$B_z = \mu_a^3 m_{66} S_6 \tag{82}$$

When a natural birefringence exists for the crystal, measurements of the other three effects Δ_4, Δ_5 and Δ_6 can be made by determining the birefringence along other directions than the Fresnel ellipsoid axes. In a direction of Z' lying in the XZ plane $\varphi = 0$, $\theta =$ variable and

$$B_{xz} = \frac{\mu^3}{2} \sqrt{\begin{aligned}&[(\beta_{11} - \beta_{22}) \cos^2\theta + (\beta_{33} - \beta_{22}) \sin^2\theta - \beta_{13} \sin^2\theta]^2 \\ &+ [2\beta_{12} \cos\theta + 2\beta_{23} \sin\theta]^2.\end{aligned}} \tag{83}$$

When a natural birefringence exists, this reduces to

$$B_{xy} \doteq \frac{\mu^3}{2}\left[\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right)\cos^2\theta + \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right)\sin^2\theta - \Delta_5 \sin 2\theta\right] \tag{84}$$

[5] "The Electro-optic Effect in Uniaxial Crystals of the Type XH_2PO_4," Robert O'B. Carpenter, *Jour. Opt. Soc. Am.*, in course of publication.

[6] Zwicker and Scherrer, *Helv. Phys. Acta.*, 17, 346 (1944).

and hence, by measuring at 45° between the two axes, one can evaluate the Δ_5 term.

Similarly, for the YZ plane, $\varphi = 90°$, $\theta =$ variable and

$$B_{yz} = \frac{\mu^3}{2}\sqrt{[-(\beta_{11} - \beta_{22}) + (\beta_{33} - \beta_{22})\sin^2\theta - \beta_{23}\sin 2\theta]^2 + [2\beta_{12}\cos\theta - 2\beta_{13}\sin\theta]^2}. \tag{85}$$

Hence, when a natural birefringence exists, we have

$$B_{yz} \doteq \frac{\mu^3}{2}\left[-\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right) + \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right)\sin^2\theta - \Delta_4\sin 2\theta\right]. \tag{86}$$

In the XY plane $\theta = 90°$, $\varphi =$ variable and

$$B_{xy} = \frac{\mu^3}{2}\sqrt{[(\beta_{11} - \beta_{12})\sin^2\varphi - (\beta_{33} - \beta_{22}) - \beta_{12}\sin 2\varphi]^2 + [2\beta_{13}\sin\varphi - \beta_{23}\cos\varphi]^2}. \tag{87}$$

Then, for natural birefringence,

$$B_{xy} \doteq \frac{\mu^3}{2}\left[\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right)\sin^2\varphi - \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right) - \Delta_6\sin 2\varphi\right]. \tag{88}$$

Hence, with measurements at 45° between the axes and with suitably applied fields and strains, the three effects Δ_4, Δ_5 and Δ_6 can be measured. Since the axes of the test specimen are turned with respect to the X, Y and Z axes, suitable transformations of the effects Δ_1 to Δ_6 with respect to the new axes will have to be made. These can be done as shown in reference (4) by means of tensor transformation formulae.

Another method for measuring the constants in Δ_4, Δ_5, Δ_6 is to measure the amount they rotate the axes of the Fresnel ellipsoid. As an example consider the z_{41} constant of ADP. For example, if we look along the X axis and apply a field in the same direction, then, in equation (74), $\theta = 90°$, $\varphi = 0$ and

$$\tan 2\psi = \frac{-2\beta_{23}}{\beta_{33} - \beta_{22}} = \frac{-2z_{41}E_1}{\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2}} = \frac{-2\mu_b^2\mu_c^2 z_{41}E_1}{(\mu_b + \mu_c)(\mu_b - \mu_c)}. \tag{89}$$

According to Carpenter, the z_{41} electro-optic constant of ADP is 6.25×10^{-7} in cgs units. $\mu_a = \mu_b = 1.5254$; $\mu_c = 1.4798$; hence the angle of rotation for a field of 30,000 volts per centimeter = 100 stat volts cm is

$$\psi = -2.25 \times 10^{-3} \text{ radians} = 7.7 \text{ minutes of arc.} \tag{90}$$

VI. Electro-optic and Photoelastic Tensors for Various Crystal Classes

Since $r_{mno} = r_{nmo}$ and $z_{mno} = z_{nmo}$ are third rank tensors similar to the h_{mij} piezoelectric tensor, they will have the same components for the various crystal classes. For the twenty crystal classes that show the electro-optic effect these tensors are given below. They are given with the crystal system they belong to, and the symmetry is designated by the Hermann-Mauguin symbol. The last number of the subscript of z designates the direction of the applied static field.

(91)

Triclinic; 1	z_{11}	z_{21}	z_{31}	z_{41}	z_{51}	z_{61}
	z_{12}	z_{22}	z_{32}	z_{42}	z_{52}	z_{62}
	z_{13}	z_{23}	z_{33}	z_{43}	z_{53}	z_{63}
Monoclinic; 2	0	0	0	z_{41}	0	z_{61}
	z_{12}	z_{22}	z_{32}	0	z_{52}	0
	0	0	0	z_{43}	0	z_{63}
Monoclinic; $\bar{2} = m$	z_{11}	z_{21}	z_{31}	0	z_{51}	0
	0	0	0	z_{42}	0	z_{62}
	z_{13}	z_{23}	z_{33}	0	z_{53}	0
Orthorhombic; 222	0	0	0	z_{41}	0	0
	0	0	0	0	z_{52}	0
	0	0	0	0	0	z_{63}
Orthorhomic; 2mm	0	0	0	0	z_{51}	0
	0	0	0	z_{42}	0	0
	z_{13}	z_{23}	z_{33}	0	0	0
Tetragonal; $\bar{4}$	0	0	0	z_{41}	z_{51}	0
	0	0	0	$-z_{51}$	z_{41}	0
	z_{13}	$-z_{13}$	0	0	0	z_{63}

Tetragonal; 4

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & z_{51} & 0 \\ 0 & 0 & 0 & z_{51} & -z_{41} & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal; $\bar{4}2m$

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & z_{41} & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{63} \end{vmatrix}$$

Tetragonal; 422

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & -z_{41} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal; 4mm

$$\begin{vmatrix} 0 & 0 & 0 & 0 & z_{51} & 0 \\ 0 & 0 & 0 & z_{51} & 0 & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 3

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & z_{41} & z_{51} & -z_{22} \\ -z_{22} & z_{22} & 0 & z_{51} & -z_{41} & -z_{11} \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 32

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & -z_{41} & -z_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 3m

$$\begin{vmatrix} 0 & 0 & 0 & 0 & z_{51} & -z_{22} \\ -z_{22} & z_{22} & 0 & z_{51} & 0 & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal; $\bar{6}$

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & 0 & 0 & -z_{22} \\ -z_{22} & z_{22} & 0 & 0 & 0 & -z_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal; $\bar{6}$m2	z_{11}	$-z_{11}$	0	0	0	0
	0	0	0	0	0	$-z_{11}$
	0	0	0	0	0	0
Hexagonal; 6	0	0	0	z_{41}	z_{51}	0
	0	0	0	z_{51}	$-z_{41}$	0
	z_{13}	z_{13}	z_{33}	0	0	0
Hexagonal; 622	0	0	0	z_{41}	0	0
	0	0	0	0	$-z_{41}$	0
	0	0	0	0	0	0
Hexagonal; 6mm	0	0	0	0	z_{51}	0
	0	0	0	z_{51}	0	0
	z_{13}	z_{13}	z_{33}	0	0	0
Cubic; 23 and $\bar{4}$3m	0	0	0	z_{41}	0	0
	0	0	0	0	z_{41}	0
	0	0	0	0	0	z_{41}

The r tensor has similar terms.

The photoelastic constants are similar to the elastic constant tensors except that $m_{rs} \neq m_{sr}$ in general. However, for the tetragonal, trigonal, hexagonal and cubic systems, Pockels found that $m_{12} = m_{21}$. This follows from the transformation equations about the Z axis which is the n fold axes for these groups. For a rotation of an angle θ about Z, the direction cosines are

$$\begin{vmatrix} \ell_1 = \dfrac{\partial x_1'}{\partial x_1} = \cos\theta & m_1 = \dfrac{\partial x_1'}{\partial x_2} = \sin\theta & n_1 = \dfrac{\partial x_1'}{\partial x_3} = 0 \\ \ell_2 = \dfrac{\partial x_2'}{\partial x_1} = -\sin\theta & m_2 = \dfrac{\partial x_2'}{\partial x_2} = \cos\theta & n_2 = \dfrac{\partial x_2'}{\partial x_3} = 0 \\ \ell_3 = \dfrac{\partial x_3'}{\partial x_1} = 0 & m_3 = \dfrac{\partial x_3'}{\partial x_2} = 0 & n_3 = \dfrac{\partial x_3'}{\partial x_3} = 1 \end{vmatrix} \tag{92}$$

Transforming the two terms $m'_{1122} = m'_{12}$ and $m'_{2211} = m'_{21}$ by the tensor transformation equation

$$m_{ijk\ell} = \frac{\partial x'_i}{\partial x_m}\frac{\partial x'_j}{\partial x_n}\frac{\partial x'_k}{\partial x_o}\frac{\partial x'_\ell}{\partial x_p} m_{mnop} \tag{93}$$

we find, for these two coefficients,

$$\begin{aligned} m'_{12} &= (m_{11} + m_{22} - 4m_{66}) \sin^2\theta \cos^2\theta + 2(m_{62} - m_{16}) \\ &\quad \sin\theta \cos^3\theta + 2(m_{61} - m_{16}) \sin^3\theta \cos\theta + m_{12} \cos^4\theta + m_{21} \sin^4\theta \\ m'_{21} &= (m_{11} + m_{22} - 4m_{66}) \sin^2\theta \cos^2\theta + 2(m_{16} - m_{62}) \\ &\quad \sin^3\theta \cos\theta + 2(m_{26} - m_{61}) \sin\theta \cos^3\theta + m_{21} \cos^4\theta + m_{12} \sin^4\theta \end{aligned} \tag{94}$$

If $m'_{12} = m'_{21}$ for all angles of rotation we must have

$$m_{16} + m_{26} = m_{61} + m_{62}$$

For all the classes that $m_{12} = m_{21}$, either $m_{26} = -m_{16}$ and $m_{62} = -m_{61}$ or else $m_{16} = m_{26} = m_{61} = m_{62} = 0$.

Now, if Z is a four-fold axis, as it is in the tetragonal and cubic systems, then, for a 90° rotation, the value of m'_{12} or m'_{21} must repeat. From the first of (92) this means that

$$m_{12} = m_{21} \text{ and } m_{21} = m_{12}$$

For a trigonal or hexagonal system additional relations are obtained between m_{66} and m_{11}, m_{22} and m_{12} in the usual manner. Hence the photoelastic matrices become, for the various crystal classes,

(95)

Triclinic 36 Constant	$\begin{vmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} & m_{16} \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} & m_{26} \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} & m_{36} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} & m_{46} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} & m_{56} \\ m_{61} & m_{62} & m_{63} & m_{64} & m_{65} & m_{66} \end{vmatrix}$	The π tensor is entirely analogous
Monoclinic 20 Constants	$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & m_{15} & 0 \\ m_{21} & m_{22} & m_{23} & 0 & m_{25} & 0 \\ m_{31} & m_{32} & m_{33} & 0 & m_{35} & 0 \\ 0 & 0 & 0 & m_{44} & 0 & m_{46} \\ m_{51} & m_{52} & m_{53} & 0 & m_{55} & 0 \\ 0 & 0 & 0 & m_{64} & 0 & m_{66} \end{vmatrix}$	The π tensor is entirely analogous

Orthorhombic 12 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & 0 \\ m_{21} & m_{22} & m_{23} & 0 & 0 & 0 \\ m_{31} & m_{32} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Tetragonal 4,4,4/*m* 9 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & m_{16} \\ m_{12} & m_{11} & m_{13} & 0 & 0 & -m_{16} \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & 0 \\ m_{61} & -m_{61} & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Tetragonal $\bar{4}2m$, 422 $4mm$, $(\bar{4}/m)mm$ 7 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & 0 \\ m_{12} & m_{11} & m_{13} & 0 & 0 & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Trigonal 3,3 11 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & m_{14} & -m_{25} & 0 \\ m_{12} & m_{11} & m_{13} & -m_{14} & m_{25} & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ m_{41} & -m_{41} & 0 & m_{44} & m_{45} & m_{52} \\ -m_{52} & m_{52} & 0 & -m_{45} & m_{44} & m_{41} \\ 0 & 0 & 0 & m_{25} & m_{14} & \frac{m_{11}-m_{12}}{2} \end{vmatrix}$$

The π tensor is analogous except that $\pi_{46} = 2\pi_{52}$ $\pi_{56} = 2\pi_{41}$ $\pi_{66} = (\pi_{11} - \pi_{12})$

Trigonal 32,3*m* 3(2/*m*) 8 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & m_{14} & 0 & 0 \\ m_{12} & m_{11} & m_{13} & -m_{14} & 0 & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ m_{41} & -m_{41} & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & m_{41} \\ 0 & 0 & 0 & 0 & m_{14} & \frac{m_{11}-m_{12}}{2} \end{vmatrix}$$

The π tensor is analogous except that $\pi_{56} = 2\pi_{41}$ $\pi_{66} = \pi_{11} - \pi_{12}$

Hexagonal $6, 6m2, 6$ $622, 6/m;$ $6mm, \frac{6}{m}mm$ 6 Constants	m_{11}	m_{12}	m_{13}	0	0	0	The π tensor is analogous except that $\pi_{66} = \pi_{11} - \pi_{12}$
	m_{12}	m_{11}	m_{13}	0	0	0	
	m_{31}	m_{31}	m_{33}	0	0	0	
	0	0	0	m_{44}	0	0	
	0	0	0	0	m_{44}	0	
	0	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	
Cubic System 23, 432 $\frac{2}{m}3, 43m, \frac{4}{m}3\frac{2}{m}$ 3 Constants	m_{11}	m_{12}	m_{12}	0	0	0	The π tensor is entirely analogous (95)
	m_{12}	m_{11}	m_{12}	0	0	0	
	m_{12}	m_{12}	m_{11}	0	0	0	
	0	0	0	m_{44}	0	0	
	0	0	0	0	m_{44}	0	
	0	0	0	0	0	m_{44}	
Isotropic Systems 2 Constants	m_{11}	m_{12}	m_{12}	0	0	0	The π tensor is analogous except that $\pi_{66} = \pi_{11} - \pi_{12}$
	m_{12}	m_{11}	m_{12}	0	0	0	
	m_{12}	m_{12}	m_{11}	0	0	0	
	0	0	0	$\frac{m_{11}-m_{12}}{2}$	0	0	
	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	0	
	0	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	

From measurement[7] on the photoelastic effects at high pressure for cubic crystals, it has become apparent that the second derivatives of equation (44) are not sufficient to represent the experimental results and derivatives up to the fourth power should be included. This extension, however, is not considered in the present paper.

VII. Photoelasticity in Isotropic Media

The photoelastic effect in isotropic solids has been used extensively in studying the stresses existing in machine parts and other pieces. For this purpose a plastic model cut in the shape of the original is used and is loaded in a similar manner to that of the machine part to be studied. Since stresses are applied, the π_i photoelastic constants are most useful. If we look along

[7] H. B. Maris, *Jour. Optical Society of Amer.*, Vol. 15, pp. 194–200, 1927.

the Z axis, the last of equations (79) shows that the birefringence is equal to

$$B_z = \frac{\mu^3}{2}\sqrt{(\beta_1 + \Delta_1 - \beta_2 - \Delta_2)^2 + 4(\Delta_6)^2} \tag{96}$$

Since, for an isotropic substance $\beta_1 = \beta_2$, we have, after substituting the value of Δ_1 and Δ_2, with the appropriate photoelastic constants from equation (95), (last tensor):

$$B_z = \frac{\mu^3}{2}(\pi_{11} - \pi_{12})\sqrt{(T_1 - T_2)^2 + 4T_6^2} \tag{97}$$

If we transform to axes rotated by an angle θ about Z, the values of T'_{11} and T'_{22} are given by

$$\begin{aligned} T'_{11} &= \cos^2\theta T_1 + 2\sin\theta\cos\theta T_6 + \sin^2\theta T_2 \\ T'_{22} &= \sin^2\theta T_1 - 2\sin\theta\cos\theta T_6 + \cos^2\theta T_2 \end{aligned} \tag{98}$$

If, now, we choose the angle θ so that T'_{11} is a maximum, we find

$$\tan 2\theta = \frac{+2T_6}{T_1 - T_2} \tag{99}$$

Inserting this value of $\tan 2\theta$ in (98) we find

$$\begin{aligned} T'_1 &= \frac{T_1 + T_2}{2} + \tfrac{1}{2}\sqrt{(T_1 - T_2)^2 + 4T_6^2} \\ T'_2 &= \frac{T_1 + T_2}{2} - \tfrac{1}{2}\sqrt{(T_1 - T_2)^2 + 4T_6^2} \end{aligned} \tag{100}$$

and, hence,

$$T'_1 - T'_2 = \sqrt{(T_1 - T_2)^2 + 4T_6^2} \tag{101}$$

Hence the birefringence obtained in stressing a material is proportional to the difference in the principal stresses. By observing the isoclinic lines of a photoelastic picture, methods[8] are available for determining the stresses in a model. A photograph[9] of a stressed disk is shown by Fig. 3. The high concentration of lines near the surface shows that the shearing stress is very high at these points. By counting the number of lines from the edge and knowing the stress optical constant, the stress can be calculated at any point.

If we apply a single stress T_1, the birefringence is given by the equation

$$B_z = \frac{\mu^3}{2}(\pi_{11} - \pi_{12})T_1 \tag{102}$$

[8] See Photoelasticity, Coker and Filon, Cambridge University Press, 1931.
[9] This photograph was taken by T. F. Osmer.

Instead of using the constants π_{11} and π_{12} it is customary to use a single constant C given by

$$B = \mu_e - \mu_o = r = CT \tag{103}$$

where the constant C is called the relative stress optical constant and r the retardation. The dimensions of C are the reciprocal of a stress and are

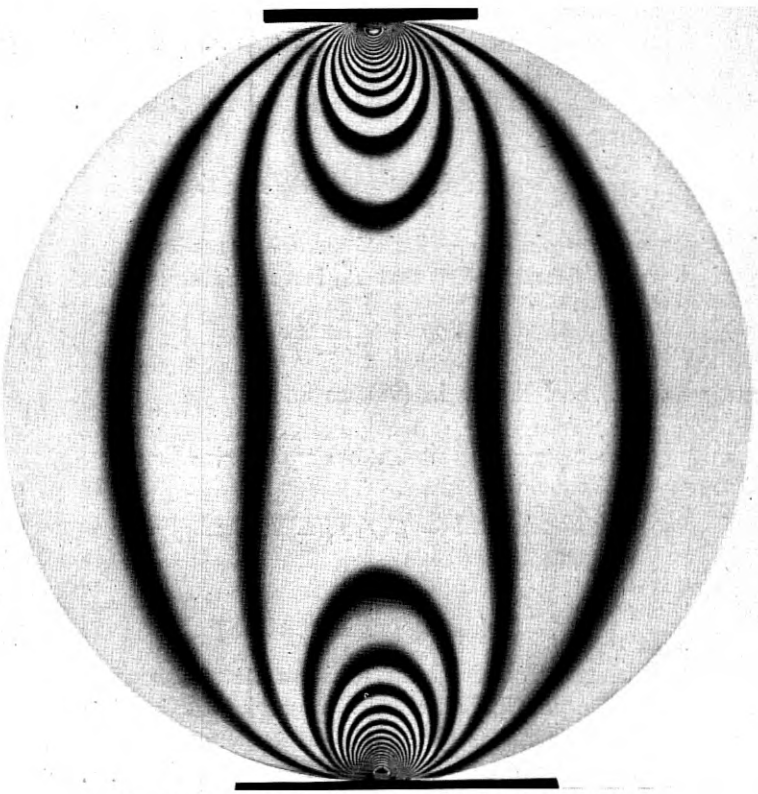


Fig. 3—Photoelastic picture of a disk in compression.

measured in cm² per dyne. A convenient unit for most purposes is one of 10^{-13} cm²/dyne; if this is used, the stress optical coefficients of most glasses are from 1 to 10 and most plastics are from 10 to 100. This unit so defined has been called the "Brewster". In terms of the Brewster, the retardation is

$$r = CTd \tag{104}$$

If C is measured in Brewsters, d in millimeters and T in bars (10^6 dynes/cm²) then r, as given by the formula, is expressed in angstrom units.

Traveling-Wave Tubes

By J. R. PIERCE

[SECOND INSTALLMENT]

CHAPTER IV

FILTER-TYPE CIRCUITS

Synopsis of Chapter

ASIDE FROM HELICES, the circuits most commonly used in traveling-wave tubes are iterated or filter-type circuits, composed of linear arrays of coupled resonant slots or cavities.

Sometimes the geometry of such structures is simple enough so that an approximate field solution can be obtained. In other cases, the behavior of the circuits can be inferred by considering the behavior of lumped-circuit analogues, and the behavior of the circuits with frequency can be expressed with varying degrees of approximation in terms of parameters which can be computed or experimentally evaluated.

In this chapter the field approach will be illustrated for some very simple circuits, and examples of lumped-circuit analogues of other circuits will be given. The intent is to present methods of analyzing circuits rather than particular numerical results, for there are so many possible configurations that a comprehensive treatment would constitute a book in itself.

Readers interested in a wider and more exact treatment of field solutions are referred to the literature.[1,2]

The circuit of Fig. 4.1 is one which can be treated by field methods. This "corrugated waveguide" type of circuit was first brought to the writer's attention by C. C. Cutler. It is composed of a series of parallel equally spaced thin fins of height h projecting normal to a conducting plane. The case treated is that of propagation of a transverse magnetic wave, the magnetic field being parallel to the length of the fins. It is assumed that the spacing ℓ is small compared with a wavelength. In Fig. 4.2, βh is plotted vs. $\beta_0 h$. Here β is the phase constant and $\beta_0 = \omega/c$ is a phase constant corresponding to the velocity of light.

[1] E. L. Chu and W. W. Hansen, "The Theory of Disk-Loaded Wave Guides," *Journal of Applied Physics*, Vol. 18, pp. 999–1008, Nov. 1947.

[2] L. Brillouin, "Wave Guides for Slow Waves," *Journal of Applied Physics*, Vol. 19, pp. 1023–1041, Nov. 1948.

For small values of $\beta_0 h$, that is, at low frequencies, very nearly $\beta = \beta_0$; that is, the phase velocity is very near to the velocity of light. The field decays slowly away from the circuit. The longitudinal electric field is small compared with the transverse electric field. In fact, as the frequency approaches zero, the wave approaches a transverse electromagnetic wave traveling with the speed of light.

At high frequencies the wave falls off rapidly away from the circuit, and the transverse and longitudinal components of electric field are almost equal. The wave travels very slowly. As the wavelength gets so short that the spacing ℓ approaches a half wavelength ($\beta\ell = \pi$) the simple analysis given is no longer valid. Actually, $\beta\ell = \pi$ specifies a cutoff frequency; the circuit behaves as a lowpass filter.

Figure 4.3 shows two opposed sets of fins such as those of Fig. 4.1. Such a circuit propagates two modes, a transverse mode for which the longitudinal electric field is zero at the plane of symmetry and a longitudinal mode for which the transverse electric field is zero at the plane of symmetry.

At low frequencies, the longitudinal mode corresponds to the wave on a loaded transmission line. The fins increase the capacitance between the conducting planes to which they are attached but they do not decrease the inductance. Figure 4.6 shows βh vs. $\beta_0 h$ for several ratios of fin height, h, to half-separation, d. The greater is h/d, the slower is the wave (the larger is β/β_0).

The longitudinal mode is like a transverse magnetic waveguide mode; it propagates only at frequencies above a cutoff frequency, which increases as h/d is increased. Figure 4.7 shows βh vs. $\beta_0 h = (\omega/c)h$ for several values of h/d. The cutoff, for which $\beta\ell = \pi$, occurs for a value of $\beta_0 h$ less than $\pi/2$. Thus, we see that the longitudinal mode has a band pass characteristic. The behavior of the longitudinal mode is similar to that of a longitudinal mode of the washer-loaded waveguide shown in Fig. 4.8. The circuit of Fig. 4.8 has been proposed for use in traveling-wave tubes.

The transverse mode of the circuit of Fig. 4.3 can also exist in a circuit consisting of strips such as those of Fig. 4.1 and an opposed conducting plane, as shown in Fig. 4.5. This circuit is analogous in behavior to the disk-on-rod circuit of Fig. 4.9. The circuit of Fig. 4.5 may be thought of as a loaded parallel strip line. That of Fig. 4.9 may be thought of as a loaded coaxial line.

Wave-analysis makes it possible to evaluate fairly accurately the transmission properties of a few simple structures. However, iterated or repeating structures have certain properties in common: the properties of filter networks.

For instance, a mode of propagation of the loaded waveguide of Fig. 4.10 or of the series of coupled resonators of Fig. 4.11 can be represented accurately at a single frequency by the ladder networks of Fig. 4.12. Further,

if suitable lumped-admittance networks are used to represent the admittances B_1 and B_2, the frequency-dependent behavior of the structures of Figs. 4.10 and 4.11 can be approximated.

It is, for instance, convenient to represent the shunt admittances B_2 and the series admittances B_1 in terms of a "longitudinal" admittance B_L and a "transverse" admittance B_T. B_L and B_T are admittances of shunt resonant circuits, as shown in Fig. 4.15, where their relation to B_1 and B_2 and approximate expressions for their frequency dependence are given. The resonant frequencies of B_L and B_T, that is, ω_L and ω_T, have simple physical meanings. Thus, in Fig. 4.10, ω_L is the frequency corresponding to equal and opposite voltages across successive slots, that is, the π mode frequency. ω_T is the frequency corresponding to zero slot voltage and no phase change along the filter, that is, the zero mode frequency.

If ω_L is greater than ω_T, the phase characteristic of this lumped-circuit analogue is as shown in Fig. 4.17. The phase shift is zero at the lower cutoff frequency ω_T and rises to π at the upper cutoff frequency ω_L. If ω_T is greater than ω_L, the phase shift starts at $-\pi$ at the lower cutoff frequency ω_L and rises to zero at the upper cutoff frequency ω_T, as shown in Fig. 4.19. In this case the phase velocity is negative. Figure 4.20 shows a measure of $(E^2/\beta^2 P)$ plotted vs. ω for $\omega_L > \omega_T$. This impedance parameter is zero at ω_T and rises to infinity at ω_L.

The structure of Fig. 4.11 can be given a lumped-circuit equivalent in a similar manner. In this case the representation should be quite accurate. We find that ω_L is always greater than ω_T and that one universal phase curve, shown in Fig. 4.27, applies. A curve giving a measure of $(E^2/\beta^2 P)$ vs. frequency is shown in Fig. 4.28. In this case the impedance parameter goes to infinity at both cutoff frequencies.

The electric field associated with iterated structures does not vary sinusoidally with distance but it can be analyzed into sinusoidal components. The electron stream will interact strongly with the circuit only if the electron velocity is nearly equal to the phase velocity of one of these field components. If θ is the phase shift per section and L is the section length, the phase constant β_m of a typical component is

$$\beta_m = (\theta + 2m\pi)/L$$

where m is a positive or negative integer. The field component for which $m = 0$ is called the fundamental; for other values of m the components are called *spatial harmonics*. Some of these components have negative phase velocities and some have positive phase velocities.

The peak field strength of any field component may be expressed

$$E = -M(V/L)$$

Here V is the peak gap voltage, L is the section spacing and M is a function of β (or β_m) and of various dimensions. For the electrode systems of Figs.

4.29, 4.30, 4.31 and 4.32 M is given by (4.69), (4.71), (4.72) and (4.73), respectively.

The factor M may be indifferently regarded as a factor by which we multiply the a-c beam current to give the induced current at the gap, or, as a factor by which we multiply the gap voltage in obtaining the field. We can go further, evaluate $E^2/\beta^2 P$ in terms of gap voltage, and use $M^2 I_0$ as the effective current, or we can use the current I_0 and take the effective field in the impedance parameter as

$$E^2 = M^2(V/\ell)^2$$

It is sometimes desirable to make use of a spatial harmonic ($m \neq 0$) instead of a fundamental, usually to (1) allow a greater resonator spacing (2) to obtain a positive phase velocity when the fundamental has a negative phase velocity (3) to obtain a phase curve for which the phase angle is nearly a constant times frequency; that is, a phase curve for which the group velocity does not change much with frequency and hence can be matched by the electron velocity over a considerable frequency range. Figure 4.33 shows how $\theta + 2\pi$ (the phase shift per section for $m = 1$) can be nearly a constant times ω even when θ is not.

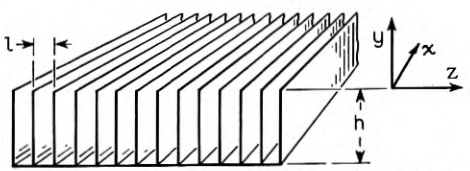


Fig. 4.1—A corrugated or finned circuit with filter-like properties.

4.1 Field Solutions

An approximate field analysis will be made for two very simple two-dimensional structures. The first of these, which is shown in Fig. 4.1, is empty space for $y > 1$ and consists of very thin conducting partitions in the y direction from $y = 0$ to $y = -h$; the partitions are connected together by a conductor in the z direction at $y = -h$. These conducting partitions are spaced a distance ℓ apart in the z direction. The structure is assumed to extend infinitely in the $+x$ and $-x$ directions.

In our analysis we will initially assume that the wavelength of the propagated wave is long compared with ℓ. In this case, the effect of the partitions is to prevent the existence of any y component of electric field below the z axis, and the conductor at $y = -h$ makes the z component of electric field zero at $y = -z$.

In some perfectly conducting structures the waves propagated are either transverse electric (no electric field component in the direction of propagation, that is, z direction) or transverse magnetic (no magnetic field com-

ponent in the z direction). We find that for the structure under consideration there is a transverse magnetic solution. We can take it either on the basis of other experience or as a result of having solved the problem that the correct form for the x component of magnetic field for $y > 0$ is

$$H_x = H_0 e^{(-\gamma y - j\beta z)} \tag{4.1}$$

Expressing the electric field in terms of the curl of the magnetic field, we have

$$j\omega\epsilon E_x = \frac{\partial Hz}{\partial y} - \frac{\partial Hy}{\partial z} = 0$$

$$j\omega\epsilon E_y = \frac{\partial Hx}{dz} - \frac{\partial Hz}{\partial x} \tag{4.2}$$

$$E_y = -\frac{\beta}{\omega\epsilon} H_0 e^{(-\gamma y - j\beta z)} \tag{4.3}$$

$$j\omega\epsilon E_z = \frac{\partial Hy}{\partial x} - \frac{\partial Hx}{\partial y} \tag{4.4}$$

$$E_z = -j\frac{\gamma}{\omega\epsilon} H_0 e^{(-\gamma y - j\beta z)} \tag{4.5}$$

We can in turn express H_x in terms of E_y and E_z

$$-j\omega\mu H_x = \frac{\partial Ez}{\partial y} - \frac{\partial Ey}{\partial z} \tag{4.6}$$

This leads to the relation

$$\beta^2 - \gamma^2 = \omega^2\mu\epsilon \tag{4.7}$$

Now, $1/\sqrt{\mu\epsilon}$ is the velocity of light, and ω divided by the velocity of light has been called β_0, so that

$$\beta^2 - \gamma^2 = \beta_0^2 \tag{4.8}$$

Between the partitions, the field does not vary in the z direction. In any space between from $y = 0$ to $y = -h$, the appropriate form for the magnetic field is

$$H_x = H_0 \frac{\cos \beta_0(y + h)}{\cos \beta_0 h} \tag{4.9}$$

From this we obtain by means of (4.4)

$$E_z = -\frac{j\beta_0}{\omega\epsilon} H_0 \frac{\sin \beta_0(y + h)}{\cos \beta_0 h} \tag{4.10}$$

Application of (4.6) shows that this is correct.

Now, at $y = 0$ we have just above the boundary

$$E_z = -j \frac{\gamma}{\omega\epsilon} H_0 e^{-j\beta z} \tag{4.11}$$

The fields in the particular slot just below the boundary will be in phase with these (we specify this by adding a factor exp $-j\beta z$ to 4.10) and hence will be

$$E_z = -\frac{j\beta_0}{\omega\epsilon} H_0 e^{-j\beta z} \tan \beta_0 h \tag{4.12}$$

From (4.11) and (4.12) we see that we must have

$$\beta_0 h \tan \beta_0 h = \gamma h \tag{4.13}$$

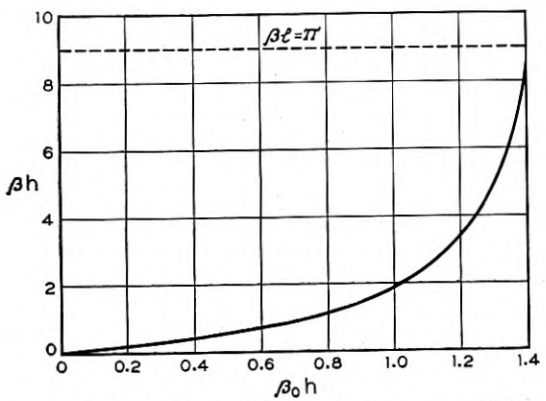


Fig. 4.2—The approximate variation of the phase constant β with frequency (proportional to $\beta_0 h$) for the circuit of Fig. 4.1. The curve is in error as $\beta\ell$ approaches π, and there is a cutoff at $\beta\ell = \pi$.

Using (4.8), we obtain

$$\beta h = \frac{\pm\beta_0 h}{\cos \beta_0 h} \tag{4.14}$$

In Fig. 4.2, βh has been plotted vs $\beta_0 h$, which is, of course, proportional to frequency. This curve starts out as a straight line, $\beta = \beta_0$; that is, for low frequencies the speed is the speed of light. At low frequencies the field falls off slowly in the y direction, and as the frequency approaches zero we have essentially a plane electromagnetic wave. At higher frequencies, $\beta > \beta_0$, that is, the wave travels with less than the speed of light, and the field falls off rapidly in the y direction. According to (4.14), β goes to infinity at $\beta_0 h = \pi/2$.

As a matter of fact, the match between the fields assumed above and below the boundary becomes increasingly bad as $\beta\ell$ becomes larger. The most rapid

alteration we can have below the boundary is one in which fields in alternate spaces follow a $+, -, +, -$ pattern. Thus, the rapid variations of field above the boundary predicted by (4.14) for values of $\beta_0 h$ which make $\beta \ell$ greater than π cannot be matched below the boundary. The frequency at which $\beta \ell = \pi$ constitutes the cutoff frequency of the structure regarded as a filter. There is another pass band in the region $\pi < \beta_0 h < 3\pi/2$, in which the ratio of E to H below the boundary has the same sign as the ratio of E to H above the boundary.

A more elaborate matching of fields would show that our expression is considerably in error near cutoff. This matter will not be pursued here; the behavior of filters near cutoff will be considered in connection with lumped circuit representations.

We can obtain the complex power flow P by integrating the Poynting vector over a plane normal to the z direction in the region $y > 0$. Let us consider the power flow over a depth W normal to the plane of the paper. Then

$$P = \frac{1}{2} \int_0^\infty \int_0^W (E_x H_y^* - E_y H_x^*) \, dx \, dy \tag{4.15}$$

Using (4.1) and (4.3), we obtain

$$\begin{aligned} P &= \frac{W}{2} \int_0^\infty \frac{\beta H_0^2}{\omega \epsilon} e^{-2\gamma y} \, dy \\ P &= \frac{1}{4} \frac{H_0^2 \beta W}{\omega \epsilon \gamma} \end{aligned} \tag{4.16}$$

We will express this in terms of E the magnitude of the z component of the field at $y = 0$, which, according to (4.5), is

$$E = \frac{\gamma}{\omega \epsilon} H_0 \tag{4.17}$$

We will also note that

$$\begin{aligned} \omega \epsilon &= \omega \sqrt{\mu \epsilon} / \sqrt{\mu / \epsilon} \\ &= (\omega / c) / \sqrt{\mu / \epsilon} = \beta_0 / \sqrt{\mu / \epsilon} \end{aligned} \tag{4.18}$$

and that

$$\sqrt{\mu / \epsilon} = 377 \text{ ohms} \tag{4.19}$$

By using (4.17)–(4.18) in connection with (4.16), we obtain

$$E^2 / \beta^2 P = (4 / \beta_0 W)(\gamma / \beta)^3 \sqrt{\mu / \epsilon} \tag{4.20}$$

We notice that this impedance is very small for low frequencies, at which

the velocity of the wave is high, and the field extends far in the y direction and becomes higher at high frequencies, where the velocity is low and the field falls off rapidly.

We will next consider a symmetrical array of two opposed sets of slots (Fig. 4.3) similar to that shown in Fig. 4.1. Two modes of propagation will be of interest. In one the field is symmetrical about the axis of physical symmetry, and in the other the fields at positions of physical symmetry are equal and opposite.

In writing the equations, we need consider only half of the circuit. It is convenient to take the z axis along the boundary, as shown in Fig. 4.4.

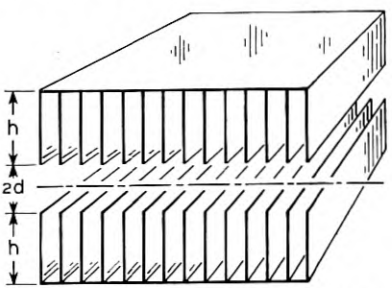


Fig. 4.3—A double finned structure which will support a transverse mode (no longitudinal electric field on axis) and a longitudinal mode (no transverse electric field on axis).

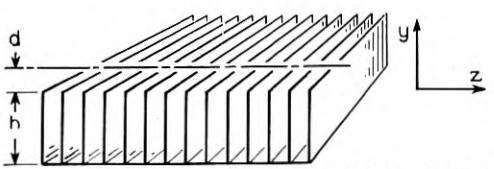


Fig. 4.4—The coordinates used in connection with the circuit of Fig. 4.3.

This puts the axis of symmetry at $y = +d$, and the slots extend from $y = 0$ to $y = -h$.

For negative values of y, (4.9), (4.10), (4.12) hold.

Let us first consider the case in which the fields above are opposite to the fields below. This also corresponds to waves in a series of slots opposite a conducting plane, as shown in Fig. 4.5. In this case the appropriate form of the magnetic field above the boundary is

$$H_x = H_0 \frac{\cosh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.21}$$

From Maxwell's equations we then find

$$E_y = -\frac{\beta}{\omega\epsilon} H_0 \frac{\cosh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.22}$$

$$E_z = -j\frac{\gamma}{\omega\epsilon} H_0 \frac{\sinh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.23}$$

$$\beta_0^2 = \beta^2 - \gamma^2 \tag{4.24}$$

At $y = 0$ we have from (4.23) and (4.12)

$$E_z = -j\frac{\gamma}{\omega\epsilon} H_0 e^{-j\beta z} \tanh \gamma d \tag{4.25}$$

$$E_z = -j\frac{\beta_0}{\omega\epsilon} H_0 e^{-j\beta z} \tan \beta_0 h \tag{4.12}$$

Hence, we must have

$$\gamma h \tanh ((d/h)\gamma h) = \beta_0 h \tan \beta_0 h \tag{4.26}$$

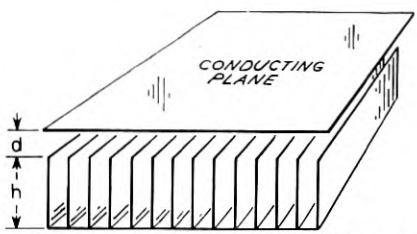


Fig. 4.5—The transverse mode of the circuit of Fig. 4.3 exists in this circuit also.

Here we have added parameter, (d/h). For any value of d/h, we can obtain γh vs $\beta_0 h$; and we can obtain βh in terms of γh by means of 4.24

$$\beta h = ((\gamma h)^2 + (\beta_0 h)^2)^{1/2} \tag{4.27}$$

We see that for small values of $\beta_0 h$ (low frequencies)

$$\gamma^2 = (h/d)\,\beta_0^2 \tag{4.28}$$

$$\beta = \beta_0 \left(\frac{h + d}{d}\right)^{1/2} \tag{4.29}$$

If we examine Fig. 4.5, to which this applies, we find (4.28) easy to explain. At low frequencies, the magnetic field is essentially constant from $y = d$ to $y = -h$, and hence the inductance is proportional to the height $h + d$. The electric field will, however, extend only from $y = 0$ to $y = d$; hence the capacitance is proportional to $1/d$. The phase constant is proportional to $\sqrt{LC}$, and hence (4.29). At higher frequencies the electric and magnetic fields vary with y and (4.29) does not hold.

We see that (4.26) predicts infinite values of γ for $\beta h = \pi/2$. As in the previous cases, cutoff occurs at $\beta \ell = \pi$.

As an example of the phase characteristic of the circuit, βh from (4.26) and (4.27) is plotted vs $\beta_0 h$ for $h/d = 0, 10, 100$ in Fig. 4.6. The curve for $h/d = 0$ is of course the same as Fig. 4.2.

If we integrate Poynting's vector from $y = 0$ to $y = d$ and for a distance W in the x direction, and multiply by 2 to take the power flow in the other half of the circuit into account, we obtain

$$E^2/\beta^2 P = (2/\beta_0 W)(\gamma/\beta)^3 \left(\frac{\sinh^2 \gamma d}{\sinh \gamma d \cosh \gamma d + \gamma d} \right) \sqrt{\mu/\epsilon} \quad (4.30)$$

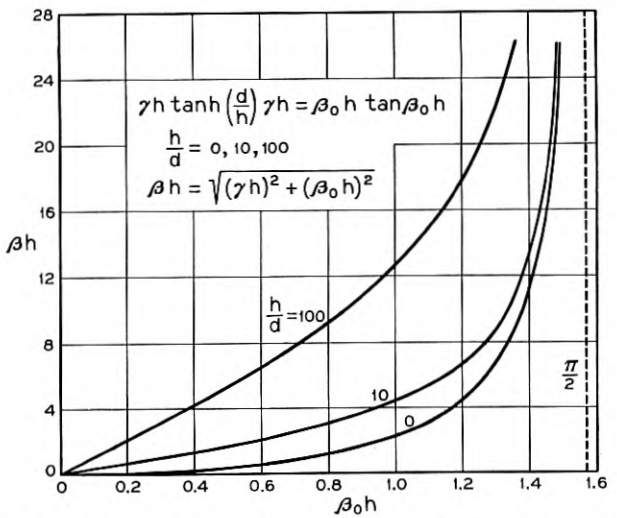


Fig. 4.6—The variation of β with frequency (proportional to $\beta_0 h$) for the transverse mode of the circuit of Fig. 4.3. Again, the curves are in error near the cutoff at $\beta\ell = \pi$.

At very low frequencies, at which (4.28) and (4.29) hold, we have

$$\begin{aligned} E^2/\beta^2 P &= (\gamma^4/\beta_0 \beta^3)(d/W) \sqrt{\mu/\epsilon} \\ E^2/\beta^2 P &= (h/d)^{1/2} (1 + d/h)^{3/2} (d/W) \sqrt{\mu/\epsilon} \end{aligned} \quad (4.31)$$

At high frequencies, for which γd is large, (4.30) approaches $\frac{1}{2}$ of the value given by (4.20). There is twice as much power because there are two halves to the circuit.

Let us now consider the case in which the field is symmetrical and E_z does not go to zero on the axis. In this case the appropriate field for $y > 0$ is

$$H_x = H_0 \frac{\sinh \gamma(d - y)}{\sinh \gamma d} e^{-j\beta z} \quad (4.32)$$

Proceeding as before, we find

$$\frac{\gamma h}{\tanh\left((d/h)\,\gamma h\right)} = \beta_0 h \tan \beta_0 h \tag{4.33}$$

We see that, in this case, for small values of γh we have

$$\beta_0 h \tanh \beta_0 h = h/d \tag{4.33a}$$

There is no transmission at all for frequencies below that specified by (4.33). As the frequency is increased above this lower cutoff frequency, γh and hence βh increase, and approach infinity at $\beta_0 h = \pi/2$. Actually, of course, the upper cutoff occurs at $\beta \ell = \pi$. In Fig. 4.7 βh is plotted vs $\beta_0 h$ for $h/d = 0$,

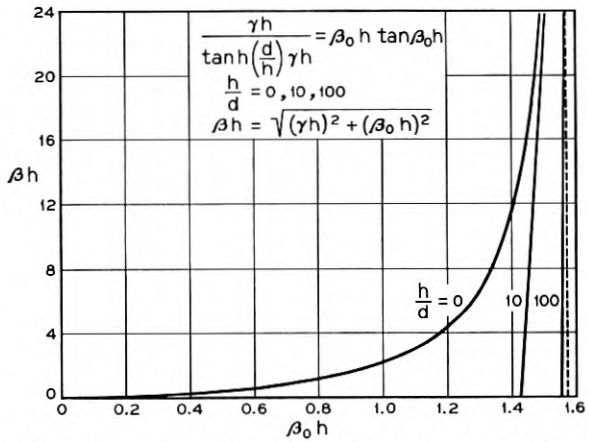


Fig. 4.7—The variation of β with frequency (proportional to $\beta_0 h$) for the longitudinal mode of the circuit of Fig. 4.3. This mode has a band pass characteristic; the band narrows as the opening of width 2d is made small compared with the fin height. Again, the curves are in error near the upper cutoff at $\beta \ell = \pi$.

10, 100. This illustrates how the band is narrowed as the opening between the slots is decreased.

By the means used before we obtain

$$E^2/\beta^2 P = (2/\beta_0 W)(\gamma/\beta)^3 \left(\frac{\cosh^2 \gamma d}{\sinh \gamma d \cosh \gamma d - \gamma d}\right) \sqrt{\mu/\epsilon} \tag{4.34}$$

We see that this goes to infinity at $\gamma d = 0$. For large values of γd it becomes the same as (4.30).

4.2 Practical Circuits

Circuits have been proposed or used in traveling-wave tubes which bear a close resemblance to those of Figs. 4.1, 4.3, 4.5 and which have very similar

properties[3]. Thus Field[4] describes an apertured disk structure (Fig. 4.8) which has band-pass properties very similar to the symmetrical mode of the circuit of Fig. 4.3. In this case there is no mode similar to the other mode, with equal and opposite fields in the two halves. Field also shows a disk-on-rod structure (Fig. 4.9) and describes a tube using it. This structure has low-

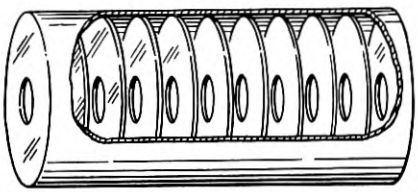


Fig. 4.8—This loaded waveguide circuit has band-pass properties similar to those of Fig. 4.7.

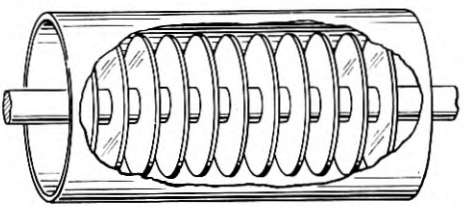


Fig. 4.9—This disk-on-rod circuit has properties similar to those of Fig. 4.6.

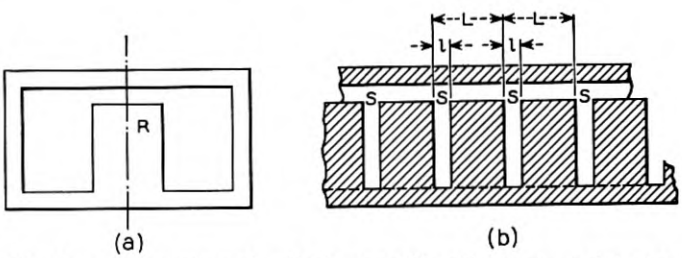


Fig. 4.10—A circuit consisting of a ridged waveguide with transverse slots or resonators in the ridge.

pass properties very similar to those of the circuit of Fig. 4.5, which are illustrated in Fig. 4.6.

Figure 4.10 shows a somewhat more complicated circuit. Here we have a rectangular waveguide, shown end on in *a* of Fig. 4.10, loaded by a longitudinal ridged portion *R*. In *b* of Fig. 4.10 we have a longitudinal cross sec-

[3] F. B. Llewellyn, *U. S. Patents* 2,367,295 and 2,395,560.

[4] Lester M. Field, "Some Slow-Wave Structures for Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 37, pp. 34–40, Jan. 1949.

tion, showing regularly spaced slots S cut in the ridge R. The slots S may be thought of as resonators.

Figure 4.11 shows in cross section a circuit made of a number of axially symmetrical reentrant resonators R, coupled by small holes H which act as inductive irises.

It would be very difficult to apply Maxwell's equations directly in deducing the performance of the structures shown in Figs. 4.10 and 4.11. Moreover, it is apparent that we can radically change the performance of

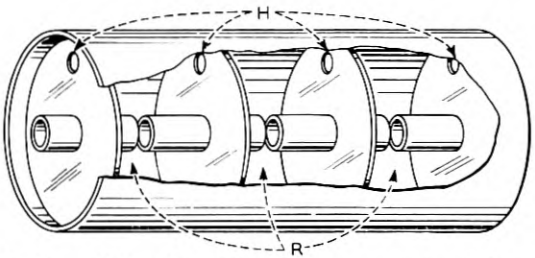


Fig. 4.11—A circuit consisting of a number of resonators inductively coupled by means of holes.

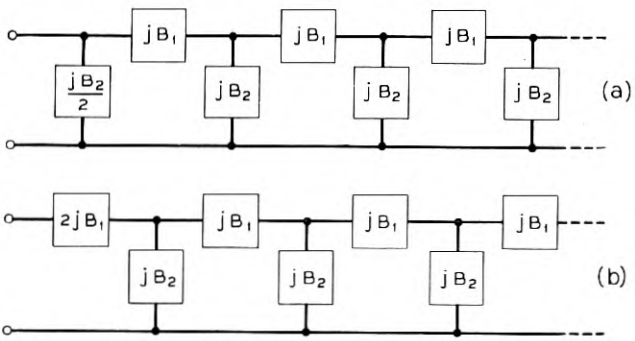


Fig. 4.12—Ladder networks terminated in π (above) and T (below) half sections. Such networks can be used in analyzing the behavior of circuits such as those of Figs. 4.10 and 4.11.

such structures by minor physical alterations as, by changing the iris size, or by using resonant irises in the circuit of Fig. 4.11, for instance.

As a matter of fact, it is not necessary to solve Maxwell's equations afresh each time in order to understand the general properties of these and other circuits.

4.3 Lumped Iterated Analogues

Consider the ladders of lossless admittances or susceptances shown in Fig. 4.12. Susceptances rather than reactances have been chosen because the

elements we shall most often encounter are shunt resonant near the frequencies considered; their susceptance is near zero and changing slowly but their reactance is near infinity.

If these ladders are continued endlessly to the right (or terminated in a reflectionless manner) and if a signal is impressed on the left-hand end, the voltages, currents and fields at corresponding points in successive sections will be in the ratio exp(-Γ) so that we can write the voltages,

$$V_n = V_0 e^{-n\Gamma} \tag{4.35}$$

If the admittances Y_1 and Y_2 are pure susceptances (lossless reactors), Γ is either purely real (an exponential decay with distance) or purely imaginary (a pass band). In this case Γ is usually replaced by $j\beta$. In order to avoid confusion of notation, we will use $j\theta$ instead, and write for the lossless case in the pass band

$$V_n = V_0 e^{-jn\theta} \tag{4.35a}$$

Thus, θ is the phase lag in radians in going from one section to the next. In terms of the susceptances,*

$$\cos\theta = 1 + B_2/2B_1 \tag{4.36}$$

We will henceforward assume that all elements are lossless.

Two characteristic impedances are associated with such iterated networks. If the network starts with a shunt susceptance $B_1/2$, as in *a* of Fig. 4.12, then we see the mid-shunt characteristic impedance K_π

$$K_\pi = 2(-B_2(B_2 + 4B_1))^{-1/2} \tag{4.37}$$

If the network starts with a series susceptance $2B_1$ we see the mid-series characteristic impedance K_T

$$K_T = \pm(1/2B_1)(-B_2 + 4B_1)/B_2)^{1/2} \tag{4.38}$$

Here the sign is chosen to make the impedance positive in the pass band.

When such networks are used as circuits for a traveling-wave tube, the voltage acting on the electron stream may be the voltage across B_2 or the voltage across B_1 or the voltage across some capacitive element of B_2 or B_1. We will wish to relate this peak voltage V to the power flow P. If the voltage across B_2 acts on the electron stream

$$V^2/P = 2K_\pi \tag{4.39}$$

If the voltage across Y_1 acts on the electron stream

$$V = I/jB_1$$

* The reader can work such relations out or look them up in a variety of books or handbooks. They are in Schelkunoff's Electromagnetic Waves.

where I is the current in B_1

$$P = |I^2| K_T/2$$

and hence

$$V^2/P = 2/B_1^2 K_T \tag{4.40}$$

$$V^2/P = -4(B_2/B_1)(-B_2(B_2 + 4B_1))^{-1/2} \tag{4.41}$$

$$V^2/P = -2(B_2/B_1)K_\pi \tag{4.42}$$

Here the sign has been chosen so as to make V^2/P positive in the pass band.

Let us now consider as an example the structure of Fig. 4.10. We see that two sorts of resonance are possible. First, if all the slots are shorted, or if no voltage appears between them, we can have a resonance in which the field between the top of the ridge R and the top of the waveguide is constant

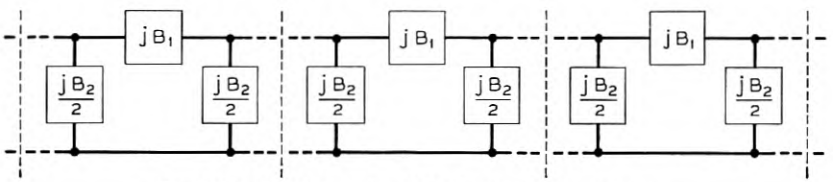


Fig. 4.13—A ladder network broken up into π sections.

all along the length, and corresponds to the cutoff frequency of the ridged waveguide. There are no longitudinal currents (or only small ones near the slots S) and hence there is no voltage across the slots and their admittance (the slot depth, for instance) does not affect the frequency of this resonance. Looking at Fig. 4.12, we see that this corresponds to a condition in which all shunt elements are open, or $B_2 = 0$. We will call the frequency of this resonance ω_T, the T standing for transverse.

There is another simple resonance possible; that in which the fields across successive slots are equal and opposite. Looking at Fig. 4.12, we see that this means that equal currents flow into each shunt element from the two series elements which are connected to it. We could, in fact, divide the network up into unconnected π sections, associating with each series element of susceptance B_1 half of the susceptance of a shunt element, that is, $B_2/2$, at each end, as shown in Fig. 4.13, without affecting the frequency of this resonance. This resonance, then, occurs at the frequency ω_L (L for longitudinal) at which

$$B_1 + B_2/4 = 0. \tag{4.43}$$

We have seen that the transverse resonant frequency, ω_T, has a clear meaning in connection with the structure of Fig. 4.10; it is (except for small

errors due to stray fields near the slots) the cutoff frequency of the waveguide without slots. Does the longitudinal frequency ω_L have a simple meaning?

Suppose we make a model of one section of the structure, as shown in Fig. 4.14. Comparing this with *b* of Fig. 4.10, we see that we have included the section of the ridged portion between two slots, and one half of a slot at each end, and closed the ends off with conducting plates *C*. The resonant frequency of this model is ω_L, the longitudinal resonant frequency defined above.

We will thus liken the structure of Fig. 4.10 to the filter network of Fig.

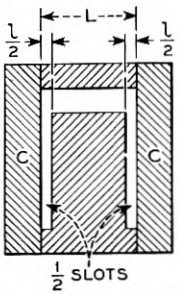


Fig. 4.14—A section which will have a resonant frequency corresponding to that for π radians phase shift per section in the circuit of Fig. 4.10.

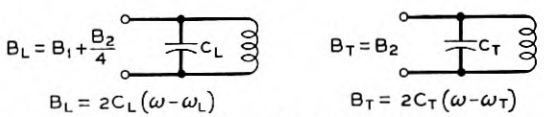


Fig. 4.15—The approximate variation with frequency (over a narrow band) of the longitudinal (B_L) transverse (B_T) susceptances of a filter network.

4.12, and express the susceptances B_1 and B_2 in terms of two susceptances B_T and B_L associated with the transverse and longitudinal resonances and defined below

$$B_T = B_2 \tag{4.44}$$

$$B_L = B_1 + B_2/4 \tag{4.45}$$

At the transverse resonant frequency ω_T, $B_T = 0$, and at the longitudinal resonant frequency ω_L, $B_L = 0$. So far, the lumped-circuit representation of the structure of Fig. 4.14 can be considered exact in the sense that at any frequency we can assign values to B_T and B_L which will give the correct values for θ and for V^2/P for the voltage across either the shunt or the series elements (whichever we are interested in).

We will go further and assume that near resonances these values of B_T and B_L behave like the admittances of shunt resonant circuits, as indicated in Fig. 4.15. Certainly we are right by our definition in saying that $B_T = 0$ at ω_T, and $B_L = 0$ at ω_L. We will assume near these frequencies a linear variation of B_T and B_L with frequency, which is very nearly true for shunt resonant circuits near resonance*

$$B_T = 2C_T(\omega - \omega_T) \tag{4.46}$$

$$B_L = 2C_L(\omega - \omega_L) \tag{4.47}$$

Here C_T can mean twice the peak stored electric energy per section length for unit peak voltage between the top of the guide and the top of the ridge R when the structure resonates in the transverse mode, and C_L can mean twice the stored energy per section length L for unit peak voltage across the top

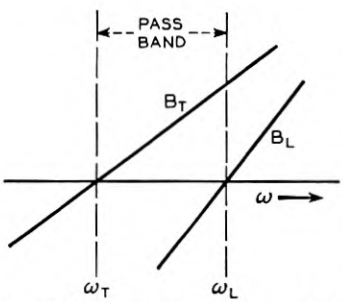


Fig. 4.16—Longitudinal and transverse susceptances which give zero radians phase shift at the lower cutoff ($\omega = \omega_T$) and π radians phase shift at the upper cutoff ($\omega = \omega_L$).

of the slot when the structure resonates in the longitudinal mode.

In terms of B_T and B_L, expression (4.36) for the phase angle θ becomes

$$\cos\theta = \frac{4B_L + B_T}{4B_L - B_T} \tag{4.48}$$

We see immediately that for real values of θ ($\cos\theta \leq 1$), B_T and B_L must have opposite signs, making the denominator greater than the numerator.

Figure 4.16 shows one possible case, in which $\omega_T < \omega_L$. In this case the pass band (θ real) starts at the lower cutoff frequency $\omega = \omega_T$ at which B_T is zero, $\cos\theta = 1$ (from (4.48)) and $\theta = 0$, and extends up to the upper cutoff frequency $\omega = \omega_L$ at which $B_L = 0$, $\cos\theta = -1$ and $\theta = \pi$.

* In case the filter has a large fractional bandwidth, it may be worth while to use the accurate lumped-circuit forms

$$B_T = \omega_T C_T(\omega/\omega_T - \omega_T/\omega) \tag{4.46a}$$

$$B_L = \omega_L C_L(\omega/\omega_L - \omega_L/\omega) \tag{4.46b}$$

The shape of the phase curves will depend on the relative rates of variation of B_T and B_L with frequency. Assuming the linear variations with frequency of (4.46) and (4.47) the shapes can be computed. This has been done for $C_L/C_T = 1, 3, 10$ and the results are shown in Fig. 4.17.

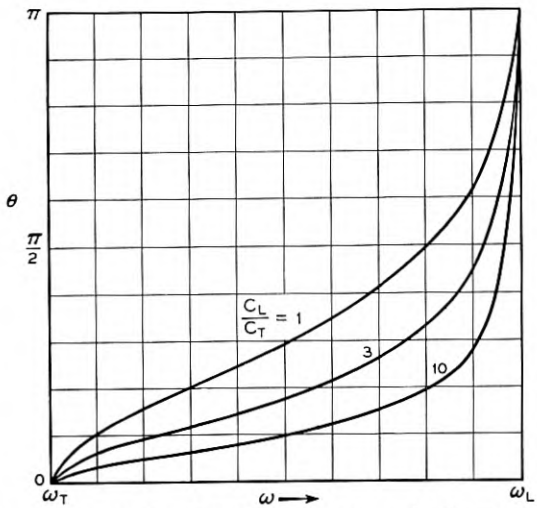


Fig. 4.17—Phase shift per section, θ, vs radian frequency ω for the conditions of Fig. 4.16.

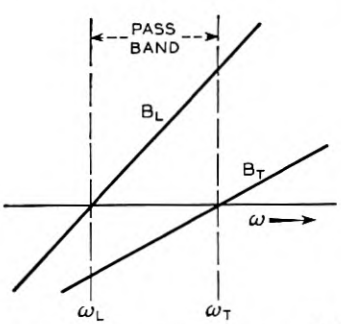


Fig. 4.18—Longitudinal and transverse susceptances which give $-\pi$ radians phase shift at the lower cutoff ($\omega = \omega_L$) and 0 degrees phase shift at the upper cutoff ($\omega = \omega_T$). This means a negative phase velocity.

It is of course possible to make $\omega_L > \omega_T$. In this case the situation is as shown in Fig. 4.18, the pass band extending from ω_L to ω_T. At $\omega = \omega_L$, $\cos\theta = -1$, $\theta = -\pi$. At $\omega = \omega_T$, $\cos\theta = 1$ and $\theta = 0$. In Fig. 4.19, assuming (4.46) and (4.47), θ has been plotted vs ω for $C_L/C_T = 1, 3, 10$.

The curves of Figs. 4.17 and 4.18 are not exact for any physical structure of the type shown in Fig. 4.10. In lumped circuit terms, they neglect coupling

between slots. They will be most accurate for structures with slots longitudinally far apart compared with the transverse dimensions, and least accurate for structures with slots close together. They do, however, form a valuable guide in understanding the performance of such structures and in evaluating the effect of the ratio of energies stored in the fields at the two cutoff frequencies.

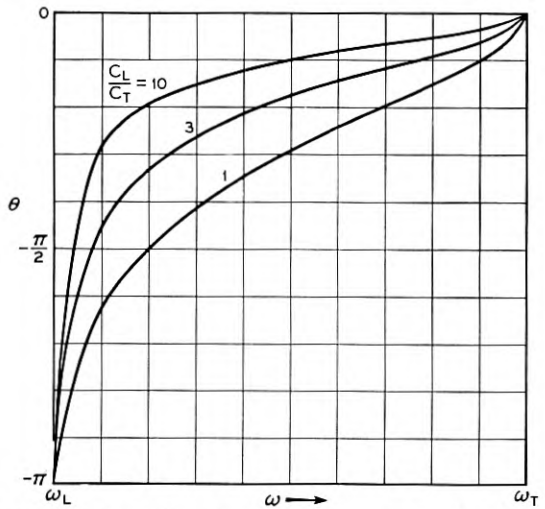


Fig. 4.19—Phase shift per section, θ, vs radian frequency, ω, for the conditions of Fig 4.18.

It is most likely that the voltages across the slots would be of most interest in connection with the circuit shown in Fig. 4.10. We can rewrite (4.41) in terms of B_T and B_L

$$V^2/P = \frac{1}{2(1 - 4B_L/B_T)(-B_T B_L)^{1/2}} \tag{4.49}$$

We see that V^2/P goes to 0 at $B_T = 0$ ($\omega = \omega_T$) and to infinity at $B_L = 0$ ($\omega = \omega_L$). In Fig. 4.20 assuming (4.46) and (4.47), $(V^2/P)(\omega_L C_L \omega_T C_T)$ is plotted vs ω for $C_L/C_T = 1, 3, 10$.

Let us consider another circuit, that shown in Fig. 4.11. We see that this consists of a number of resonators coupled together inductively. We might draw the equivalent circuits of these resonators as shown in Fig. 4.21. Here L and C are the effective inductance and the effective capacitance of the resonators without irises. They are chosen so that the resonant frequency ω_0 is given by

$$\omega_0 = \sqrt{LC} \tag{4.50}$$

and the variation of gap susceptance B with frequency is

$$\partial B/\partial\omega = 2C \tag{4.51}$$

The arrows show directions of current flow when the currents in the gap capacitances are all the same.

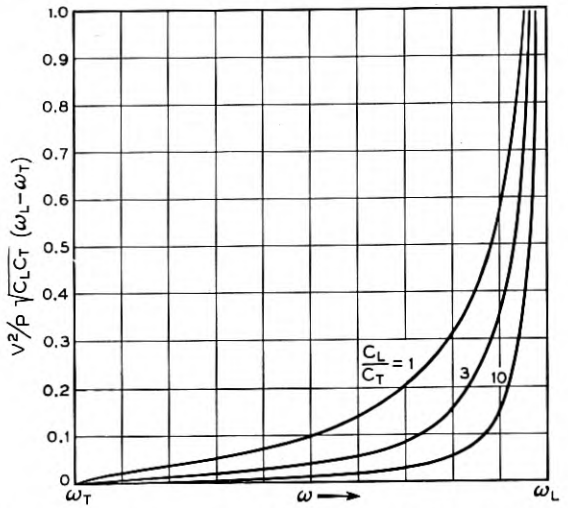


Fig. 4.20—A quantity proportional to $(E^2/\beta^2 P)$ vs ω for the conditions of Figs. 4.16 and 4.17.

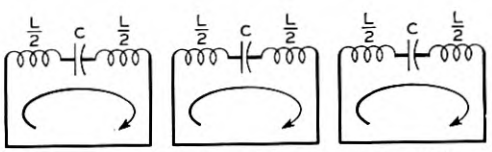


Fig. 4.21—A representation of the resonators of Fig. 4.11.

We can now represent the circuit of Fig. 4.11 by interconnecting the circuits of Fig. 4.21 by means of inductances L_M of Fig. 4.22. This gives a suitable representation, but one which is open to a minor objection: the gap capacitance does not appear across either a shunt or a series arm.

It is important to notice that there is another equally good representation, and there are probably many more. Suppose we draw the resonators as shown in Fig. 4.23 instead of as in Fig. 4.21. The inductance L and capacitance C are still properly given by 4.50 and 4.51. We can now interconnect the resonators inductively as shown in Fig. 4.24.

We should note one thing. In Fig. 4.21, the currents which are to flow in the common inductances of Fig. 4.22 flow in opposite directions when the

gap currents are in the same directions. In the representation of Fig. 4.23 the currents which will flow in the common inductances of Fig. 4.24 have been drawn in opposite directions, and we see that the currents in the gap capacitances flow alternately up and down. In other words, in Fig. 4.24, every other gap appears inverted. This can be taken into account by adding a phase angle $-\pi$ to θ as computed from (4.48).

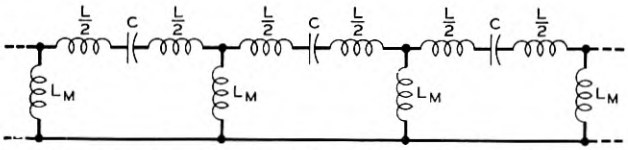


Fig. 4.22—The resonators of Fig. 4.11 coupled inductively.

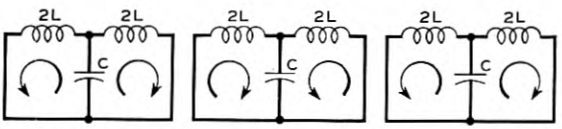


Fig. 4.23—Another representation of the resonators of Fig. 4.11.

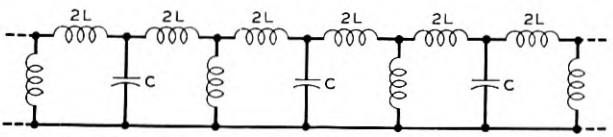


Fig. 4.24—Figure 4.23 with inductive coupling added.

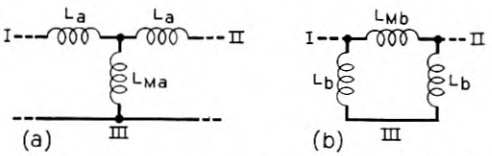


Fig. 4.25—A $T - \pi$ transformation used in connection with the circuit of Fig. 4.24.

Now, the T configuration of inductances in a of Fig. 4.25 can be replaced by the π configuration, b of Fig. 4.25. Imagine I and II to be connected together and a voltage to be applied between them and III. We see that

$$L_b = L_a + 2L_{Ma} \tag{4.52}$$

Imagine a voltage to be applied between I and II. We see that

$$1/L_a = 1/L_b + 2/L_{Mb} \tag{4.53}$$

If $L_{Ma} \ll L_a$, then L_b will be nearly equal to L_a and $L_{Mb} \gg L_b$.

By means of such a $T - \pi$ transformation we can redraw the equivalent circuit of Fig. 4.24 as shown in Fig. 4.26. The series susceptance B_1 is now

that of L_1, and the shunt susceptance is now that of the shunt resonant circuit consisting of C_2 (the effective capacitance of the resonators) and L_2.

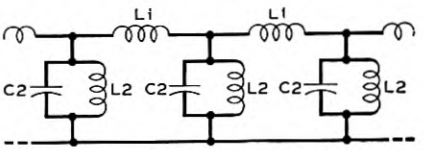


Fig. 4.26—The final representation of the circuit of Fig. 4.11.

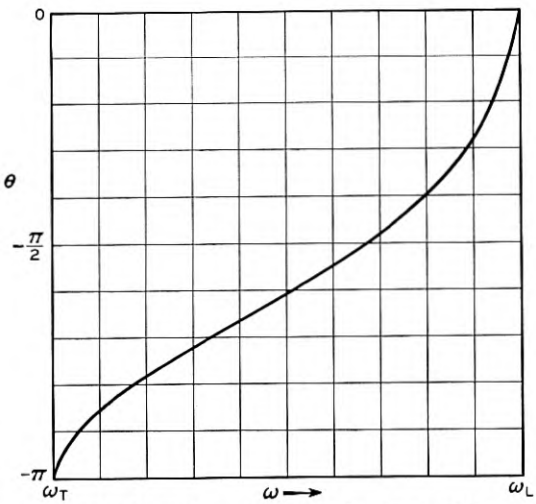


Fig. 4.27—The phase characteristic of the circuit of Fig. 4.11.

The transverse resonance, $B_2 = 0$, occurs at a frequency

$$\omega_T = \sqrt{C_2 L_2} \tag{4.54}$$

Near this frequency the transverse susceptance is given by

$$B_T = 2C_2(\omega - \omega_T) \tag{4.55}$$

The longitudinal resonance occurs at a frequency

$$\omega_L = \sqrt{2C_2 L_1 L_2/(L_1 + 2L_2)} \tag{4.56}$$

and near ω_L,

$$B_L = C_2(\omega - \omega_L) \tag{4.57}$$

These are just the forms we found in connection with the structure of Fig. 4.10; but we see that, in the case of the circuit of Fig. 4.11, the effective transverse capacitance is always twice the effective longitudinal capacitance ($C_L/C_T = 1/2$ in Fig. 4.19), and that $\omega_L > \omega_T$ for attainable volume of L_1.

We obtain θ vs ω by adding $-\pi$ to the phase angle from 4.48, using (4.55) and (4.57) in obtaining B_T and B_L. The phase angle vs. frequency is shown in Fig. 4.27. As the irises are made larger, the bandwidth, $\omega_L - \omega_T$, becomes larger, largely by a decrease in ω_L.

The voltage of interest is that across C_2, that is, that across the gap. From (4.37), (4.44), (4.45), (4.55) and (4.57) we obtain

$$V^2/P = 2/(-B_T B_L)^{1/2} \tag{4.58}$$

$$V^2/P = (\sqrt{2}/C_2)((\omega_L - \omega)(\omega - \omega_T))^{-1/2} \tag{4.59}$$

This goes to infinity at both $\omega = \omega_L$ and $\omega = \omega_T$. In Fig. 4.28, $(V^2/P)C_2\sqrt{\omega_L \omega_T}$ is plotted vs ω. This curve represents the performance of all narrow band structures of the type shown in Fig. 4.11.

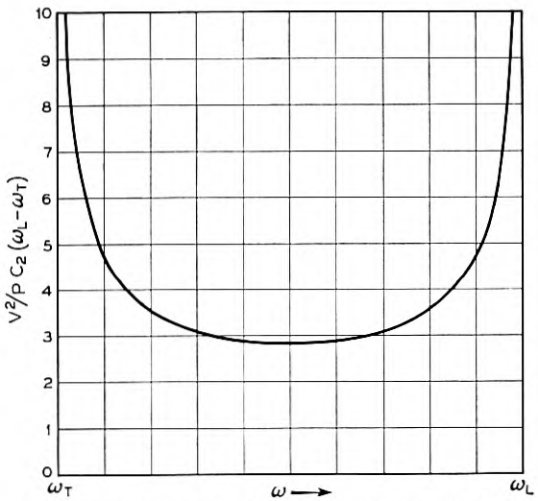


Fig. 4.28—A quantity proportional to $(E^2/\beta^2 P)$ for the circuit of Fig. 4.11, plotted vs radian frequency ω.

In a structure such as that shown in Fig. 4.11, there is little coupling between sections which are not adjacent, and hence the lumped-circuit representation used is probably quite accurate, and is certainly more accurate than in structures such as that shown in Fig. 4.10.

Other structures could be analyzed, but it is believed that the examples given above adequately illustrate the general procedures which can be employed.

4.4 Traveling Field Components

Filter-type circuits produce fields which are certainly not sinusoidal with distance. Indeed, with a structure such as that shown in Fig. 4.11, the elec-

trons are acted upon only when they are very near to the gaps. It is possible to analyze the performance of traveling-wave tubes on this basis[5]. The chief conclusion of such an analysis is that highly accurate results can be obtained by expressing the field as a sum of traveling waves and taking into account only the wave which has a phase velocity near to the electron velocity. Of course this is satisfactory only if the velocities of the other components are quite different from the electron velocity (that is, different by a fraction several times the gain parameter C).

As an example, consider a traveling-wave tube in which the electron stream passes through tubular sections of radius a, as shown in Fig. 4.29, and is acted upon by voltages appearing across gaps of length ℓ spaced L apart.

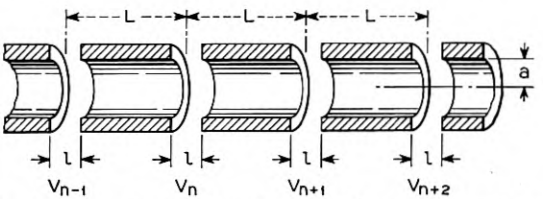


Fig. 4.29—A series of gaps in a tube of inside radius a. The gaps are ℓ long and are spaced L apart. Voltages V_n, etc., act across them.

A wave travels in some sort of structure and produces voltages across the gaps such that that across the nth gap, V, is

$$V_n = V_0 e^{-jn\theta} \tag{4.60}$$

where n is any integer.

We analyze this field into traveling-wave components which vary with distance as $\exp(-j\beta_m z)$ where

$$\beta_m = (\theta + 2m\pi)/L \tag{4.61}$$

where m is any positive or negative integer. Thus, the total field will be

$$E = \sum_{m=-\infty}^{\infty} E_m = \sum_{m=-\infty}^{\infty} A_m e^{-j\beta_m z} I_0(\gamma_m r) \tag{4.62}$$

$$\gamma_m^2 = \beta_m^2 - \beta_0^2 \tag{4.63}$$

Here $I_0(\gamma_m r)$ is a modified Bessel function, and γ_m has been chosen so that (4.62) satisfies Maxwell's equations.

[5] J. R. Pierce and Nelson Wax, "A Note on Filter-Type Traveling-Wave Amplifiers," *Proc. I.R.E.*, Vol. 37, pp. 622–625, June, 1949.

We will evaluate the coefficients by the usual means of Fourier analysis. Suppose we let $z = 0$ at the center of one of the gaps. We see that

$$\begin{aligned}\int_{-L/2}^{L/2} EE^* \, dz &= \sum_{m=-\infty}^{\infty} \int_{-L/2}^{L/2} A_m A_m^* I_0^2(\gamma_m r) \, dz \\ &= \sum_{m=-\infty}^{\infty} A_m A_m^* I_0^2(\gamma_m r) L\end{aligned} \tag{4.64}$$

All of the terms of the form $E_m E_p$, $p \neq m$ integrate to zero because the integral contains a term $\exp(-j2\pi(p - m)/L)z$.

Let us consider the field at the radius r. This is zero along the surface of the tube. We will assume with fair accuracy that it is constant and has a value $-V/\ell$ across the gap. Thus we have also at $r = a$,

$$\begin{aligned}\int_{-L/2}^{L/2} EE^* \, dz &= -(V/\ell) \sum_{m=-\infty}^{\infty} \int_{-\ell/2}^{\ell/2} A_m^* e^{-j\beta_m z} I_0(\gamma_m a) \, dz \\ &= -(V/\ell) \sum_{m=-\infty}^{\infty} (A_m^*) I_0(\gamma_m a) \left(\frac{e^{-j\beta_m \ell/2} - e^{j\beta_m \ell/2}}{j\beta} \right)\end{aligned} \tag{4.65}$$

We can rewrite this

$$\int_{-L/2}^{L/2} EE^* \, dz = -(V/\ell) \sum_{m=-\infty}^{\infty} A_m^* I_0(\gamma_m a) \frac{\sin(\beta_m \ell/2)}{(\beta_m \ell/2)} \tag{4.66}$$

By comparison with (4.64) we see that

$$A_m = -(V/L)(\sin(\beta_m \ell/2)/(\beta_m \ell/2))(1/I_0(\gamma a)) \tag{4.67}$$

This is the magnitude of the mth field component on the axis. The magnitude of the field at a radius r would be $I_0(\gamma r)$ times this.

The quantity $\beta_m \ell$ is an angle which we will call θ_g, the gap angle. Usually we are concerned with only a single field component, and hence can merely write γ instead of γ_m. Thus, we say that the magnitude E of the travelling field produced by a voltage V acting at intervals L is

$$E = -M(V/L) \tag{4.68}$$

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{I_0(\gamma r)}{I_0(\gamma a)} \tag{4.69}$$

$$\theta g = \beta \ell \tag{4.70}$$

The factor M is called the gap factor or the modulation coefficient*. For slow waves, γ is very nearly equal to β, and we can replace γr and γa by βr and βa. For unattenuated waves, M is a real positive number; and,

* This factor is often designated by β, but we have used β otherwise.

for the slowly varying waves with which we deal, we will always consider M as a real number.

The gap factor for some other physical arrangements is of interest. At a distance y above the two-dimensional array of strip electrodes shown in Fig. 4.30

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} e^{-\gamma y} \tag{4.71}$$

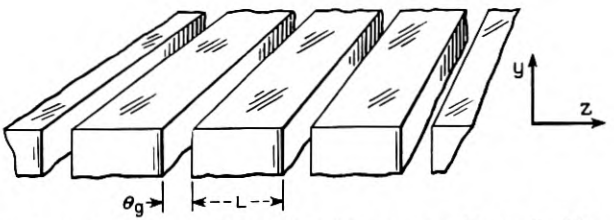


Fig. 4.30—A series of slots θ_g radians long separated by walls L long.

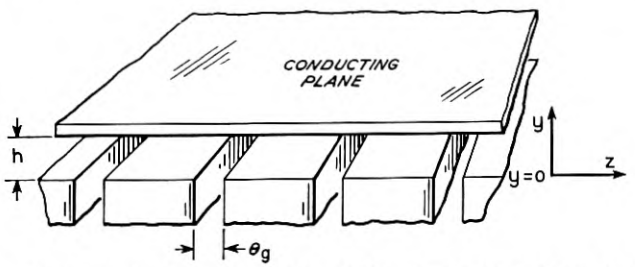


Fig. 4.31—A system similar to that of Fig. 4.30 but with the addition of an opposed conducting plane.

If we add a conducting plane a at $y = h$, as in Fig. 4.31,

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{\sinh \gamma(h - y)}{\sinh \gamma h} \tag{4.72}$$

For a symmetrical two-dimensional array, as shown in Fig. 4.32, with a separation of 2 h in the y direction and the fields above equal to the fields below

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{\cosh \gamma y}{\cosh \gamma h} \tag{4.73}$$

4.5 Effective Field and Effective Current

In Section 4.4 we have expressed a field component or "effective field" in terms of circuit voltage by means of a gap-factor or modulation coeffi-

cient M. This enables us to make calculations in terms of fields and currents at the electron stream.

The gap factor can be used in another way. A voltage appears across a gap, and the electron stream induces a current at the gap. At the electron stream the power P_1, produced in a distance L by a convection current i with the same z-variation as the field component considered, acting on the field component is

$$\begin{aligned} P_1 &= -Ei^*L \\ &= +(MV)i^* \end{aligned} \tag{4.74}$$

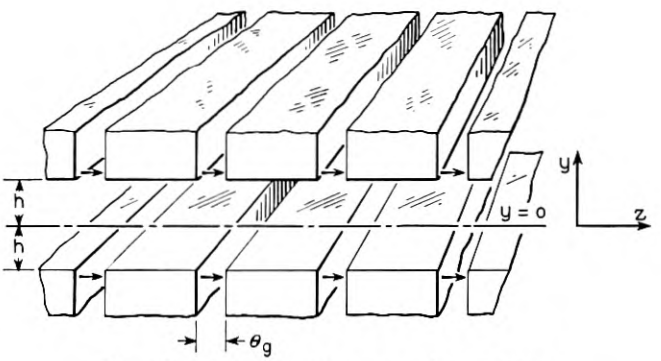


Fig. 4.32—A system of two opposed sets of slots.

At the circuit we observe some impressed current I flowing against the voltage V to produce a power

$$P_2 = VI^* \tag{4.75}$$

By the conservation of energy, these two powers must be the same, and we deduce that

$$I^* = Mi^* \tag{4.76}$$

or, since we take M as a real number

$$I = Mi \tag{4.77}$$

Thus, we have our choice of making calculations in terms of the beam current and a field component or effective field, or in terms of circuit voltage and an effective current, and in either case we make use of the modulation coefficient M.

Our gain parameter C^3 will be

$$C^3 = (V/L)^2M^2I_0/8\beta^2V_0$$

where V is circuit voltage. We can regard this in two ways. We can think of $-(V/L)M$ as the effective field at the location of the current I_0, or we can think of M^2I_0 as the effective current referred to the circuit.

If we have a broad beam of electrons and a constant current density J_0 we compute (essentially as in Chapter III) a value of C^3 by integrating

$$C^3 = (1/8\beta^2V_0)J_0(V/L)^2 \int M^2 \, d\sigma \tag{4.78}$$

where $d\sigma$ is an element of area. We can think of the result in terms of an effective field E_e

$$E_e^2 = (V/L)^2 \frac{\int M^2 \, d\sigma}{\sigma} \tag{4.79}$$

where σ is the total beam area, and a total current σJ_0, or we can think of the integral (4.77) in terms of an effective current I_0 given by

$$I_0 = J_0 \int M^2 \, d\sigma \tag{4.80}$$

and the voltage at the circuit.

Of course, these same considerations apply to distributed circuits. Sometimes it is most convenient to think in terms of the total current and an effective field (as we did in connection with helices in Chapter III) and sometimes it is most convenient to think of the field at the circuit and an effective current. Either concept refers to the same mathematics.

4.6 Harmonic Operation

Of the field components making up E in (4.62) it is customary to regard the $m = 0$ component, for which $\beta = \theta/L$, as the *fundamental* field component, and the other components as *harmonic* components. These are sometimes called *Hartree harmonics*. If the electron speed is so adjusted that the interaction is with the $m = 0$ or fundamental component we have fundamental operation; if the electron speed is adjusted so that we have interaction with a harmonic component, we have harmonic operation.

There are several reasons for using harmonic operation in connection with filter-type circuits. For one thing the fundamental component may appear to be traveling backwards. Thus, for circuits of the type shown in Fig. 4.11, we see from Fig. 4.27 that θ is always negative. Now, in terms of the velocity v

$$\beta = \omega/v = \theta/L \tag{4.81}$$

and if θ is negative, v must be negative. However, consider the $m = 1$ component

$$\beta = \omega/v = (2\pi + \theta)/L \tag{4.82}$$

We see that, for this component, v is positive.

The interaction of electrons with backward-traveling field components will be considered later. Here it will merely be said that, in order to avoid interaction with waves traveling in both directions, one must avoid having the electron speed lie near both the speed of a forward component and the speed of a backward component.

In order that the fundamental component be slow, θ must be large or L must be small. The largest value of θ is that near one edge of the band, where θ approaches π. Thus, the largest fundamental value of β is π/L, and to make

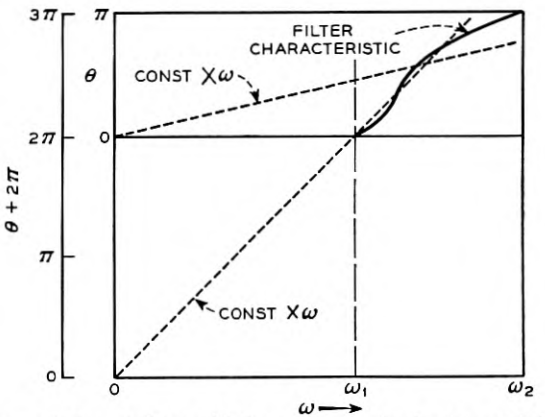


Fig. 4.33—The variation of phase with frequency for the fundamental (0 to π over the band) and a spatial harmonic (2π to 3π over the band). The dotted lines show ω divided by the electron velocity for the two cases. For amplification over a broad band the dotted curve should not depart much from the filter characteristic.

β large with $m = 0$ we must make L small and put the resonators very close together. This may be physically difficult or even impossible in tubes for very high frequencies. The alternative is to use a harmonic component, for which $\beta = (2m\pi + \theta)/L$.

Another reason for using harmonic operation is to achieve broad-band operation. The phase of a filter-type circuit changes by π radians between the lower cutoff frequency ω_1 and the upper cutoff frequency ω_2†. Now, for the wave velocity to be near to the electron velocity over a good part of the band, β must be nearly a constant times ω. Figure 4.33 shows how this can be approximately true for the $m = 1$ component even when it obviously won't be for the $m = 0$ or fundamental component. Similarly, for a filter with a narrower fractional bandwidth and hence a steeper curve of θ vs ω, a larger value of m might give a nearly constant value of v.

† The phase of some filters changes more than this, but they don't seem good candidates for traveling-wave tube circuits.

CHAPTER V

GENERAL CIRCUIT CONSIDERATIONS

Synopsis of Chapter

IN CHAPTERS III AND IV, helices and filter-type circuits have been considered. Other slow-wave circuits have been proposed, as, for instance, wave guides loaded continuously with dielectric material. One may ask what the best type of circuit is, or, indeed, in just what way do bad circuits differ from good circuits.

So far, we have as one criterion for a good circuit a high impedance, that is, a high value of $E^2/\beta^2 P$. If we want a broad-band amplifier we must have a constant phase velocity; that is, β must be proportional to frequency. Thus, two desirable circuit properties are: high impedance and constancy of phase velocity.

Now, $E^2/\beta^2 P$ can be written in the form

$$E^2/\beta^2 P = E^2/\beta^2 W v_g$$

where W is the stored energy per unit length for a field strength E, and v_g is the group velocity.

One way of making $E^2/\beta^2 P$ large is to make the stored energy for a given field strength small. In an electromagnetic wave, half of the stored energy is electric and half is magnetic. Thus, to make the total stored energy for a given field strength small we must make the energy stored in the electric field small. The energy stored in the electric field will be increased by the presence of material of a high dielectric constant, or by the presence of large opposed metallic surfaces, as in the circuits of Figs. 4.8 and 4.9. Thus, such circuits are poor as regards circuit impedance, however good they may be in other respects.

If the stored energy for a given field strength is held constant, $E^2/\beta^2 P$ may be increased by decreasing the group velocity. It is the phase velocity v which should match the electron speed. The group velocity v_g is given in terms of the phase velocity by (5.12). We see that the group velocity may be much smaller than the phase velocity if $-\partial v/\partial\omega$ is large. It is, for instance, a low group velocity near cutoff that accounts for the high impedance regions exhibited in Figs. 4.20 and 4.28. We remember, however, that, if the phase velocity of the circuit of a traveling-wave tube changes with frequency, the tube will have a narrow bandwidth, and thus the high

impedances attained through large values of $-\partial v/\partial\omega$ are useful over a narrow range of frequency only.

If we consider a broad electron stream of current density J_0, the highest effective value of $E^2/\beta^2 P$, and hence the highest value of C, will be attained if there is current everywhere that there is electric field, and if all of the electric field is longitudinal. This leads to a limiting value of C, which is given by (5.23). There λ_0 is the free-space wavelength. The nearest practical approach to this condition is perhaps a helix of fine wire flooded inside and outside with electrons.

In many cases, it is desirable to consider circuits for use with a narrow beam of electrons, over which the field may be taken as constant. As the helix is a common as well as a very good circuit, it might seem desirable to use it as a standard for comparison. However, the group velocity of the helix differs a little from the phase velocity, and it seems desirable instead to use a sort of hypothetical circuit or field for which the stored energy is almost the same as in the helix, but for which the group velocity is the same as the phase velocity. This has been referred to in the text as a "forced sinusoidal field." In Fig. 5.3, $(E^2/\beta^2 P)^{1/3}$ for the forced sinusoidal field is compared with $(E^2/\beta^2 P)^{1/3}$ for the helix.

Several other circuits are compared with this: the circular resonators of Fig. 5.4 (the square resonators of Fig. 5.4 give nearly the same impedance) and the resonant quarter-wave and half-wave wires of Figs. 5.6 and 5.7. The comparison is made in Fig. 5.8 for three voltages, which fix three phase velocities. In each case it is assumed that in some way the group velocity has been made equal to the phase velocity. Thus, the comparison is made on the basis of stored energies. The field is taken as the field at radius a (corresponding to the surface of the helix) in the case of the forced sinusoidal field, and at the point of highest field in the case of the resonators.

We see from Figs. 5.8 and 5.3 that a helix of small radius is a very fine circuit.

In circuits made up of a series of resonators, the group velocity can be changed within wide limits by varying the coupling between resonators, as by putting inductive or capacitive irises between them. Thus, even circuits with a large stored energy can be made to have a high impedance by sacrificing bandwidth.

The circuits of Fig. 5.4 have a large stored energy because of the large opposed surfaces. The wires of Fig. 5.6 have a small stored energy associated entirely with "fringing fields" about the wires. The narrow strips of Fig. 5.5 have about as much stored energy between the opposed flat surfaces as that in the fringing field, and are about as good as the half-wave wires of Fig. 5.7.

An actual circuit made up of resonators such as those of Fig. 5.4 will be

worse than Fig. 5.8 implies. Thus, there is a decrease of $(E^2/\beta^2P)^{1/3}$ due to wall thickness. Thickening the flat opposed walls of the resonators decreases the spacing between the opposed surfaces, increases the capacitance and hence increases the stored energy for a given gap voltage. In Fig. 5.9 the factor f by which $(E^2/\beta^2P)^{1/3}$ is reduced is plotted vs. the ratio of the wall thickness t to the resonator spacing L.

There is a further reduction of effective field because of the electrical length, θ in radians, of the space between opposed resonator surfaces. The lower curve in Fig. 5.10 gives a factor by which $(E^2/\beta^2P)^{1/3}$ is reduced because of this. If the resonator spacing, θ_t in radians, is greater than 2.33 radians, it is best to make the opening, or space between the walls, only 2.33 radians long by making the opposed disks forming the walls very thick.

There is of course a further loss in effective field, both in the helix and in circuits made up of resonators, because of the falling-off of the field toward the center of the aperture through which the electrons pass. This was discussed in Chapter IV.

Finally, it should be pointed out that the fraction of the stored energy dissipated in losses during each cycle is inversely proportional to the Q of the circuit or of the resonators forming it. The distance the energy travels in a cycle is proportional to the group velocity. Thus, for a given Q the signal will decay more rapidly with distance if the group velocity is lowered (to increase E^2/β^2P). Equations (5.38), (5.42) and (5.44) pertain to attenuation expressed in terms of group velocity. The table at the end of the chapter shows that a circuit made up of resonators and having a low enough group velocity to give it an impedance comparable with that of a helix can have a very high attenuation.

5.1 Group and Phase Velocity

Suppose we use a broad video pulse $F(t)$, containing radian frequencies p lying in the range 0 to p_0, to modulate a radio-frequency signal of radian frequency ω which is much larger than p_0, so as to give a radio-frequency pulse $f(t)$

$$f(t) = e^{j\omega t}F(t) \tag{5.1}$$

the functions $F(t)$ and $f(t)$ are indicated in Fig. 5.1.

$F(t)$, which is a real function of time, can be expressed by means of its Fourier transform in terms of its frequency components

$$F(t) = \int_{p_0}^{p_0} A(p)e^{jpt}\,dp \tag{5.2}$$

Here $A(p)$ is a complex function of p, such that $A(-p)$ is the complex conjugate of $A(p)$ (this assures that $F(t)$ is real).

With $F(t)$ expressed as in (5.2), we can rewrite (5.1)

$$f(t) = \int_{-p_0}^{p_0} A(p)e^{j(\omega+p)t}\,dp \tag{5.3}$$

Now, suppose, as indicated in Fig. 5.2, we apply the *r-f* pulse $f(t)$ to the input of a transmission system of length L with a phase constant β which

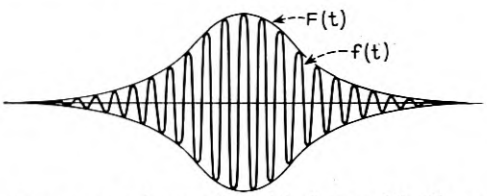


Fig. 5.1—A radio-frequency pulse varying with time as $f(t)$. The envelope varies with time as $F(t)$. The pulse might be produced by modulating a radio-frequency source with $F(t)$.

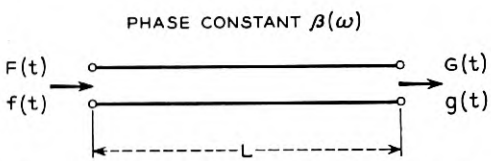


Fig. 5.2—When the pulse of Fig. 5.1 is applied to a transmission system of length L and phase constant $\beta(\omega)$ (a function of ω), the output pulse $g(t)$ has an envelope $G(t)$.

is a function of frequency. Let us assume that the system is lossless. The output $g(t)$ will then be

$$g(t) = \int_{-p_0}^{p_0} A(p)e^{j((\omega+p)t-\beta L)}\,dp \tag{5.4}$$

We have assumed that p_0 is much smaller than ω. Let us assume that over the range $\omega - p_0$ to $\omega + p_0$, β can be adequately represented by

$$\beta = \beta_0 + \frac{\partial\beta}{\partial\omega}\,p \tag{5.5}$$

In this case we obtain

$$g(t) = e^{j(\omega t-\beta_0 L)} \int_{-p_0}^{p_0} A(p)\, e^{jp(t-(\partial\beta/\partial\omega)L)}\,dp \tag{5.6}$$

The envelope at the output is

$$G(t) = \int_{-p_0}^{p_0} A(p)\, e^{jp(t-(\partial\beta/\partial\omega)L)}\,dp \tag{5.7}$$

By comparing this with (5.2) we see that

$$G(t) = F\left(t - \frac{\partial\beta}{\partial\omega} L\right) \tag{5.8}$$

In other words, the envelope at the output is of the same shape as at the input, but arrives a time τ later

$$\tau = \frac{\partial\beta}{\partial\omega} L \tag{5.9}$$

This implies that it travels with a velocity v_g

$$v_g = L/\tau = \left(\frac{\partial\beta}{\partial\omega}\right)^{-1} \tag{5.10}$$

This velocity is called the group velocity, because in a sense it is the velocity with which the group of frequency components making up the pulse travels down the circuit. It is certainly the velocity with which the energy stored in the electric and magnetic fields of the circuit travels; we could observe physically that, if at one time this energy is at a position x, a time t later it is at a position $x + v_g t$.

If the attenuation of the transmission circuit varies with frequency, the pulse shape will become distorted as the pulse travels and the group velocity loses its clear meaning. It is unlikely, however, that we shall go far wrong in using the concept of group velocity in connection with actual circuits.

We have used earlier the concept of phase velocity, which we have designated simply as v. In terms of phase velocity,

$$\beta = \frac{\omega}{v} \tag{5.11}$$

We see from (5.10) that in terms of phase velocity v the group velocity v_g is

$$v_g = v\left(1 - \frac{\omega}{v}\frac{\partial v}{\partial\omega}\right)^{-1} \tag{5.12}$$

For interaction of electrons with a wave to give gain in a traveling-wave tube, the electrons must have a velocity near the phase velocity v. Hence, for gain over a broad band of frequencies, v must not change with frequency; and if v does not change with frequency, then, from (5.12), $v_g = v$.

We note that the various harmonic components in a filter-type circuit have different phase velocities, some positive and some negative. The group

velocity is of course the same for all components, as they are all aspects of one wave. Relation (4.61) is consistent with this:

$$\beta_m = (\theta + 2m\pi)/L \tag{4.61}$$

$$1/v_g = \partial\beta_m/\partial\omega = (\partial\theta/\partial\omega)/L \tag{5.13}$$

5.2 Gain and Bandwidth in a Traveling-Wave Tube

We can rewrite the impedance parameter E^2/β^2P in terms of stored energy per unit length W for a field strength E, and a group velocity v_g. If W is the stored energy per unit length, the power flow P is

$$P = Wv_g \tag{5.14}$$

and, accordingly, we have

$$E^2/\beta^2P = E^2/\beta^2Wv_g \tag{5.15}$$

And, for the gain parameter, we will have

$$C = (E^2/\beta^2Wv_g)^{1/3}(I_0/8V_0)^{1/3} \tag{5.16}$$

For example, we see from Fig. 4.20 that E^2/β^2P for the circuit of Fig. 4.10 goes to infinity at the upper cut-off. From Fig. 4.17 we see that $\partial\theta/\partial\omega$, and hence $1/v_g$, go to infinity at the upper cutoff, accounting for the infinite impedance. We see also that $\partial\theta/\partial\omega$ goes to infinity at the lower cutoff, but there the slot voltage and hence the longitudinal field also go to zero and hence E^2/β^2P does not go to infinity but to zero instead.

In the case of the circuit of Fig. 4.11, the gap voltage and hence the longitudinal field are finite for unit stored energy at both cutoffs. As $\partial\theta/\partial\omega$ is infinite at both cutoffs, V^2/P and hence E^2/β^2P go to infinity at both cutoffs, as shown in Fig. 4.28.

To get high gain in a traveling-wave tube at a given frequency and voltage (the phase velocity is specified by voltage) we see from (5.16) that we must have either a small stored energy per unit length for unit longitudinal field, or a small group velocity, v_g.

To have amplification over a broad band of frequencies we must have the phase velocity v substantially equal to the electron velocity over a broad band of frequencies. This means that for very broad-band operation, v must be substantially constant and hence in a broad-band tube the group velocity will be substantially the same as the phase velocity.

If the group velocity is made smaller, so that the gain is Increased, the range of frequencies over which the phase velocity is near to the electron velocity is necessarily decreased. Thus, for a given phase velocity, as the group velocity is made less the gain increases but the bandwidth decreases.

Particular circuits can be compared on the basis of (E^2/β^2P) and band-

width. We have discussed the impedance and phase or velocity curves in Chapters III and IV. Field[1] has compared a coiled waveguide structure with a series of apertured disks of comparable dimensions. Both of these structures must have about the same stored energy for a given field strength. He found the coiled waveguide to have a low gain and broad bandwidth as compared with the apertured disks. We explain this by saying that the particular coiled waveguide he considered had a higher group velocity than did the apertured disk structure. Further, if the coiled waveguide could be altered in some way so as to have the same group velocity as the apertured disk structure it would necessarily have substantially the same gain and bandwidth.

In another instance, Mr. O. J. Zobel of these Laboratories evaluated the effect of broad-banding a filter-type circuit for a traveling-wave tube by m-derivation. He found the same gain for any combination of m and bandwidth which made $v = v_g(\partial v/\partial\omega = 0)$. We see this is just a particular instance of a general rule. The same thing holds for any type of broad-banding, as, by harmonic operation.

5.3 A Comparison of Circuits

The group velocity, the phase velocity and the ratio of the two are parameters which are often easily controlled, as, by varying the coupling between resonators in a filter composed of a series of resonators. Moreover, these parameters can often be controlled without much affecting the stored energy per unit length. For instance, in a series of resonators coupled by loops or irises, such as the circuit of Fig. 4.11, the stored energy is not much affected by the loops or irises unless these are very large, but the phase and group velocities are greatly changed by small changes in coupling.

Let us, then, think of circuits in terms of stored energy, and regard the phase and group velocities and their ratio as adjustable parameters. We find that, when we do this, there are not many essentially different configurations which promise to be of much use in traveling-wave tubes, and it is easy to make comparisons between extreme examples of these configurations.

5.3a Uniform Current Density throughout Field

Suppose we have a uniform current density J_0 wherever there is longitudinal electric field. We might approximate this case by flooding a helix of very fine wire with current inside and outside, or by passing current through a series of flat resonators whose walls were grids of fine wire.

[1] Lester M. Field, "Some Slow-Wave Structures for Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 37, pp. 34–40, January 1949.

In the latter case, if resonators had parallel walls of very fine mesh normal to the direction of electron motion there would be substantially no transverse electric field. All the electric field representing stored energy would act on the electron stream. In this case, we would have

$$W = \frac{\epsilon}{2} \int E^2 \, d\Sigma \tag{5.17}$$

Here $d\Sigma$ is an elementary area normal to the direction of propagation. W given by this expression is the total electric and magnetic stored energy per unit length. Where E is less than its peak value, the magnetic energy makes up the difference.

In evaluating $E^2 I_0$ in (5.16) we will have as an effective value

$$(EI_0)_{\text{eff}} = J_0 \int E \, d\Sigma \tag{5.18}$$

Hence, we will have for the gain parameter C

$$C = \left(\frac{J_0 \int E^2 \, d\Sigma}{\left(\frac{\omega}{v}\right)^2 \left(\frac{\epsilon}{2} \int E^2 \, d\Sigma\right) v_g (8V_0)} \right)^{1/3}$$

$$C = \left(\frac{J_0}{4 \left(\frac{\omega}{v}\right)^2 \epsilon v_g V_0} \right)^{1/3} \tag{5.19}$$

It is of interest to put this in a slightly different form. Suppose λ_0 is the free-space wavelength. Then

$$\frac{\omega}{v} = \frac{2\pi}{\lambda_0} \frac{c}{v} \tag{5.20}$$

where c is the velocity of light

$$c = 3 \times 10^{10} \text{ cm/sec} = 3 \times 10^8 \text{ m/sec}$$

Further, we have for synchronism between the electron velocity u_0 and the phase velocity v

$$v^2 = 2\eta V_0 \tag{5.21}$$

Also

$$c = 1/\sqrt{\mu\epsilon}$$

$$\epsilon = 1/c\sqrt{\mu/\epsilon} \tag{5.22}$$

$$\sqrt{\mu/\epsilon} = 377 \text{ ohms}$$

Using (5.20), (5.21), (5.22) in connection with (5.19), we obtain

$$C = \left(\frac{\eta\sqrt{\mu/\epsilon}\ J_0\lambda_0^2}{16\pi^2 c v_g}\right)^{1/3} = 11.16\ (J_0\lambda_0^2/v_g)^{1/3} \tag{5.23}$$

We have in (5.23) an expression for the gain parameter C in case longitudinal fields only are present and in case there is a uniform current density J_0 wherever there is a longitudinal field.

In a number of cases, as in case of a large-diameter helix, or of a resonator with large apertures, the stored energy due to the transverse field is about equal to that due to the longitudinal field and C will be $2^{-1/3}$ times as great as the value of C given by (5.23). Thus, the value of C given by (5.23), or even $2^{-1/3}$ times this, represents an unattainable ideal. It is nevertheless of interest in indicating how limiting behavior depends on various parameters. For instance, we see that if the wavelength λ_0 is made shorter, a higher current density must be used if C is not to be lowered; for a constant C the current density must be such as to give a constant current through a square a wavelength on a side.

In the table below, some values of C have been computed from (5.23) for various wavelengths and current densities. The broad-band condition of equal phase and group velocities has been assumed, and the voltage has been taken as 1,000 volts.

Wavelength Cm \ Amp/cm²	.1	1
5	.060	.130
.5	.013	.028

For larger voltages, C will be smaller. C can of course be made larger by making the group velocity smaller than the phase velocity.

Of course, if the electron stream does not pass through some portions of the field, C will be smaller than given by (5.23). C will also be less if there are "harmonic" field components which do not vary in the z direction as $\exp(j\omega z/v)$.

5.3b Narrow Beams

Usually, no attempt is made to fill the entire field with electron flow even though this is necessary in getting a large value of C for a given current density. Instead a narrow electron beam is shot through a region of high

field. We then wish to relate the peak field strength to the stored energy in comparing various circuits.

Let us first consider a helically conducting sheet of radius a. The upper curve of Fig. 5.3 shows $(E^2/\beta^2 P)^{1/3}(v/c)^{1/3}$ vs. βa. In obtaining this curve it was assumed that $v \ll c$, so that γ can be taken as equal to β. The field E is the longitudinal field at the surface of the helically conducting cylinder. Figure 5.3 can be obtained from Fig. 3.4 by multiplying $F(\gamma a)$ by $(I_0(\gamma a))^{2/3}$ to give a curve valid for the field at $r = a$.

The helix has a very small circumferential electric field which represents "useless" stored energy. The lower curve of Fig. 5.3 is based on the stored electric energy of an axially symmetrical sinusoidal field impressed at the radius a.† This field has no circumferential component but is otherwise the

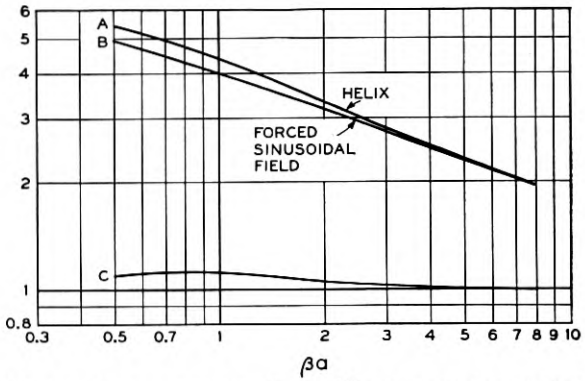


Fig. 5.3—The impedance parameter $(E^2/\beta^2 P)^{1/3}$ compared for a helically conducting sheet (A) and a forced sinusoidal field (B) with a group velocity equal to the phase velocity. The helix has a higher impedance because the phase velocity is higher than the group velocity by a radio shown to the $\frac{1}{3}$ power by curve C.

same as the electric field of the helix (again assuming $v \ll c$). We can imagine such a field propagating because of an inductive sheet at the radius a, which provides stored magnetic energy enough to make the electric and magnetic energies equal. The quantity plotted vs. βa is $(E^2/\beta^2 P)^{1/3}$ $(v/c)^{1/3}$ $(v_g/v)^{1/3}$.

The forced sinusoidal field is not the field of some particular circuit for which a certain group velocity v_g corresponds to a given phase velocity v. Hence, the factor $(v_g/v)^{1/3}$ is included in the ordinate, so that the curve will be the same no matter what group velocity is assumed. For the helically conducting sheet, a definite group velocity goes with a given phase velocity. In Fig. 5.3, the ordinate of the curve for the helically conducting sheet does not contain the factor $(v_g/v)^{1/3}$. If, for instance, we assume $v_g = v$

† See Appendix III.

in connection with the curve for the forced sinusoidal field, then the two ordinates are both $(E^2/\beta^2P)^{1/3}$ $(v/c)^{1/3}$ and the curve for the sheet is higher than that for the forced field because, for the helically conducting sheet,

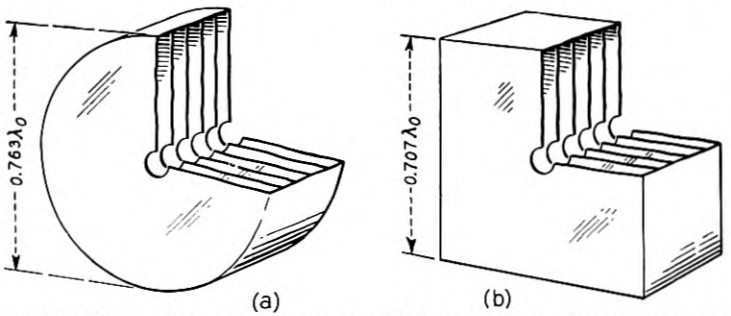


Fig. 5.4—Pillbox and rectangular resonators. When a number of resonators are coupled one to the next, a filter-type circuit is formed.

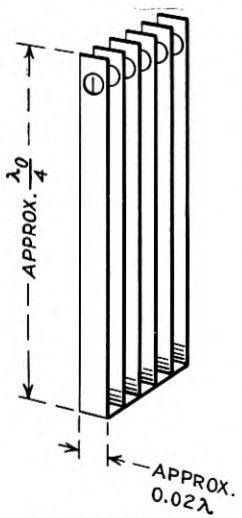


Fig. 5.5—Resonators with the opposing parallel surfaces reduced to lower stored energy and increase impedance.

$v_g < v$ for small values of γa. Curve C shows $(v/v_g)^{1/3}$ for the sheet vs. βa. Aside from the influence of group velocity, we might have expected the curve for the sheet to be a little lower than that for the forced field because of the energy associated with the transverse electric field component of the sheet. This, however, becomes small in comparison with the transverse magnetic component when $v \ll c$, as we have assumed.

Various other circuits will be compared, using the impressed sinusoidal field as a sort of standard of reference.

One of the circuits which will be considered is a series of flat resonators coupled together to make a filter. Figure 5.4a shows a series of very thin pillboxes with walls of negligible thickness. A small central hole is provided for the electron stream, and the field E is to be measured at the edge of this hole. The diameter is chosen to obtain resonance at a wavelength λ_0. Figure 5.4b shows a similar series of flat square resonators.

For the round resonators it is found that*

$$(E^2/\beta^2P)^{1/3} = 5.36\ (v/c)^{1/3}\ (v/v_g)^{1/3} \tag{5.24}$$

for the square resonators*

$$(E^2/\beta^2P)^{1/3} = 5.33\ (v/c)^{1/3}\ (v/v_g)^{1/3} \tag{5.25}$$

For practical purposes these are negligibly different.

* See Appendix III.

Suppose we wanted to improve on such circuits by reducing the stored energy. An obvious procedure would be to cut away most of the flat opposed surfaces as shown in Fig. 5.5. This reduces the energy stored between the resonator walls, but results in energy storage outside of the open edges, energy associated with a "fringing field."

Going to an extreme, we might consider an array of closely spaced very fine wires, as shown in Fig. 5.6. Here there are no opposed flat surfaces, and all of the electric field is a fringing field; we have reached an irreducible minimum of stored energy in paring down the resonator.

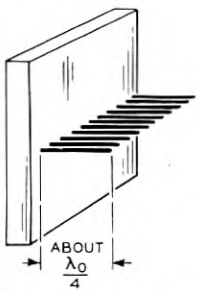


Fig. 5.6—Quarter-wave wires, which have a minimum of stored energy.

The structure of Fig. 5.6 has not been analyzed exactly, but that of Fig. 5.7 has. In Fig. 5.7, we have an array of fine, closely spaced half-wave wires between parallel planes.* This should have roughly twice the stored energy of Fig. 5.6, and we will estimate $(E^2/\beta^2 P)^{1/3}$ for Fig. 5.6 on this basis. We obtain in Appendix III:

For the half-wave wires,

$$(E^2/\beta^2 P)^{1/3} = 6.20\ (v/v_g)^{1/3} \qquad (5.25)$$

and hence for the quarter-wave wires, approximately

$$(E^2/\beta^2 P)^{1/3} = 7.81\ (v/v_g)^{1/3} \qquad (5.26)$$

As we have noted, (v/c), which appears in the expression for $(E^2/\beta^2 P)^{1/3}$ for the sinusoidal field impressed at radius a and in (5.24) and (5.25), is a

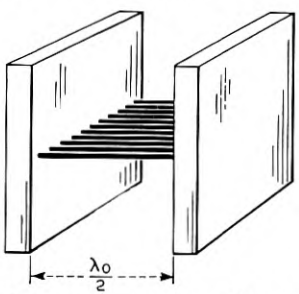


Fig. 5.7—Half-wave wires between parallel planes. The stored energy can be calculated for this configuration, assuming the wires to be very fine. The circuit does not propagate a wave unless added coupling is provided.

function of the accelerating voltage. Figure 5.8 makes a comparison between the sinusoidal field impressed at a radius a, curve A; the flat resonators, either circular or square, B; the half-wave wires, C; and the quarter-

* There is no transverse magnetic wave propagation along such a circuit unless extra coupling or loading is provided. Behavior of nonpropagating circuits in the presence of an electron stream is considered in Section 4 of Chapter XIV.

wave wires C'. In all cases, it is assumed that the coupling is so adjusted as to make $(v_g/v) = 1$ (broad-band condition).

What sort of information can we get from the curves of Fig. 5.8? Consider the curves for 1,000 volts. Suppose we want to cut down the opposed areas of resonators, as indicated in Fig. 5.5, so as to make them as good as half-wave wires (curve C). The edge capacitance in Fig. 5.5 will be about equal to that for quarter-wave wires (curve C'). Curve C' is about 3.7 times as high as curve B, and hence represents only about $(1/3.7)^3 = .02$ as much capacitance. If we make the opposed area in Fig. 5.5 about .01 that in Fig. 5.4a or b, the capacitance* between opposed surfaces will equal the edge

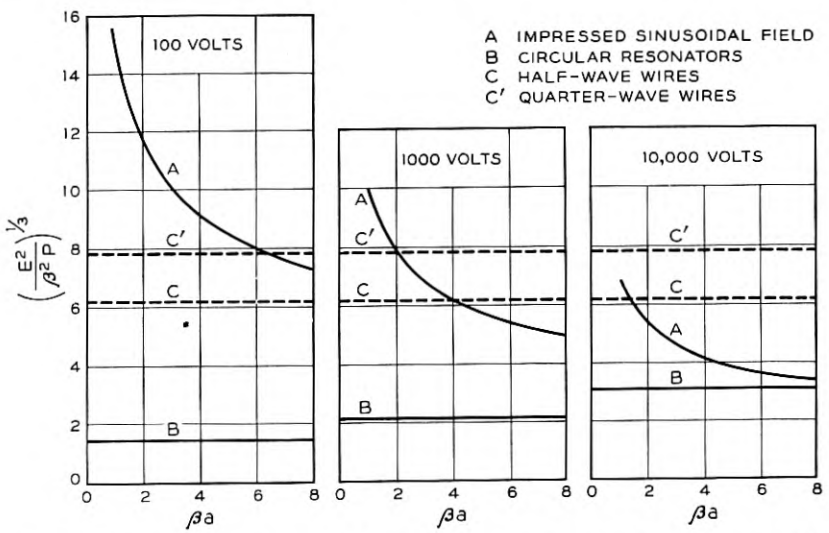


Fig. 5.8—Comparisons in terms of impedance parameter of an impressed sinusoidal field (A), circular resonators (B), half-wave wires (C) and quarter-wave wires (C') assuming the group and phase velocities to equal the electron velocity. The radius of the impressed sinusoidal field is a.

capacitance and the total stored energy will be twice that for quarter-wave wires, or equal to that for half-wave wires. This area is shown approximately to scale relative to Fig. 5.4 in Fig. 5.5. Thus, at 1,000 volts the resonant strips of Fig. 5.5 are about as good as fine, closely spaced half-wave wires.

Suppose again that we wish at 1,000 volts to make the gain of the resonators of Fig. 5.4 (or of a coiled waveguide) as good as that for a helix with $\beta a = 3$. For $\beta a = 3$ the helix curve A is about 3.2 times as high as the resona-

* This takes into account a difference in field distribution—that in Fig. 5.4b.

tor curve B. As $(E^2/\beta^2P)^{1/3}$ varies as $(v/v_g)^{1/3}$, we must adjust the coupling between resonators so as to make

$$v_g = v/(3.2)^3 = .031\ v$$

in order to make $(E^2/\beta^2P)^{1/3}$ the same for the resonators as for the helix. From (5.12) we see that this means that a change in frequency by a fraction .002 must change v by a fraction .06. Ordinarily, a fractional variation of v of $\pm.03$ would cause a very serious falling off in gain. At 3,000 mc the total frequency variation of .002 times in v would be 6 mc. This is then a measure of the bandwidth of a series of resonators used in place of a helix for which $\beta a = 3$ and adjusted to give the same gain.

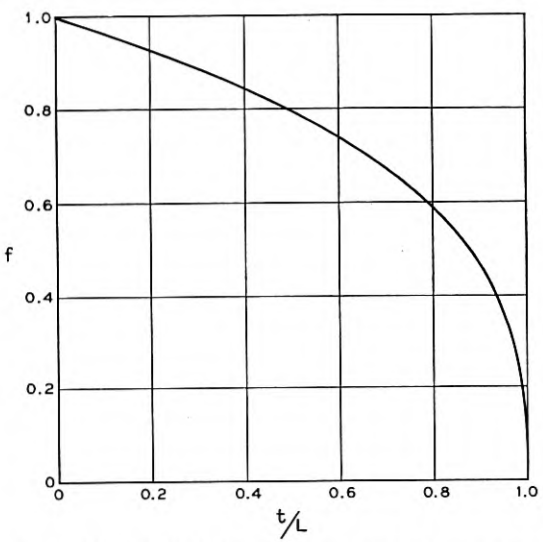


Fig. 5.9—The factor f by which $(E^2/\beta^2P)^{1/3}$ for a series of resonators such as those of Fig. 5.4 is reduced because of wall thickness t, in relation to gap spacing L.

5.4 Physical Limitations

In Section 3.3b the resonators were assumed to be very thin and to have walls of zero thickness. Of course the walls must have finite thickness, and it is impractical to make the resonators extremely thin. The wall thickness and the finite transit time across the resonators both reduce E^2/β^2P.

5.4a Effect of Wall Thickness

Consider the resonators of Fig. 5.4. Let L be the spacing between resonators ($1/L$ resonators per unit length), and t be the wall thickness. Thus, the gap length is $(L - t)$. Suppose we keep L and the voltage across each

resonator constant, so as to keep the field constant, but vary t. The capacitance will be proportional to $(L - t)^{-1}$ and, as the stored energy is the voltage squared times the capacitance, we see that $(E^2/\beta^2 P)^{1/3}$ will be reduced by a factor f,

$$f = (1 - t/L)^{1/3} \tag{5.27}$$

The factor f is plotted vs. t/L in Fig. 5.9.

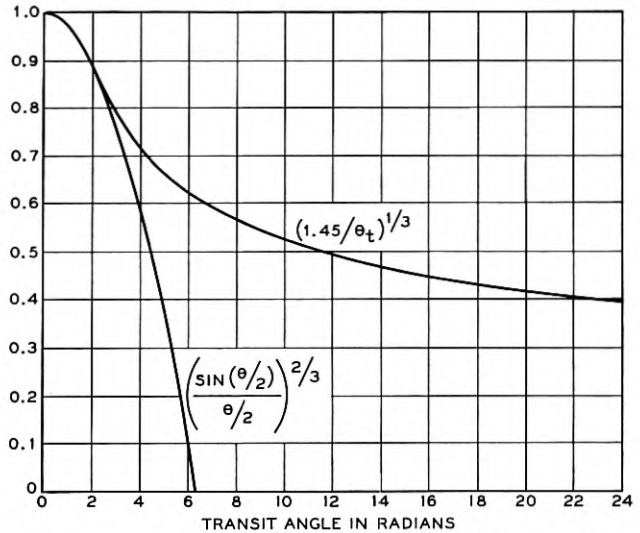


Fig. 5.10—The lower curve shows the factor by which $E^2/\beta^2 P$ is reduced by gap length, θ in radians. If the gap spacing is greater than 2.33 radians, it is best to make the gap 2.33 radians long. Then the upper curve applies.

5.4b Transit Time

As it is impractical to make the resonators infinitely thin, there will be some transit angle θ_g across the resonator, where

$$\theta_g = \beta \ell \tag{5.28}$$

Here ℓ is the space between resonator walls, or, the length of the gap. If we assume a uniform electric field between walls, the gap factor M, that is, the ratio of peak energy gained in electron volts to peak resonator voltage, or the ratio of the magnitude of the sinusoidal field component produced to that which would be produced by the same number of infinitely thin gaps with the same voltages, will be (from (4.69) with $r = a$)

$$M = \frac{\sin(\theta_g/2)}{\theta_g/2} \tag{5.29}$$

For a series of resonators θ_g long with infinitely thin walls $E^2/\beta^2 P$ will be less than the values given by (5.24) and (5.25) by a factor $M^{2/3}$. This is plotted vs. θ_g in Fig. 5.10.

5.4c Fixed Gap Spacing

Suppose it is decided in advance to put only one gap in a length specified by the transit angle θ_t. How wide should the gap be made, and how much will $E^2/\beta^2 P$ be reduced below the value for very thin resonators and infinitely thin walls?

Let us assume that all the stored energy is energy stored between parallel planes separated by the gap thickness, expressed in radians as θ or in distance as L

$$\theta_t = \beta \ell$$
$$\theta_g = \beta L$$

Here ℓ is the gap spacing and L is the spacing between resonators.

From Section 4.4 of Chapter IV we see that if V is the gap voltage, the field strength E is given by

$$E = MV/L$$

The stored energy per unit length, W, will be

$$W = W_0 V^2/\ell L \tag{5.30}$$

Here W_0 is a constant depending on the cross-section of the resonators. Thus, for unit field strength, the stored energy will be

$$\begin{aligned} W &= W_0 L/\ell M^2 \\ W &= W_0(\theta_t/\theta_g)(\theta_g/2)^2/\sin^2(\theta_g/2) \end{aligned} \tag{5.31}$$

We see that W_0 is merely the value of W when $\theta_t = \theta_g$ and $\theta_g = 0$, or, for zero wall thickness and very thin resonators. Thus, the ratio W/W_0 relates the actual stored energy per unit length per unit field to this optimum stored energy for resonators of the same cross section.

For $\theta_t < 2.33$, W/W_0 is smallest (best) for $\theta_g = \theta_t$ (zero wall thickness). For larger values of θ_t, the optimum value of θ_g is 2.33 radians and for this optimum value

$$(W_0/W)^{1/3} = (1.450/\theta_t)^{1/3} \tag{5.32}$$

If $\theta_t < 2.33$, it is thus best to make $\theta_g = \theta_t$. Then $(E^2/\beta^2 P)^{1/3}$ is reduced by the factor $[\sin(\theta/2)/(\theta/2)]^{2/3}$, which is plotted in Fig. 5.10. If $\theta_t > 2.33$, it is best to make $\theta = 2.33$. Then $(E^2/\beta^2 P)^{1/3}$ is reduced from the

value for thin resonators with infinitely thin walls by a factor given by (5.32), which is plotted vs. θ_t in Fig. 5.10.

If there are edge effects, the optimum gap spacing and the reduction in $(E^2/\beta^2 P)^{1/3}$ will be somewhat different. However, Fig. 5.10 should still be a useful guide.

In case of wide gap separation (large θ_t), there would be some gain in using reentrant resonators, as shown in Fig. 4.11, in order to reduce the capacitance. How good can such a structure be? Certainly, it will be worse than a helix. Consider merely the sections of metal tube with short gaps, which surround the electron beam. The shorter the gaps, the greater the capacitance. The space outside the beam has been capacitively loaded, which tends to reduce the impedance. This capacitance can be thought of as being associated with many spatial harmonics in the electric field, which do not contribute to interaction with the electrons.

5.5 Attenuation

Suppose we have a circuit made up of resonators with specified unloaded Q.† The energy lost per cycle is

$$W_L = 2\pi W_S/Q \tag{5.33}$$

In one cycle, however, a signal moves forward a distance L, where

$$L = v_0/f \tag{5.34}$$

The fractional energy loss per unit distance, which we will call 2α, is

$$2\alpha = \frac{W_L}{W_S}\frac{1}{L} \tag{5.35}$$

whence

$$\alpha = \frac{\omega}{2Qv_0} \tag{5.36}$$

So defined, α is the attenuation constant, and the amplitude will decay along the circuit as $\exp(-\alpha z)$.

The wavelength, λ, is given by

$$\lambda = v/f = 2\pi v/\omega \tag{5.37}$$

The loss per wavelength in db is

$$\text{db/wavelength} = 20 \log_{10} \exp(\alpha\lambda)$$

$$\text{db/wavelength} = \frac{27.3}{Q}\frac{v}{v_0} \tag{5.38}$$

† Disregarding coupling losses, the circuit and the resonantors will both have this same Q.

We see that, for given values of v and Q, decreasing the group velocity, which increases $E^2/\beta^2 P$, also increases the attenuation per wavelength.

5.5a Attenuation of Circuits

For various structures, Q can be evaluated in terms of surface resistivity, R, the intrinsic resistance of space, $\sqrt{\mu/\epsilon} = 377$ ohms, and varous other parameters. For instance, Schelkunoff[2] gives for the Q of a pill-box resonator

$$Q = \frac{1.20(\sqrt{\mu/\epsilon}/R)}{1 + a/h} \tag{5.39}$$

Here a is the radius of the resonator and h is the height. If we express the radius in terms of the resonant wavelength λ_0 ($a = 1.2\lambda_0/\pi$), we obtain

$$Q = \frac{\pi(\sqrt{\mu/\epsilon}/R)(v/c)}{(1 + h/a)n} \tag{5.40}$$

Here n is the number of resonators per wavelength (assuming the walls separating the resonators to be of negligible thickness); thus

$$n = h/\lambda = (h/\lambda_0)(c/v) \tag{5.41}$$

From (5.40) and (5.38) we obtain for a series of pill-box resonators

$$\text{db/wavelength} = 8.68(R/\sqrt{\mu/\epsilon})(c/v_g)(1 + h/a)n \tag{5.42}$$

In Appendix III an estimate of the Q of an array of fine half-wave parallel wires is made by assuming conduction in one direction with a surface resistance R. On this basis, Q is found to be

$$Q = (\sqrt{\mu/\epsilon}/R)(v/c) \tag{5.43}$$

and hence

$$\text{db/wavelength} = 27.3(R/\sqrt{\mu/\epsilon})(c/v_g) \tag{5.44}$$

For non-magnetic materials, surface resistance varies as the square root of the resistivity times the frequency. The table below gives R for copper and db/wavelength for pill-box resonators for $h/a \ll 1$ (5.42) and for wires (5.44) for several frequencies

f, mc	R, Ohms	(db/wavelength)/ (c/v_g)	
		Pill-box Resonators	Wires
3,000	.0142	$3.3 \times 10^{-4}n$	10.3×10^{-4}
10,000	.0260	$6.0 \times 10^{-4}n$	18.1×10^{-4}
30,000	.0450	$10.4 \times 10^{-4}n$	32.6×10^{-4}

In Section 3.3b a circuit made up of resonators, with a group velocity .031 times the phase velocity, was discussed. Suppose such a circuit were

[2] Electromagnetic Waves, S. A. Schelkunoff, Van Nostrand, 1943. Page 269.

used at 1,000 volts ($c/v = 16.5$), were 40 wavelengths long, and had three copper resonators per wavelength. The total attenuation in db is given below

f, mc	Attenuation, db
3,000	21
10,000	38
30,000	67

CHAPTER VI

THE CIRCUIT DESCRIBED IN TERMS OF NORMAL MODES

Synopsis of Chapter

IN CHAPTER II, the field produced by the current in the electron stream, which was assumed to vary as exp $(-\Gamma z)$, was deduced from a simple model in which the electron stream was assumed to be very close to an artificial line of susceptance B and reactance X per unit length. Following these assumptions, the voltage per unit length was found to be that of equation (2.10) and the field E in the z direction would accordingly be Γ times this, or

$$E = \frac{\Gamma^2 \Gamma_1 K}{\Gamma_1^2 - \Gamma^2} i \tag{6.1}$$

Here we will remember that Γ_1 is the natural propagation constant of the line, and K is the characteristic impedance.

We further replaced K by a quantity

$$E^2/\beta^2 P = 2K \tag{6.2}$$

where E is the field produced by a power flow P, and β is the phase constant of the line. For a lossless line, Γ_1 is a pure imaginary and

$$\beta^2 = -\Gamma_1^2 \tag{6.3}$$

From (6.1) and (6.2) we obtain

$$E = \frac{\Gamma^2 \Gamma_1 (E^2/\beta^2 P)}{2(\Gamma_1^2 - \Gamma^2)} i \tag{6.4}$$

To the writer it seems intuitively clear that the derivation of Chapter II is correct for waves with a phase velocity small compared with the velocity of light, and that (6.4) correctly gives the part of the field associated with the excitation of the circuit. However, it is clear that there are other field components excited; a bunched electron stream will produce a field even in the absence of a circuit. Further, many legitimate questions can be raised. For instance, in Chapter II capacitive coupling only was considered. What about mutual inductance between the electron stream and the inductances of the line?

The best procedure seems to be to analyze the situation in a way we know to be valid, and then to make such approximations as seem reasonable. One approximation we can make is, for instance, that the phase velocity of the wave is quite small compared with the speed of light, so that

$$| \Gamma_1 |^2 \gg \beta_0^2 = (\omega/c)^2 \tag{6.5}$$

In this chapter we shall consider a lossless circuit which supports a group of transverse magnetic modes of wave propagation. The finned structure of Fig. 4.3 is such a circuit, and so are the circuits of Figs. 4.8 and 4.9 (assuming that the fins are so closely spaced that the circuit can be regarded as smooth). It is assumed that waves are excited in such a circuit by a current in the z direction varying with distance as $\exp(-\Gamma z)$ and distributed normal to the z direction as a function of x and y, $\hat{J}(x, y)$. Such a current might arise from the bunching at low signal levels of a broad beam of electrons confined by a strong magnetic field so as not to move appreciably normal to the z direction.

The structure considered may support transverse electric waves, but these can be ignored because they will not be excited by the impressed current.

In the absence of an impressed current, any field distribution in the structure can be expressed as the sum of excitations of a number of pairs of normal modes of propagation. For one particular pair of modes, the field distribution normal to the z direction can be expressed in terms of a function $\hat{\pi}_n(x, y)$ and the field components will vary in the z direction as $\exp(\pm\Gamma_n z)$. Here the $+$ sign gives one mode of the pair and the $-$ sign the other. If Γ_n is real the mode is *passive*; the field decays exponentially with distance. If Γ_n is imaginary the mode is *active*; the field pattern of the mode propagates without loss in the z direction.

An impressed current which varies in the z direction as $\exp(-\Gamma z)$ will excite a field pattern which also varies in the z direction as $\exp(-\Gamma z)$, and as some function of x and y normal to the z direction. We may, if we wish, regard the variation of the field normal to the z direction as made up of a combination of the field patterns of the normal modes of propagation, the patterns specified by the functions $\hat{\pi}_n(x, y)$. Now, a pattern specified by $\hat{\pi}_n(x, y)$ coupled with a variation $\exp(\pm\Gamma_n z)$ in the z direction satisfies Maxwell's equations and the boundary conditions imposed by the circuit with *no* impressed current. If, however, we assume the same variation with x and y but a variation as $\exp(-\Gamma z)$ with z, Maxwell's equations will be satisfied only if there is an impressed current having a distribution normal to the z direction which also can be expressed by the function $\hat{\pi}_n(x, y)$.

Suppose we add up the various forced modes in such relative strength and phase that the total of the impressed currents associated with them is equal to the actual impressed current. Then, the sum of the fields of these

modes is the actual field produced by the actual impressed current. The field is so expressed in (6.44) where the current components J_n are defined by (6.36).

If it is assumed that there is only one mode of propagation, and if it is assumed that the field is constant over the electron flow, (6.44) can be put in the form shown in (6.47). For waves with a phase velocity small compared with the velocity of light, this reduces to (6.4), which was based on the simple circuit of Fig. 2.3.

Of course, actual circuits have, besides the one desired active mode, an infinity of passive modes and perhaps other active modes as well. In Chapter VII a way of taking these into account will be pointed out.

Actual circuits are certainly not lossless, and the fields of the helix, for instance, are not purely transverse magnetic fields. In such a case it is perhaps simplest to assume that the modes of propagation exist and to calculate the amount of excitation by energy transfer considerations. This has been done earlier[1], at first subject to the error of omitting a term which later[2] was added. In (6.55) of this chapter, (6.44) is reexpressed in a form suitable for comparison with this earlier work, and is found to agree.

Many circuits are not smooth in the z direction. The writer believes that usually small error will result from ignoring this fact, at least at low signal levels.

6.1 Excitation of Transverse Magnetic Modes of Propagation by a Longitudinal Current

We will consider here a system in which the natural modes of propagation are transverse magnetic waves. The circuit of Fig. 4.3, in which a slow wave is produced by finned structures, is an example. We will remember that the modes of propagation derived in Section 4.1 of Chapter IV were of this type. We will consider here that any structure the circuit may have (fins, for instance) is fine enough so that the circuit may be regarded as smooth in the z direction.

Any transverse electric modes which may exist in the structure will not be excited by longitudinal currents, and hence may be disregarded.

The analysis presented here will follow Chapter X of Schelkunoff's Electromagnetic Waves.

The divergence of the magnetic field H is zero. As there is no z component of field, we have

[1] J. R. Pierce, "Theory of the Beam-Type Traveling-Wave Tube," *Proc. I.R.E.*, Vol. 35, pp. 111–123, February, 1947.

[2] J. R. Pierce, "Effect of Passive Modes in Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 36, pp. 993–997, August, 1948.

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} = 0 \tag{6.6}$$

This will be satisfied if we express the magnetic field in terms of a "stream function", π

$$H_x = \frac{\partial \pi}{\partial y} \tag{6.7}$$

$$H_y = -\frac{\partial \pi}{\partial x} \tag{6.8}$$

π can be identified as the z component of the vector potential (the vector potential has no other components).

We will assume π to be of the form

$$\pi = \hat{\pi}\,(x, y)e^{-\Gamma z} \tag{6.9}$$

Here $\hat{\pi}\,(x, y)$ is a function of x and y only, which specifies the field distribution in any x, y plane.

We can apply Maxwell's equations to obtain the electric fields

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = j\omega\epsilon E_x$$

Using (6.7) and (6.8), and replacing differentiation with respect to z by multiplication by $-\Gamma$, we find

$$E_x = \frac{j\Gamma}{\omega\epsilon}\frac{\partial \pi}{\partial x} \tag{6.10}$$

Similarly

$$E_y = \frac{j\Gamma}{\omega\epsilon}\frac{\partial \pi}{\partial y} \tag{6.11}$$

We see that in an x, y plane, a plane perpendicular to the direction of propagation, the field is given as the gradient of a scalar potential V

$$V = (-j\Gamma/\omega\epsilon)\pi \tag{6.12}$$

This is because we deal with transverse magnetic waves, that is, with waves which have no longitudinal or z component of magnetic field. Thus, a closed path in an x, y plane, which is normal to the direction of propagation, will link no magnetic flux, and the integral of the electric field around such a path will be zero.

We can apply the curl relation and obtain E_z

$$\begin{aligned} \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= j\omega\epsilon E_z \\ E_z &= \frac{j}{\omega\epsilon}\left(\frac{\partial^2 \pi}{\partial x^2} + \frac{\partial^2 \pi}{\partial y^2}\right) \end{aligned} \tag{6.14}$$

Applying Maxwell's equations again, we have

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = j\omega\mu H_x$$
$$\frac{j}{\omega\epsilon}\frac{\partial}{\partial y}\left(\frac{\partial^2\hat{\pi}}{\partial x^2} + \frac{\partial^2\hat{\pi}}{\partial y^2}\right) + \frac{j\Gamma^2}{\omega\epsilon}\frac{\partial\hat{\pi}}{\partial y} = -j\omega\mu\frac{\partial\hat{\pi}}{\partial y} \tag{6.15}$$

This is certainly true if

$$\frac{\partial^2\hat{\pi}}{\partial x^2} + \frac{\partial^2\pi}{\partial y^2} = -(\Gamma^2 + \beta_0^2)\hat{\pi} \tag{6.16}$$

$$\beta_0 = \omega\sqrt{\mu\epsilon} = \omega/c \tag{6.17}$$

We find that this satisfies the other curl E relations as well.

From (6.16) and (6.14) we see that

$$E_z = (-j/\omega\epsilon)(\Gamma^2 + \beta_0^2)\hat{\pi}(x, y)e^{-\Gamma z} \tag{6.18}$$

For a given physical circuit, it will be found that there are certain real functions $\hat{\pi}_n(x, y)$ which are zero over the conducting boundaries of the circuit, assuring zero tangential field at the surface of the conductor, and which satisfy (6.16) with some particular value of Γ, which we will call Γ_n. Thus, as a particular example, for a square waveguide of width W some (but not all) of these functions are

$$\hat{\pi}_n(x, y) = \cos(n\pi y/W)\cos(n\pi x/W) \tag{6.19}$$

where n is an integer. We see from (6.10), (6.11) and (6.18) that this makes E_x, E_y and E_z zero at the conducting walls $x = \pm W/2$, $y = \pm W/2$.

Each possible real function $\hat{\pi}_n(x, y)$ is associated with two values of Γ_n, one the negative of the other. The Γ_n's are the natural propagation constants of the normal modes, and the $\hat{\pi}_n$'s are the functions giving their field distribution in the x, y plane. The $\hat{\pi}_n$'s can be shown to be orthogonal, at least in typical cases. That is, integrating over the region in the x, y plane in which there is field

$$\iint \hat{\pi}_n(x, y)\,\hat{\pi}_m(x, y)\,dx\,dy = 0$$
$$n \neq m \tag{6.20}$$

For a lossless circuit the various field distributions fall into two classes: those for which Γ_n is imaginary, called active modes, which represent waves which propagate without attenuation; and those for which Γ_n is real, which change exponentially with amplitude in the z direction but do not change in phase. The latter can be used to represent the disturbance in a waveguide below cutoff frequency, for instance.

If Γ_n is imaginary (an active mode) the power flow is real, while if Γ_n is real (a passive mode) the power flow is imaginary (reactive or "wattless" power).

The spatial distribution functions $\hat{\pi}_n$ and the corresponding propagation constants Γ_n are a means for specifying the electrical properties of a physical structure, just as are the physical dimensions which describe the physical structure and determine the various $\hat{\pi}_n$'s and Γ_n's. In fact, if we know the various π_n's and Γ_n's, we can determine the response of the structure to an impressed current without direct reference to the physical dimensions.

In terms of the $\hat{\pi}_n$'s and Γ_n's, we can represent any unforced disturbance in the circuit in the form

$$\sum_n \hat{\pi}_n(x, y)[A_n e^{-\Gamma_n z} + B_n e^{\Gamma_n z}] \tag{6.21}$$

Here A_n is the complex amplitude of the wave of the nth spatial distribution traveling to the right, and B_n the complex amplitude of the wave of the same spatial distribution traveling to the left.

It is of interest to consider the power flow in terms of the amplitude, A_n or B_n. We can obtain the power flow P by integrating the Poynting vector over the part of the x, y plane within the conducting boundaries

$$\begin{aligned} P &= \frac{1}{2}\int\int EXH^* \, ds \\ P &= \frac{1}{2}\int\int (E_x H_y^* - E_y H_x^*) \, dx \, dy \end{aligned} \tag{6.22}$$

By expressing the fields in terms of the stream function, we obtain

$$P = A_n A_n^* \left(\frac{-j\Gamma_n}{2\omega\epsilon}\right) \iint \left[\left(\frac{\partial \hat{\pi}_n}{\partial x}\right)^2 + \left(\frac{\partial \hat{\pi}_n}{\partial y}\right)^2\right] dx \, dy \tag{6.23}$$

We can transform this by integrating by parts (essentially Green's theorem). Thus

$$\int_{x_1}^{x_2} \frac{\partial \hat{\pi}_n}{\partial x} \frac{\partial \hat{\pi}_n}{\partial x} \, dx = \hat{\pi}_n \frac{\partial \hat{\pi}_n}{\partial x}\Bigg|_{x_1}^{x_2} - \int_{x_1}^{x_2} \hat{\pi}_n \frac{\partial^2 \hat{\pi}_n}{\partial x^2} \, dx \tag{6.24}$$

Here x_1 and x_2, the limits of integration, lie on the conducting boundaries where $\hat{\pi}_n = 0$, and hence the first term on the right is zero. Doing the same for the second term in (6.23), we obtain

$$P_n = A_n A_n^* \left(\frac{-j\Gamma_n}{2\omega\epsilon}\right) \iint \hat{\pi}_n \left(\frac{\partial^2 \hat{\pi}_n}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n}{\partial y^2}\right) dx \, dy \tag{6.25}$$

By using (6.16), we obtain

$$P_n = A_n A_n^* \left(\frac{j\Gamma_n}{2\omega\epsilon}\right)(\Gamma_n^2 + \beta_0^2) \iint (\hat{\pi}_n)^2 \, dx \, dy \tag{6.26}$$

It is also of interest to express the z component of the nth mode, E_{zn}, explicitly. For the wave traveling to the right we have, from (6.18),

$$E_{zn} = A_n \left(\frac{-j}{\omega\epsilon}\right)(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y) \tag{6.27}$$

Let the field at some particular position, say, $x = y = 0$, be E_{zn0}. Then

$$A_n = \frac{j\omega\epsilon E_{zn0}}{(\Gamma_n^2 + \beta_0^2)\,\hat{\pi}_n(0, 0)} \tag{6.28}$$

and from (6.26)

$$P_n = (E_{zn0} E_{zn0}^*) \frac{-j\omega\epsilon\Gamma_n}{2\hat{\pi}_n^2(0, 0)(\Gamma_n^2 + \beta_0^2)} \iint [\hat{\pi}_n(x, y)]^2 \, dx \, dy \tag{6.29}$$

We can rewrite this

$$\frac{E_{zn0} E_{zn0}^*}{(-\Gamma_n^2)P_n} = \frac{2\hat{\pi}_n^2(0, 0)(\Gamma_n^2 + \beta_0^2)}{-j\omega\epsilon\Gamma_n(-\Gamma_n^2) \iint [\hat{\pi}_n(x, y)]^2 \, dx \, dy} \tag{6.30}$$

For an active mode in a lossless circuit, Γ_n is a pure imaginary, and the negative of its square is the square of the phase constant. Thus, for a particular mode of propagation we can identify (6.30) with the circuit parameter $E^2/\beta^2 P$ which we used in Chapter II.

Let us now imagine that there is an impressed current J which flows in the z direction and has the form

$$J = \hat{J}(x, y)e^{-\Gamma z} \tag{6.31}$$

According to Maxwell's equations we must have

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\epsilon E_z + J \tag{6.32}$$

Now, we will assume that the fields are given by some overall stream function π which varies with x and y and with z as $\exp(-\Gamma z)$.

In terms of this function π, H_x, H_y and E_x, E_y will be given by relations (6.7), (6.8), (6.10), (6.11). However, the relation used in obtaining E_z is not valid in the presence of the convection current. Instead of (6.16) we have

$$\begin{aligned} \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= j\omega\epsilon E_z + J \\ E_z &= \frac{j}{\omega\epsilon}\left(\frac{\partial^2\pi}{\partial x^2} + \frac{\partial^2\pi}{\partial y^2}\right) + \frac{j}{\omega\epsilon} J \end{aligned} \tag{6.33}$$

Again applying the relation

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -j\omega\mu H_x$$

we obtain

$$\frac{\partial^2 \pi}{\partial x^2} + \frac{\partial^2 \pi}{\partial y^2} = -(\Gamma^2 + \beta_0^2)\,\pi - J \tag{6.34}$$

We will now divide both π and J into the spatial distributions characteristic of the normal unforced modes.
Let

$$\hat{J}(x, y) = \sum_n J_n \hat{\pi}_n(x, y) \tag{6.35}$$

$$J_n = \frac{\iint \hat{J}(x, y)\hat{\pi}_n(x, y)\,dx\,dy}{\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy} \tag{6.36}$$

This expansion is possible because the π_n's are orthogonal. Let

$$\hat{\pi} = e^{-\Gamma z} \sum_n C_n \hat{\pi}_n(x, y) \tag{6.37}$$

Here there is no question of forward and backward waves; the forced excitation has the same z-distribution as the forcing current.

For the nth component, we have, from (6.16),

$$\frac{\partial^2 \hat{\pi}_n(x, y)}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n(x, y)}{\partial y^2} = -(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y) \tag{6.38}$$

From (6.34) we must also have

$$C_n\left(\frac{\partial^2 \hat{\pi}_n(x, y)}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n(x, y)}{\partial y^2}\right) = -C_n(\Gamma^2 + \beta_0^2)\hat{\pi}_n(x, y) - J_n \hat{\pi}_n(x, y) \tag{6.39}$$

Accordingly, we must have

$$C_n = \frac{J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.40}$$

The overall stream function is thus

$$\pi = e^{-\Gamma z} \sum_n \frac{\hat{\pi}_n(x, y) J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.41}$$

From (6.33) and (6.34) we see that

$$E_z = \frac{-j}{\omega\epsilon}(\Gamma^2 + \beta_0^2)\pi \tag{6.42}$$

So

$$E_z = e^{-\Gamma z} \sum \frac{-j(\Gamma^2 + \beta_0^2)\hat{\pi}_n(x, y)J_n}{\omega\epsilon(\Gamma_n^2 - \Gamma^2)} \tag{6.43}$$

$$E_z = \frac{-j(\Gamma^2 + \beta_0^2)}{\omega\epsilon} e^{-\Gamma z} \sum \frac{\hat{\pi}_n(x, y)J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.44}$$

6.2 Comparison with Results of Chapter II

Let us consider a case in which there is only one mode of propagation, characterized by $\hat{\pi}_1(x, y)$, Γ_1, and a case in which the current flows over a region in which $\hat{\pi}_1(x, y)$ has a constant value, say, $\hat{\pi}_1(0, 0)$. This corresponds to the case of the transmission line which was discussed in Chapter II.

We take only the term with the subscript 1 in (6.44) and (6.30). Combining these equations, we obtain for the field at 0, 0

$$E_z = \frac{(E^2/\beta^2 P)(\Gamma^2 + \beta_0^2)}{(\Gamma_1^2 + \beta_0^2)} \frac{\Gamma_1^3 J_1 \iint [\hat{\pi}_1(x, y)]^2 \, dx \, dy}{2\hat{\pi}_1(0, 0)} \tag{6.45}$$

We have from (6.36)

$$J_1 = \frac{\pi_1(0, 0)}{\iint [\hat{\pi}_1(x, y)]^2 \, dx \, dy} \tag{6.46}$$

From (6.45) and (6.46) we obtain

$$E_z = \frac{(\Gamma^2 + \beta_0^2)\Gamma_1^3(E^2/\beta^2 P)}{2(\Gamma_1^2 + \beta_0^2)(\Gamma_1^2 - \Gamma^2)} Je^{-\Gamma z} \tag{6.47}$$

Let us compare this with (6.4), which came from the transmission line analogy of Chapter II, identifying E_z and J with E and i. We see that, for slow waves for which

$$\beta_0^2 \ll |\Gamma_1^2| \tag{6.48}$$

$$\beta_0^2 \ll |\Gamma^2| \tag{6.49}$$

(6.47) becomes the same as (6.4). It was, of course, under the assumption that the waves are slow that we obtained (2.10), which led to (6.4).

6.3 Expansion Rewritten in Another Form

Expression (6.44) can be rewritten so as to appear quite different. We can write

$$\Gamma^2 + \beta_0^2 = \Gamma^2 - \Gamma_n^2 + \Gamma_n^2 + \beta_0^2$$

Thus, we can rewrite the expression for E_z as

$$E_z = e^{-\Gamma z}\left((-j/\omega\epsilon)\sum_n \frac{(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y)J_n}{\Gamma_n^2 - \Gamma^2} + (j/\omega\epsilon)\sum_n \hat{\pi}_n(x, y)J_n\right) \tag{6.50}$$

The second term in the brackets is just $j/\omega\epsilon$ times the impressed current, as we can see from (6.35). The first term can be rearranged

$$(-j/\omega\epsilon)(\Gamma_n^2 + \beta_0^2)J_n = \frac{(-j/\omega\epsilon)(\Gamma_n^2 + \beta_0^2)\iint \hat{\pi}_n(x, y)J(x, y)\,dx\,dy}{\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy} \tag{6.51}$$

Referring back to (6.29), let Ψ_n be twice the power P_n carried by the unforced mode when the field strength is

$$|E_{zn0}| = 1 \tag{6.52}$$

Further, let us choose the $\hat{\pi}_n$'s so that, at some specified position, $x = y = 0$,

$$_n(0, 0) = 1 \tag{6.53}$$

Then

$$\Psi_n = \frac{-j\omega\epsilon\Gamma_n}{\Gamma_n^2 + \beta_0^2}\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy \tag{6.54}$$

Using this in connection with (6.51), we obtain

$$E_z = e^{-\Gamma z}\left(-\sum_n \frac{\Gamma_n\hat{\pi}_n(x, y)\iint \hat{\pi}_n(x, y)\hat{J}(x, y)\,dx\,dy}{\Psi_n(\Gamma_n^2 - \Gamma^2)} + (j/\omega\epsilon)\hat{J}(x, y)\right) \tag{6.55}$$

An expression for the forced field in terms of the parameters of the normal modes was given earlier[1,2]. In deriving this expression, the existence of a set of modes was assumed, and the field at a point was found as an integral over the disturbances induced in the circuit to the right and to the left and propagated to the point in question. Such a derivation applies for lossy and mixed waves, while that given here applies for lossless transverse-magnetic waves only.

The earlier derivation[1] leads to an expression identical with (6.55) except that Ψ_n^* appears in place of Ψ_n. In this earlier derivation a sign was implicitly assigned to the direction of flow of reactive power (which really doesn't flow at all!) by saying that the reactive power flows in the direction in which the amplitude decreases. If we had assumed the reactive power to flow in the direction in which the amplitude increases, then, with the same definition of Ψ_n, for a passive mode Ψ_n^* would have been replaced by $-\Psi_n^*$ which is equal to Ψ_n (for a passive mode, Ψ_n is imaginary).

In deriving (6.55), no such ambiguity arose, because the power flow was identified with the complex Poynting vector for the particular type of wave considered. In any practical sense, Ψ is merely a parameter of the circuit, and it does not matter whether we call Im Ψ reactive power flow to the right or to the left.

The existence of a derivation of (6.55) not limited in its application to lossless transverse magnetic waves is valuable in that practical circuits often have some loss and often (in the case of the helix, for instance) propagate mixed waves.

6.4 Iterated Structures

Many circuits, such as those discussed in Chapter IV, have structure in the z direction. Expansions such as (6.55) do not strictly apply to such structures. We can make a plausible argument that they will be at least useful if all field components except one differ markedly in propagation constant from the impressed current. In this case we save the one component which is nearly in synchronism with the impressed current and hope for the best.

APPENDIX III

STORED ENERGIES OF CIRCUIT STRUCTURES

A3.1 Forced Sinusoidal Field

If $v \ll c$, the field can be very nearly represented inside the cylinder of radius a by

$$V = V_0 \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} = \frac{E}{j\beta} \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} \tag{1}$$

and outside by

$$V = V_0 \frac{K_0(\gamma r)}{K(\gamma a)} e^{-j\beta z} \tag{2}$$

Inside

$$\frac{\partial V}{\partial r} = \beta \frac{I_1(\beta r)}{I_0(\beta a)} e^{-j\beta z} V_0 \tag{3}$$

$$\frac{\partial V}{\partial z} = -j\beta \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} V_0 \tag{4}$$

Outside

$$\frac{\partial V}{\partial r} = -\beta \frac{K_1(\beta r)}{K_0(\beta a)} e^{-j\beta z} V_0 \tag{5}$$

$$\frac{\partial V}{\partial z} = -j\beta \frac{K_0(\beta r)}{K_0(\beta a)} e^{-j\beta z} V_0 \tag{6}$$

Because there is a sinusoidal variation in the z direction, the average stored electric energy per unit length will be

$$W_E = \left(\frac{1}{2}\right)\left(\frac{\epsilon}{2}\right) \int_{r=0}^{\infty} [(E_{r\,\max})^2 + (E_{z\,\max})^2](2\pi r\, dr) \tag{7}$$

Here $E_{r\ max}$ and $E_{z\ max}$ are maximum values at $r = a$. The total electric plus magnetic stored energy will be twice this. This gives

$$W = \frac{\pi\epsilon(\gamma a)^2}{2\gamma^2}\left[\frac{I_0^2 - I_0 I_2}{I_0^2} + \frac{K_0 K_2 - K_0^2}{K_0^2}\right] E^2$$

$$W = \frac{\pi\epsilon\gamma a}{\gamma^2}\left[\frac{I_1}{I_0} + \frac{K_1}{K_0}\right] E^2 \tag{8}$$

$$(E^2/\beta^2 P)^{1/3} = (c/v)^{1/3}(v/v_g)^{1/3}\left[\frac{120}{\beta a\left(\frac{I_1}{I_0} + \frac{K_1}{K_0}\right)}\right]^{1/3} \tag{9}$$

A3.2 Pill-Box Resonators

Schelkunoff gives on page 268 of Electromagnetic Waves an expression for the peak electric energy stored in a pill-box resonator, which may be written as

$$.135\ \pi\ \epsilon\ a^2 h E^2$$

Here a is the radius of the resonator and h is the axial length. For a series of such resonators, the peak stored electric energy per unit length, which is also the average electric plus magnetic energy per unit length, is

$$W = .135\ \pi\ \epsilon\ a^2 E^2 \tag{10}$$

For resonance

$$a = 1.2\lambda_0/\pi \tag{11}$$

Whence

$$W = .0618\ \epsilon\lambda_0^2 E^2 \tag{12}$$

And

$$(E^2/\beta^2 P)^{1/3} = 5.36\ (v/v_g)^{1/3}\ (v/c)^{1/3} \tag{13}$$

The case of square resonators is easily worked out.

A3.3 Parallel Wires

Let us consider very fine very closely spaced half-wave parallel wires with perpendicular end plates.

If z is measured along the wires, and y perpendicular to z and to the direction of propagation, the field is assumed to be

$$E_x = E\cos\beta x e^{\pm\beta y}\cos\frac{2\pi}{\lambda_0}z$$

$$E_y = E\sin\beta x e^{\pm\beta y}\cos\frac{2\pi}{\lambda_0}z \tag{14}$$

Here the + sign applies for $y < 0$ and the − sign for $y > 0$. We will then find that

$$W = 2W_E = \frac{\epsilon E^2 \lambda_0}{2} \int_0^\infty e^{-2\beta y}\, dy$$

$$W = \frac{\epsilon \lambda_0}{4\beta} E^2 \tag{15}$$

and

$$(E^2/\beta^2 P)^{1/3} = 6.20\ (v/v_g)^{1/3} \tag{16}$$

The surface charge density σ on one side of the array of wires (say, $y > 0$) is given by the y component of field at $y = 0$.

$$\sigma = \epsilon E_y = \epsilon E \sin \beta x \cos \frac{2\pi}{\lambda_0} z \tag{17}$$

This is related to the current I (flowing in the z direction) per unit distance in the x direction by

$$\frac{\partial I}{\partial z} = -\frac{\partial \sigma}{\partial t} \tag{18}$$

From (18) and (17) we obtain for the current on one side of the array

$$I = -\frac{j\omega\lambda_0 \epsilon}{2\pi} E \sin \beta x \sin \frac{2\pi}{\lambda_0} z \tag{19}$$

If we use the fact that $\omega\lambda_0/2\pi = c$ and $c\,\epsilon = 1/\sqrt{\mu/\epsilon}$, we obtain

$$I = \frac{-jE}{\sqrt{\mu/\epsilon}} \sin \beta x \sin \frac{2\pi}{\lambda_0} z \tag{20}$$

If R is the surface resistivity of either side ($y > 0$, $y < 0$) of the wires, when the wires act as a resonator (a standing wave) the average power lost per unit length for both sides is

$$P = \tfrac{1}{8} R \lambda_0 E^2/(\mu/\epsilon) \tag{21}$$

In this case the stored electric energy is half the value given by (15), and we find

$$Q = (\sqrt{\mu/\epsilon}/R)\ (v/c) \tag{22}$$

Factors Affecting Magnetic Quality*

By R. M. BOZORTH

IN THE preparation of magnetic materials for practical use it is important to know how to obtain products of the best quality and uniformity. In the scientific study of magnetism the goal is to understand the relation between the structure and composition on the one hand and the magnetic properties on the other. From both standpoints it is necessary to know the principal factors which influence magnetic behavior. These are briefly reviewed here.

The properties depend on chemical composition, fabrication and heat-treatment. Some properties, such as saturation magnetization, change only slowly with chemical composition and are usually unaffected by fabrication or heat treatment. On the contrary, permeability, coercive force and hysteresis loss are highly sensitive and show changes which are extreme among all the physical properties. Properties may thus be divided into *structure-sensitive* and *structure-insensitive* groups. As an example, Fig. 1 shows magnetization curves of permalloy after it has been (a) cold rolled, (b) annealed and cooled slowly, and (c) annealed and cooled rapidly. The maximum permeability varies with the treatment over a range of about 20 fold, while the saturation induction is the same within a few per cent. Structure sensitive properties such as permeability depend on small irregularities in atomic spacings, which have little effect on properties such as saturation induction.

Some of the more common sensitive and insensitive properties are listed in Table I. The principal physical and chemical factors which affect these properties are listed in column 3. Their various effects will now be briefly discussed and illustrated.

Phase Diagram

Some of the most drastic changes in properties occur when the fabrication or heat treatment has brought about a change in structure of the material. For this reason the phase diagram or constitutional diagram is of the utmost importance in relation to the preparation and properties of magnetic materials. As an example consider the phase diagram of the binary iron-cobalt alloys of Fig. 2. Here the various areas show the phases, of different

*This article is the substance of Chapter II of a book entitled "Ferromagnetism" to be published early in 1951 by D. Van Nostrand Company, Inc.

composition or structure, which are stable at the temperatures and compositions indicated. The α phase has the body-centered-cubic crystal structure characteristic of iron. At 910°C it transforms into the face-centered phase γ, and at 1400° into the δ phase, which has the same structure as the α phase. At about 400°C cobalt transforms, on heating, from the ϵ phase (hexagonal structure) into the γ phase.

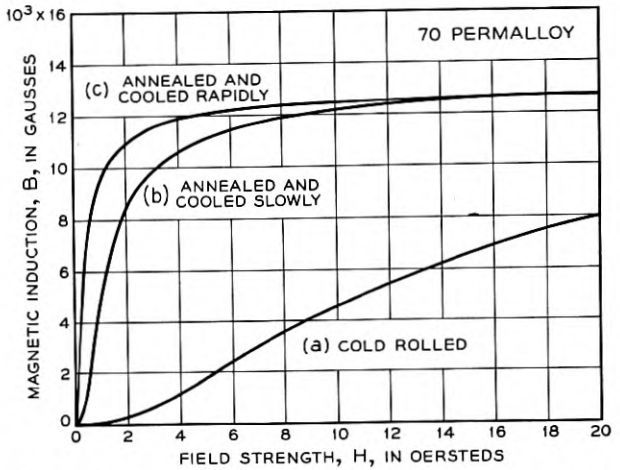


Fig. 1—Effect of mechanical and heat treatment on the magnetization curve of 70 permalloy (70% Ni, 30% Fe).

TABLE I

Properties Commonly Sensitive or Insensitive to Small Changes in Structure, and Some of the Factors which Effect Such Changes

Structure-Insensitive Properties	Structure-Sensitive Properties	Factors Affecting the Properties
I_s, Saturation Magnetization	μ, Permeability	Composition (gross)
θ, Curie Point	H_c Coercive Force	Impurities
λ_s, Magnetostriction at Saturation	W_h Hysteresis Loss	Strain
K, Crystal Anisotropy Constant		Temperature
		Crystal Structure
		Crystal Orientation

The dotted lines indicate the Curie point, at which the material becomes non-magnetic.

In between the areas corresponding to the single phases α, γ, δ and ϵ there are two-phase regions in which two crystal structures co-exist, some of the crystal grains having one structure and others the other. Such a two-phase structure is usually evident upon microscopic or X-ray examina-

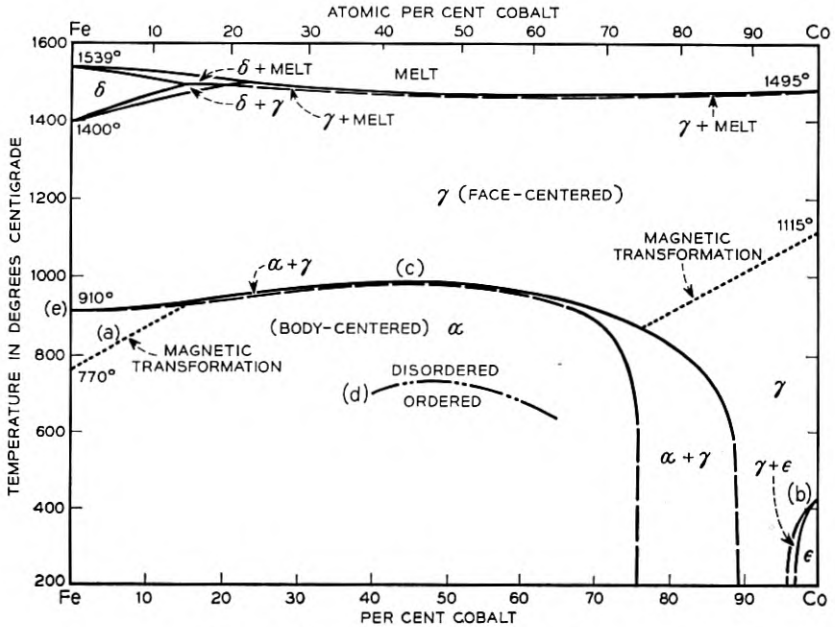


Fig. 2—Phase diagram of iron-cobalt alloys.

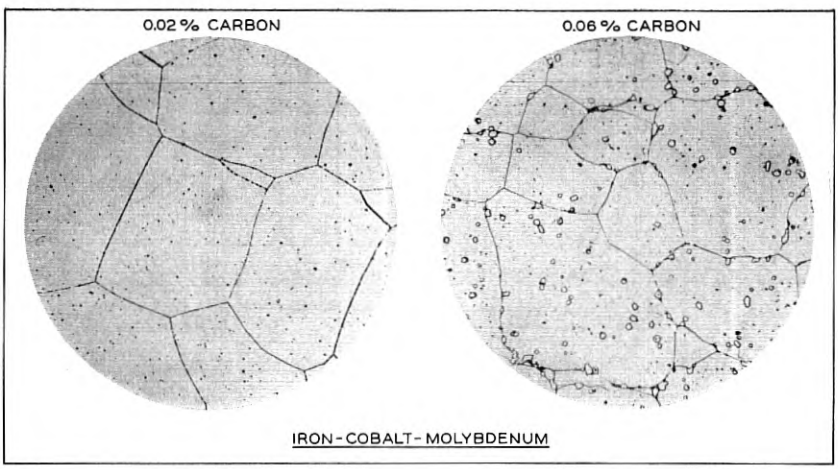


Fig. 3—Photomicrographs of remalloy (12% Co, 17% Mo, 71% Fe) showing the precipitation of a second phase in the specimen containing an excess of carbon (0.06%) Courtesy of E. E. Thomas. Magnification: (a) 50 times, (b) 200 times.

tion. Microphotographs of a single-phase alloy and a two-phase alloy of iron-cobalt-molybdenum are reproduced in Fig. 3 (a) and (b).

The diagram of Fig. 2 shows several kinds of changes that affect the magnetic properties. At (a) the material becomes non-magnetic on heating, without change in phase. At (b) there is a change of phase, both phases

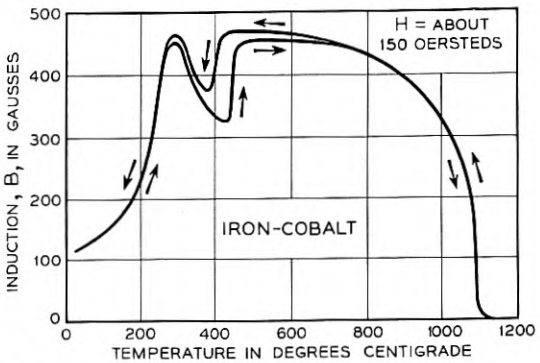


Fig. 4—Effect of phase transformation of cobalt on magnetization with a constant field of 150 oersteds. Both phases magnetic. *Masumoto.*

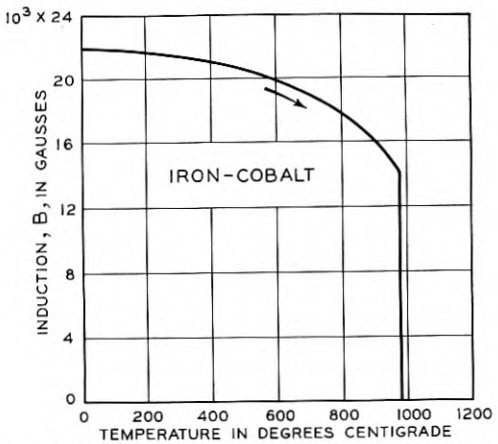


Fig. 5—Phase transformation in iron-cobalt alloy (50% Co). High-temperature phase is non-magnetic.

being magnetic. Figure 4 shows the changes in magnetic properties that occur during this latter transition; they are due partly to the high local strains that result from the change in structure, and partly to the difference in the crystal structures of the two phases. At (c) there is a change from a ferromagnetic to a non-magnetic phase, and Fig. 5 shows the rapid change in magnetization that occurs when the temperature rises in this area. At

(d) the α phase becomes ordered on cooling, i.e., the iron and cobalt atoms tend to distribute themselves regularly among the various atom positions so that each atom is surrounded by atoms of the other kind. This phenomenon is especially important in connection with the properties of iron-aluminum and manganese-nickel alloys.

The transition at (e) is entirely in the non-magnetic region but it has its influence on the properties of iron at room temperature. If iron is cooled very slowly through (e), the internal strains caused by the change in structure will be relieved by diffusion of the metal atoms, but if the cooling is too rapid there will not be sufficient time for strain relief. Practically this means that to obtain high permeability in iron it must be annealed for some time below 900°C, or cooled slowly through this temperature so that diffusion will have time to occur. In most ferromagnetic materials diffusion occurs at a reasonably rapid rate only at temperatures above about 500 to 600°C.

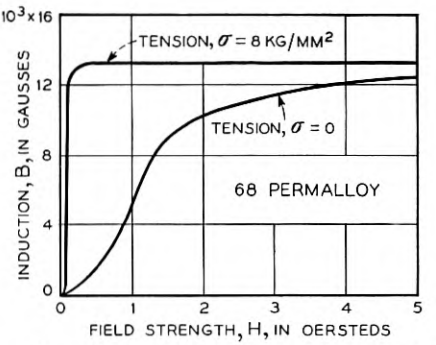


Fig. 6—Effect of tension on the magnetization curve of 68 permalloy.

The effect of a homogeneous *strain* on the magnetization curve can be observed in a simple way, as by applying tension to an annealed wire and then measuring B and H. The effect of tension on some materials is to increase the permeability and on other materials to decrease it, as shown in Fig. 6. Compression usually causes a change in the opposite sense.

The internal strains resulting from *plastic deformation* of the material, brought about by stressing beyond the elastic limit, as by pulling, rolling or drawing, almost always reduce the permeability. The material is then under rather severe local strains similar to those present after phase change, and these strains are different in magnitude and direction in different places in the material and have quite different values at points close together. Strains of this kind can usually be relieved by annealing; therefore, metal that has been fabricated by plastic deformation is customarily annealed to raise its permeability. Figure 1 shows the effect of annealing a permalloy strip that has been cold-rolled to 15 per cent of its original thickness.

The *temperature* also is effective in changing permeability and other properties, even when no change in phase occurs. Figure 7 shows the rapidity with which the initial permeability decreases as the Curie point is approached. For this material, Ferroxcube III, a zinc manganese ferrite ($ZnMnFe_4O_8$), the Curie point is not far above room temperature.

The effect of *impurities* may be illustrated by the B vs H curves for iron containing various amounts of carbon. Curve (a) of Fig. 8 is for a mild

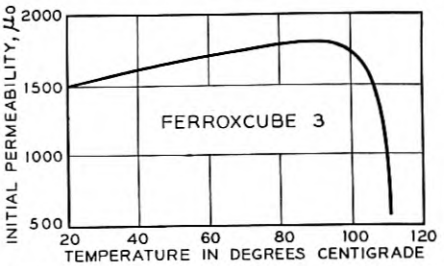


Fig. 7—Variation of initial permeability of Ferroxcube 3, showing maximum at temperature just below the Curie temperature.

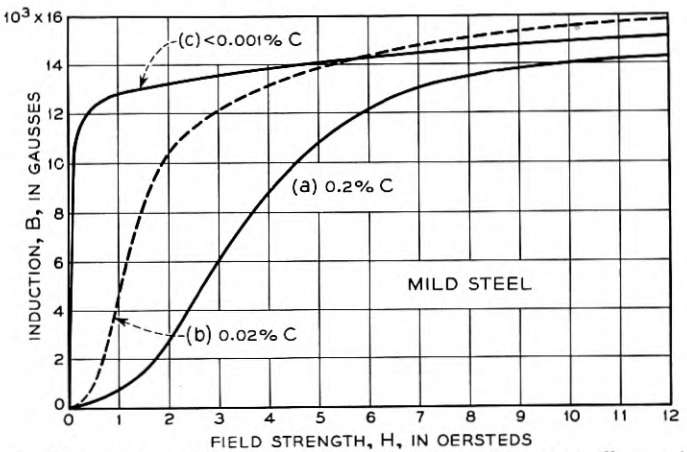


Fig. 8—Effect of impurities on magnetic properties of iron. Annealing at 1400°C in hydrogen reduces the carbon content from about 0.02 per cent to less than 0.001 per cent.

steel having 0.2 per cent carbon, (b) is for the iron commonly used in electromagnetic apparatus—it contains about 0.02 per cent carbon and is annealed at about 900°C. When this same iron is purified by heating for several hours at 1400°C in hydrogen, the carbon is reduced to less than 0.001 per cent and other impurities are removed, and curve (c) is obtained.

Finally, Fig. 9 shows that large differences in permeability may be found by simply varying the *direction of measurement* of the magnetic properties in a single specimen. The material is a single crystal of iron containing about 4 per cent silicon, and the directions in which the properties are measured

are [100] (parallel to one of the crystal axes), and [111] (as far removed as possible from an axis). The magnetic properties in the two directions are different because different "views" of the atomic arrangement are obtained in the two directions.

Production of Magnetic Materials

In the preparation of magnetic materials for either laboratory or commercial use there are many processes which influence the chemical and physical

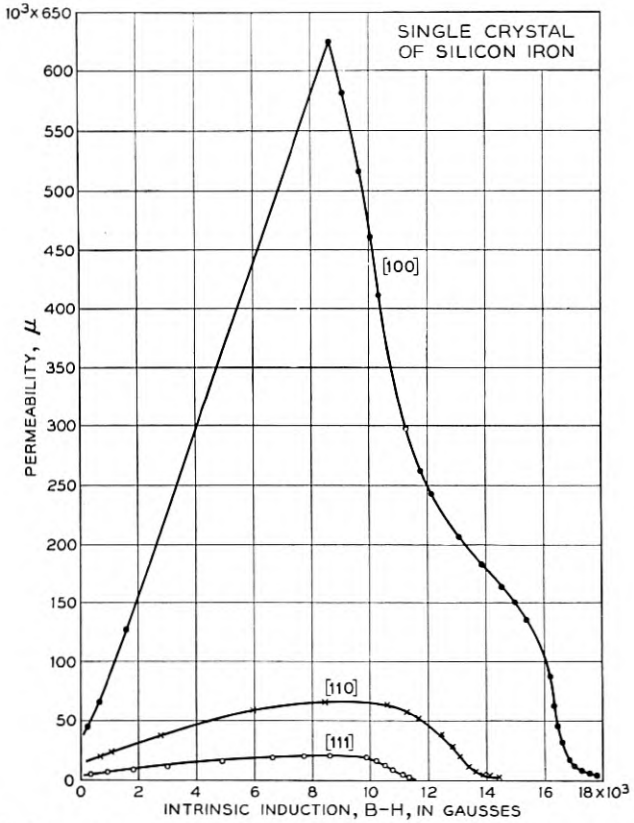


Fig. 9—Dependence of permeability on crystallographic direction. *Williams.*

structure of the product. The selection of raw materials, the melting and casting, the fabrication and the heat treatment, are all important and must be carried out with a proper knowledge of the metallurgy of the material. A brief description of the common practices is now given. For further discussion the reader is referred to more detailed metallurgical books and articles.

Melting and Casting

For experimental investigation of magnetic materials in the laboratory, the raw materials easily obtainable on the market are generally satisfactory. When high purity is desirable specially prepared materials and crucibles must be used and the atmosphere in contact with the melt must be controlled. The impurities that have the greatest influence on the magnetic properties of high permeability materials are the non-metallic elements,

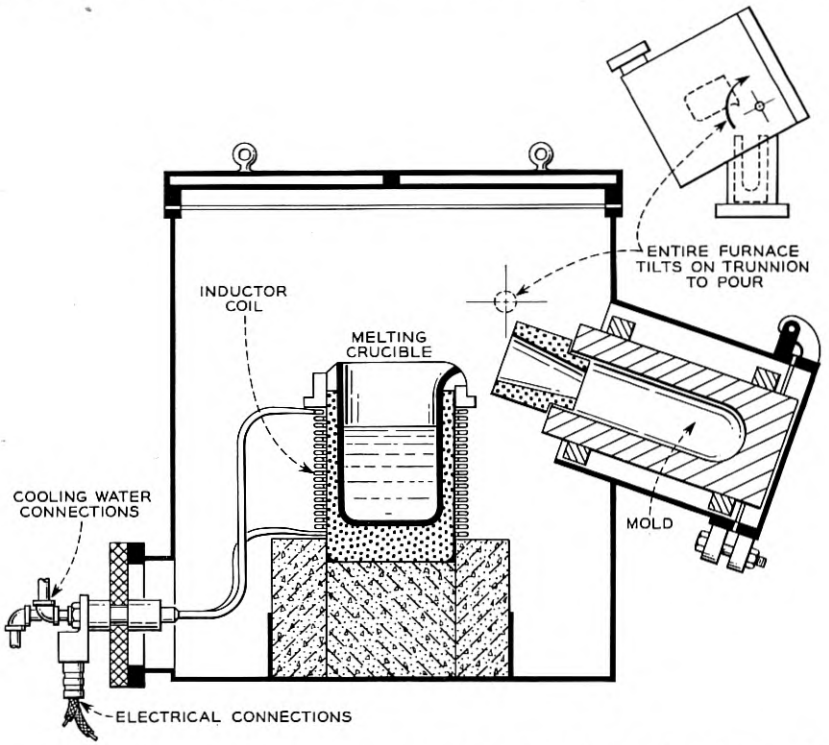


Fig. 10—Induction furnace designed for small melts in controlled atmosphere, as designed by J. H. Scaff and constructed by the Ajax Northrup Company.

particularly oxygen, carbon and sulfur, and the presence of these impurities is therefore watched carefully and their analyses are carried out with special accuracy. Impurities are likely to change in important respects during the melting and pouring on account of reactions of the melt with the atmosphere, the slag or the crucible lining, or because of reactions taking place among the constituents of the metal.

Melting of small lots (10 pounds) is best carried out in a high-frequency induction furnace. Figure 10 shows such a furnace designed for melting ten to fifty pounds, and casting by tilting the furnace, the whole operation being

carried out in a controlled atmosphere. High-frequency currents (usually 1,000 to 2,000 cycles/sec but sometimes much higher) are passed through the water-cooled copper coils, and the alternating magnetic field so produced

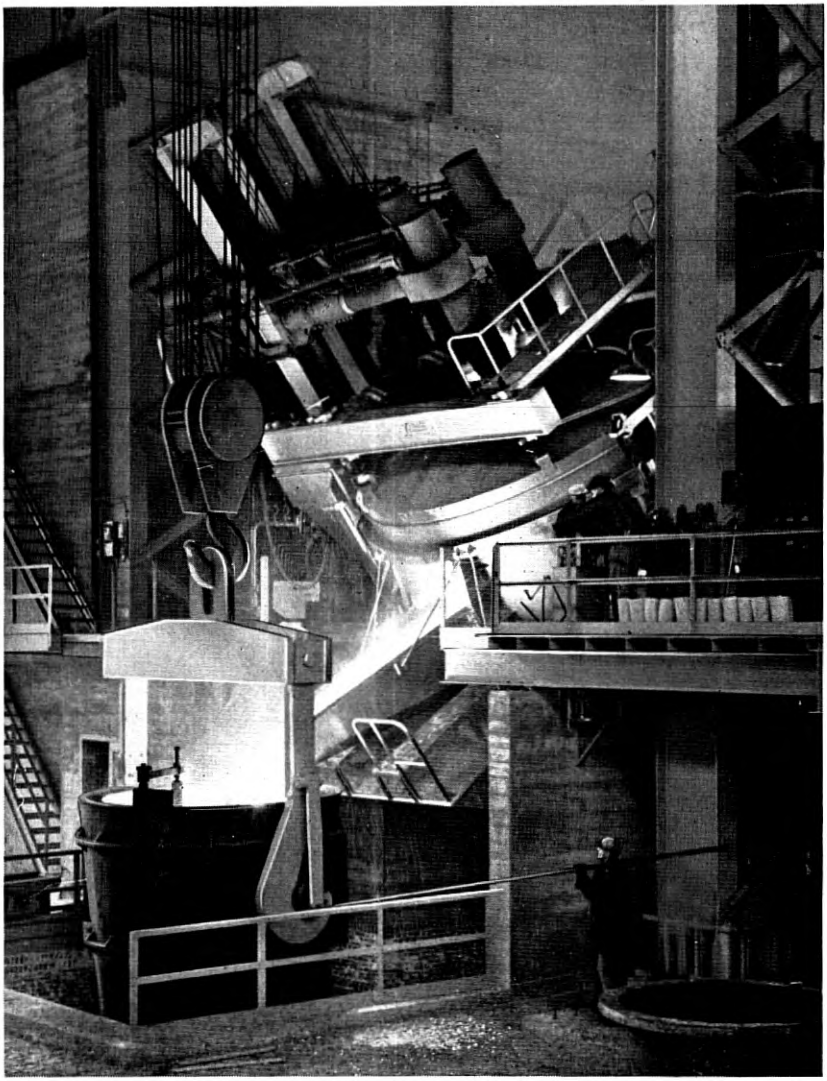


Fig. 11—Arc furnace for large commercial melts. Courtesy of J. S. Marsh of the Bethlehem Steel Company.

heats the charge by inducing eddy currents in it. Crucibles are usually composed of alumina or magnesia.

On a commercial scale melts of silicon-iron are usually made in the open

hearth furnace, in which pig-iron and scrap are refined and ferro-silicon added. The furnace capacity may be as large as 100 tons. Sometimes silicon-iron, and usually iron-nickel alloys, are melted in the arc furnace, in amounts varying from a few tons to 50 tons. A photograph of such a furnace, in the position of pouring, is shown in Fig. 11. The heat is produced in the arc drawn between large carbon electrodes immersed in the metal, the current sometimes rising to over 10,000 amperes. By tipping the furnace the melt is poured into a ladle, and from this it is poured into cast-iron molds through a valve-controlled hole in the ladle bottom. Special-purpose alloys, including permanent magnets, are prepared commercially in high-

TABLE II

Heats of Formation and Other Properties of Some Oxides (Sachs and Van Horn[4])

Oxide	Heat of formation (Kilo-cal per gram atom of metal)	Melting Point (°C)	Density (g/cm³)
CaO	152	>2500	3.4
BeO	144	>2500	3.0
MgO	144	2800	3.65
Li_2O	141	>1700	2.0
Al_2O_3	127	2050	3.5
V_2O_2	116	1970	4.9
TiO_2	109	1640	4.3
Na_2O	101	*	2.3
SiO_2	95	1670	2.3
B_2O_3	94	580	1.8
MnO	91	1650	5.5
ZrO_2	89	2700	5.5
ZnO	85	*	5.5
P_2O_5	73	*	2.4
SnO_2	68	1130	6.95
FeO	66	1420	5.7
NiO	58	**	7.45

* Sublimes.
** Decomposes before melting.

frequency induction furnaces or in arc furnaces in quantities ranging from a fraction of a ton to several tons.

Slags are commonly used when melting in air, both to protect from oxidation and to reduce the amounts of undesirable impurities. Common protective coverings are mixtures of lime, magnesia, silica, fluorite, alumina, and borax in varying proportions. In commercial production different slags are used at different stages, to refine the melt; e.g., iron oxide may be used to decarburize and basic oxides to desulfurize.

Melting in vacuum requires special technique that has been described in some detail by Yensen.[1] Commercial use has been described by Rohn[2] and others.[3] Melting in hydrogen has been used on an experimental scale in both

[1] T. D. Yensen, *Trans. A.I.E.E.* 34, 2601–41 (1915).
[2] W. Rohn, Heraeus Vacuumschmelze, Albertis, Hanau, 356–80 (1933).
[3] W. Hessenbruch and K. Schichtel, *Zeits. f. Metallkunde* 36, 127–30 (1944).

high-frequency and resistance-wound furnaces. In commercial furnaces Rohn has used hydrogen and vacuum alternately before pouring, for purification in the melt, in low-frequency induction furnaces having capacities of several tons.

Just before casting a melt of a high-permeability alloy such as iron nickel, a deoxidizer may be added, e.g. aluminum, magnesium, calcium or silicon, in an amount averaging around 0.1 per cent. The efficacy of a deoxidizer is measured by its heat of formation, and this is given for the common ele-

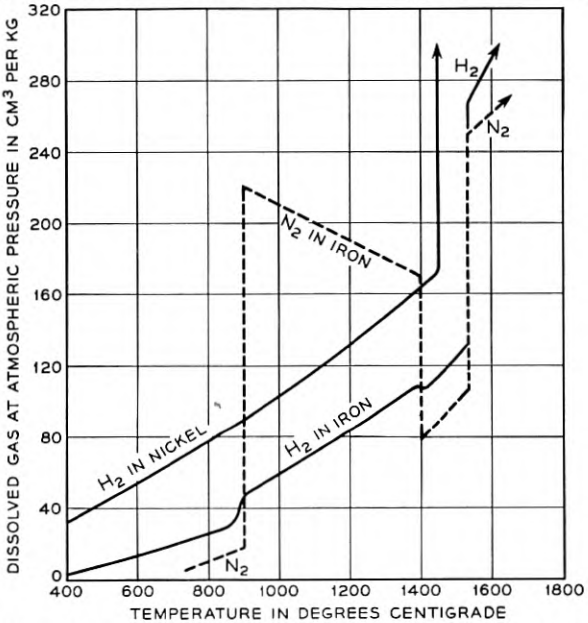


Fig. 12—Solubility of some gases in iron and nickel at various temperatures. *Sieverts.*

ments in Table II, taken from Sachs and Van Horn.[4] Also several tenths of a per cent of manganese may be put in to counteract the sulfur so that the material may be more readily worked; the manganese sulfide so formed collects into small globular masses which do not interfere seriously with the magnetic or mechanical properties of most materials.

Ordinarily a quantity of gas is dissolved in molten metal, and this is likely to separate during solidification and cause unsound ingots. The solubilities of some gases in iron and nickel have been determined by Sieverts[5] and others and are given in Fig. 12, adapted from the compilation by Dushman.[6] The characteristic decrease of solubility during freezing is apparent. Most

[4] G. Sachs and K. R. Van Horn, Practical Metallurgy, Am. Soc. Metals, Cleveland (1940).
[5] A. Sieverts, *Zeits. f. Metallkunde* 21, 37–46 (1929).
[6] S. Dushman, Vacuum Technique, Wiley, New York (1949).

of the gases given off by magnetic metals during heating are formed from the impurities carbon, oxygen, nitrogen and sulfur; CO is usually given off in greatest amount from cast metal, and some N_2 and H_2 are also found. Refining of the melt is therefore of obvious advantage, and the furnace of Fig. 10 is especially useful for this purpose.

Small ingots are sometimes made by cooling in the crucible. Usually, however, ingots are poured into cast iron molds for subsequent reduction by rolling, etc.; permanent magnet or other materials are often cast in sand

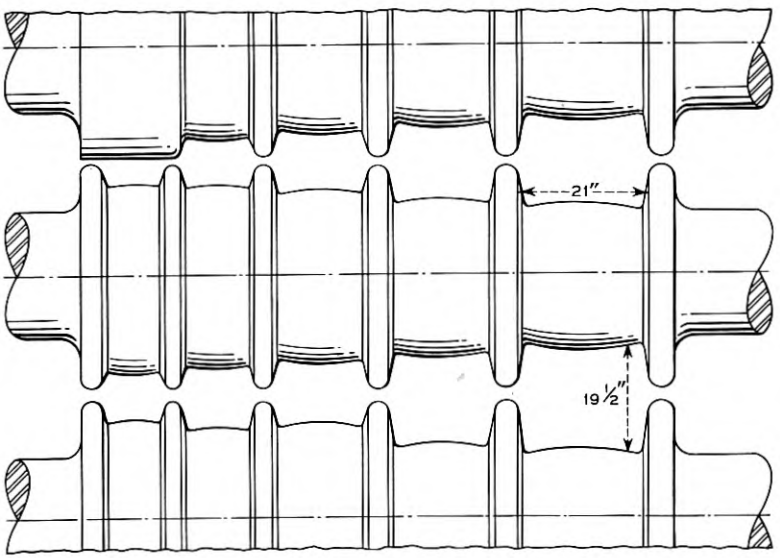


Fig. 13—Design of rolls in a blooming mill for hot reduction of ingots to rod. *Carnegie Illinois Steel Corp.*

in shapes which require only nominal amounts of machining or grinding for use in apparatus or in testing. Special techniques are used for specific materials.

Other considerations important in the melting and pouring of ingots are proper mixing in the melt, the temperature of pouring, mold construction, inclusions of slag, segregation, shrinkage, cracks, blow holes, etc.

Fabrication

Magnetic materials require a wide variety of modes of fabrication, which can best be discussed in connection with the specific materials. The methods include hot and cold rolling, forging, swaging, drawing, pulverization, elec-

trodeposition, and numerous operations such as punching, pressing and spinning. In the commercial fabrication of ductile material it is common practice to start the reduction in a breakdown or blooming mill (Fig. 13) after heating the ingot to a high temperature (1200° to 1400°C). Large ingots, of several tons weight, are often led to the mill before they have cooled below the proper temperature. The reduction is continued as the metal cools, in a rod or flat rolling mill, depending on the desired form of the final product. When the thickness is decreased to 0.2 to 0.5 inch the material has usually cooled below the recrystallization temperature. Because of the difficulty in handling hot sheets or rod of small thickness, they are rolled at or near room temperature, with intermediate annealings if necessary to soften or to develop the proper structure. In experimental work, rod is often swaged instead of rolled.

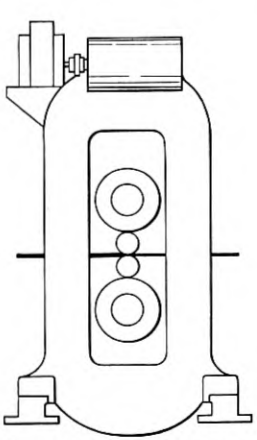
In recent years the outstanding trends in methods of fabricating materials have been toward the construction of the multiple-roll rolling mill for rolling thin strip, and the continuous strip mill for high-speed production on a large scale. Figure 14 shows the principle of construction of a typical 4-high mill ((a) and (b)), and of two special mills ((c) and (d)). In the 20-high Rohn[7] mill and 12-high Sendzimir[8] mill the two working rolls are quite small (0.2 to one inch in diameter). These are each backed by two larger rolls and these in turn by others as indicated. In the Rohn mill (c), power is supplied to the two smallest rolls and the final bearing surfaces are at the ends of the largest rolls. In the Sendzimir mill (d) the power is supplied to the rolls of intermediate size and the bearing surfaces are distributed along the whole length of the largest rolls so that no appreciable bending of the rolls occurs. The small rolls reduce the thickness of thin stock with great efficiency, and the idling rolls permit the application of high pressure. In the Steckel mill power is used to pull the sheet through the rolls, which are usually 4-high with small working rolls.

The continuous strip mill is an arrangement of individual mills such that the strip is fed continuously from one to another and may be undergoing reduction in thickness in several mills simultaneously. Figure 15 shows a mill of this kind, used for cold reduction, with 6 individual mills in tandem.

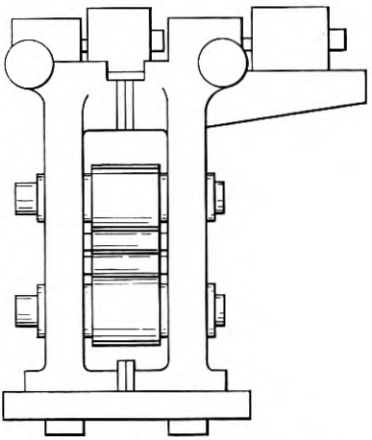
For magnetic testing numerous forms of specimens are required for various kinds of tests; these include strips for standard tests for transformer sheet, rings or parallelograms for conventional ballistic tests, "pancakes" of thin tape spirally wound for measurement by alternating current, ellipsoids for high field measurements, and many others. The various forms are

[7] W. Rohn, Heraeus Vacuumschmelze, Albertis, Hanau, 381–7 (1933).
[8] T. Sendzimir, *Iron and Steel Engr.* 23, 53–9 (1946).

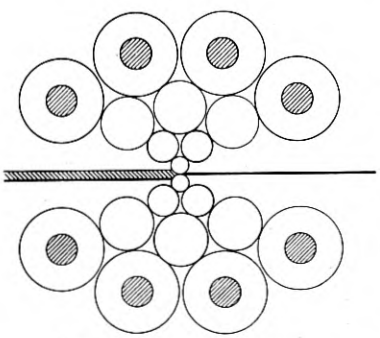
required to study or eliminate the effects of eddy-currents, demagnetizing fields and directional effects and to simulate the use of material in apparatus. Most of the needs arizing in commerce and in experimental investigation are filled by strips or sheets of thicknesses from 0.002 inch to 0.1 inch from



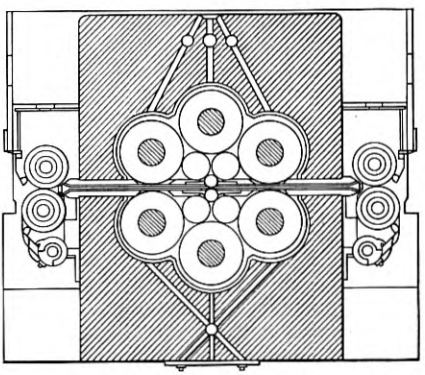
(a) 4-HIGH MILL, SIDE VIEW



(b) 4-HIGH MILL, END VIEW



(c) 20-HIGH ROHN MILL



(d) 12-HIGH SENDZIMIR MILL

Fig. 14—Arrangement of rolls in mills used for reduction of thin sheet: (a) and (b) conventional 4-high mill; (c) Rohn 20-high; (d) Sendzimir 12-high.

which coils can be wound or parts cut, by rods from which relay cores or other forms can be made, by powdered material used for pressing into cores for coils for inductive loading, and by castings for permanent magnets or other objects which may be machined or ground to final shape.

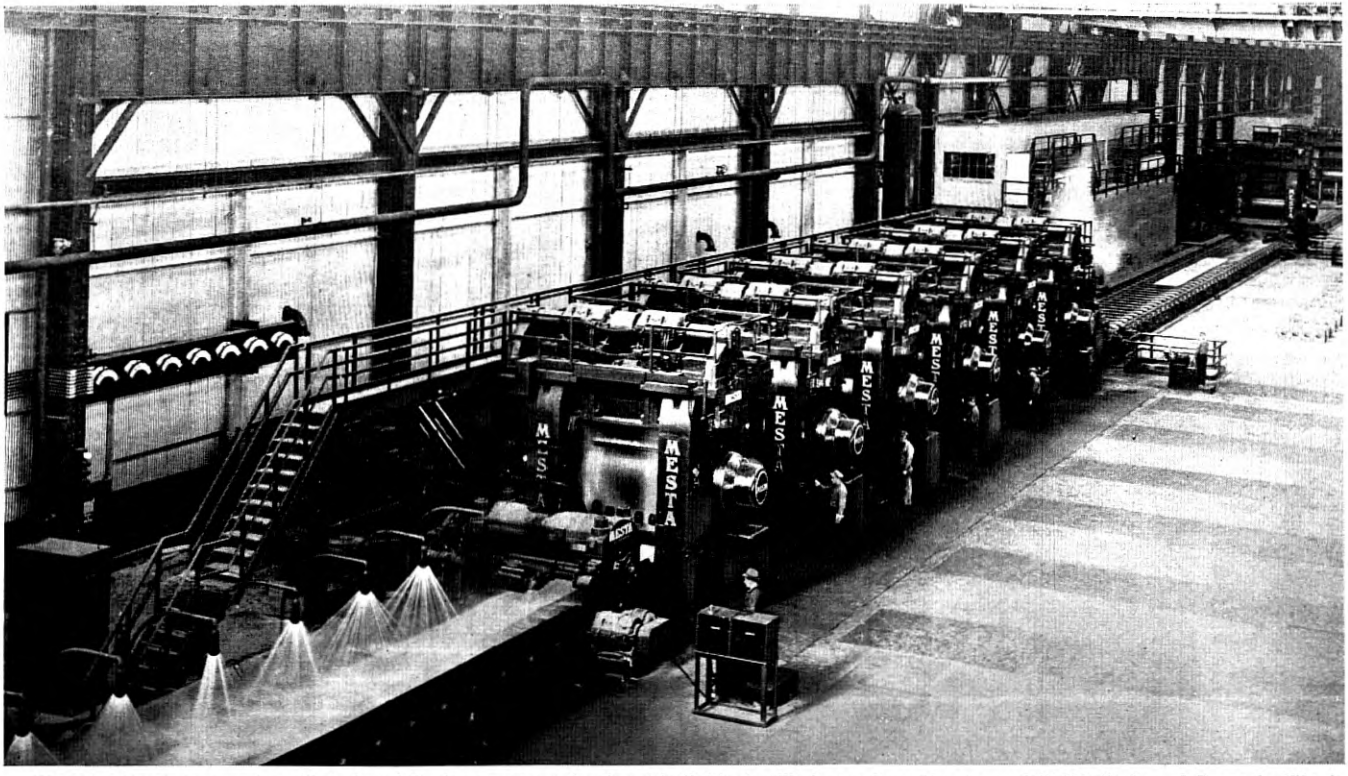


Fig. 15—Continuous strip mill designed for large output, having 6 individual mills in tandem. Courtesy of C. W. Stoker of Carnegie Illinois Steel Corp.

Heat-Treatment

High permeability materials are annealed primarily to relieve the internal strains introduced during fabrication. On the contrary permanent magnet materials are heat-treated to *introduce* strains by precipitating a second phase. Heat-treatments are decidedly characteristic of the materials and their intended uses and are best discussed in detail in connection with them.

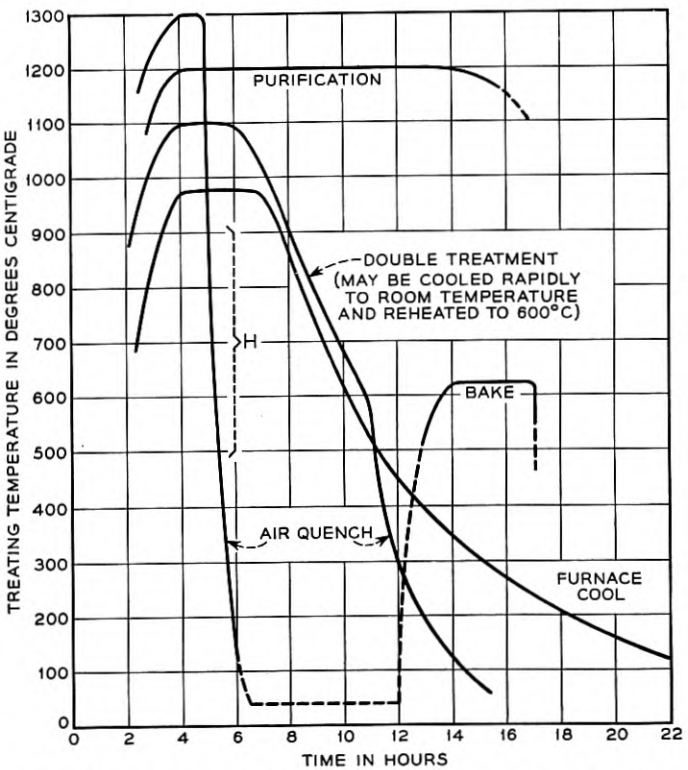


Fig. 16—Some common heat treatments for magnetic materials.

Figure 16 shows some of the commonest treatments in the form of temperature-time curves. The purpose of these various heating and cooling cycles, and typical materials subjected to them, may be listed as follows:

(1) Relief of internal strains due to fabrication or phase-changes (furnace cool). Magnetic iron.

(2) Increase of internal strains by precipitation hardening (air quench and bake). Alnico type of permanent magnets.

(3) Purification by contact with hydrogen or other gases. Silicon-iron (cold rolled), hydrogen-treated iron, Supermalloy.

There are also special treatments, such as those used for "double-treated" permalloy, "magnetically annealed" permalloy, and perminvar.

Occasionally it is necessary to homogenize a material by maintaining the

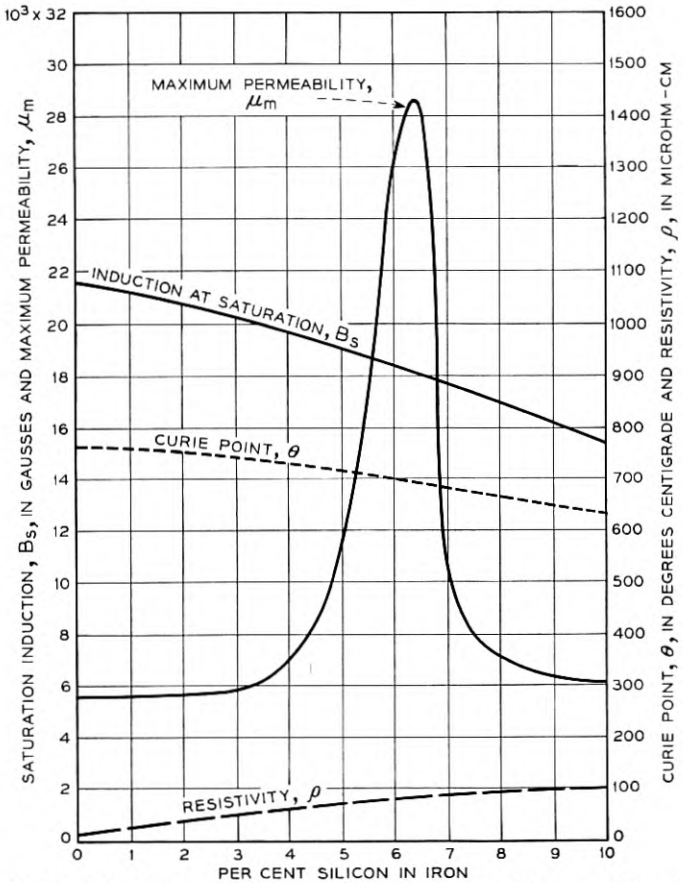


Fig. 17—Variation of some properties of iron-silicon alloys with composition: B_s, saturation intrinsic induction; θ, magnetic transformation point; ρ, electrical resistivity; μ_m, maximum permeability as determined by Miss M. Goertz.

temperature just below the freezing point for many hours. Heat-treatments also may affect grain size and crystal orientation.

Furnaces for heat-treating have various designs that will not be considered here. A modern improvement has been the use of globar (silicon carbide) heating elements that permit treatment at 1300 to 1350°C in an atmosphere of hydrogen or air.

Further discussion of "Metallurgy and Magnetism" is given in an excellent small book of this title by Stanley.[9]

Effect of Composition

Gross Chemical Composition

The effect of composition on magnetic properties will now be considered, using as examples the more important binary alloys of iron with silicon,

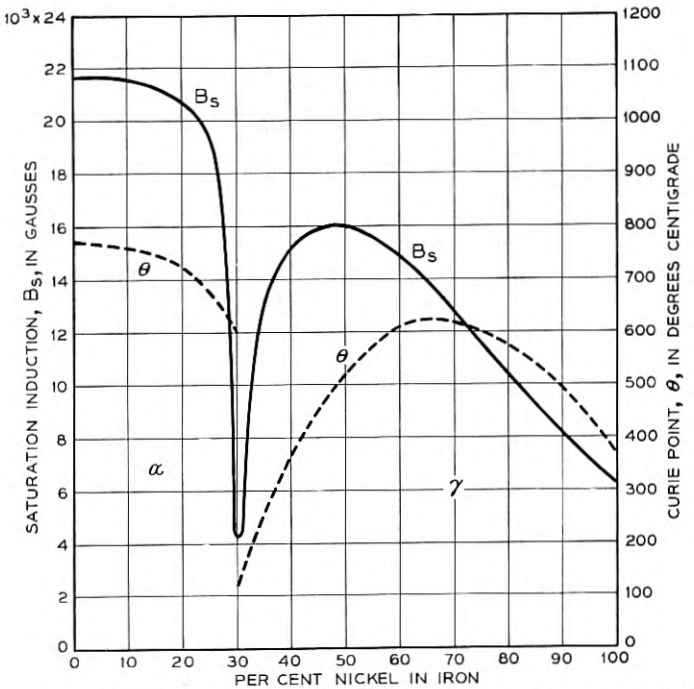


Fig. 18—Variation of B_s and θ with the composition of iron-nickel alloys.

nickel or cobalt, on which are based the most useful and interesting materials. The iron-silicon alloys are used commercially without additions, the iron-nickel and iron-cobalt alloys are most useful in the ternary form; and many special alloys, for example material for permanent magnets, contain four or five components.

Figure 17 shows four important properties of the iron-silicon alloys of low silicon content, after they have been hot rolled and annealed. The commercial alloys (3 to 5% silicon) are the most useful because they have the best

[9] J. K. Stanley, Metallurgy and Magnetism, Am. Soc. Metals, Cleveland (1949).

combination of properties of various kinds. The properties shown in the figure are important in determining the best balance: the maximum permeability, μ_m, only indirectly (it is a good measure of hysteresis loss and maximum field necessary in use), and the Curie point, θ, only in a minor role. The saturation B_s, permeability, and resistivity ρ, should all be as high as possible. B_s, θ and ρ are structure insensitive, and vary with composition in a characteristically smooth way, practically independent of heat treatment; μ_m depends on heat treatment (strain), impurities and crystal orientation. There are no phase changes to give sudden changes with composition of properties measured at room temperatures.

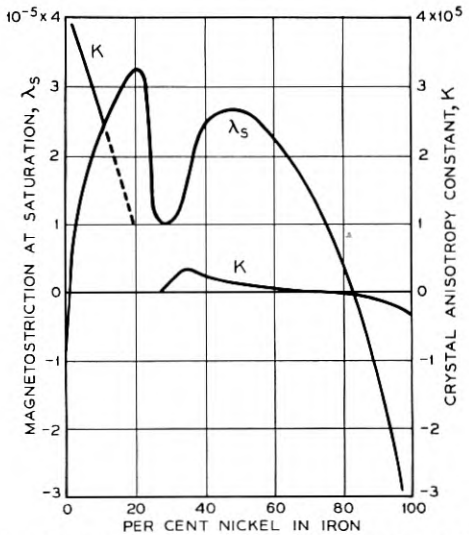


Fig. 19—Variation of saturation magnetostriction, λ_s, and crystal anisotropy, K, with the composition of iron-nickel alloys.

Some of the properties of the iron-nickel alloys are given in Figs. 18 and 19. The change in phase from α to γ at about 30 per cent nickel is responsible for the breaks at this composition. The permeabilities, μ_0 and μ_m, (Fig. 20) show characteristically the effect of heat treatment. The maxima are closely related to the points at which the saturation magnetostriction, λ_s, and crystal anisotropy, K, pass through zero (Fig. 19).

Additions of molybdenum, chromium, copper and other elements are made to enhance the desirable properties of the iron-nickel alloys.

The iron-cobalt alloys, some properties of which are shown in Fig. 21, are usually used when high inductions are advantageous. The unusual course of the saturation induction curve, with a maximum greater than that for any other material, is of obvious theoretical and practical importance. The sud-

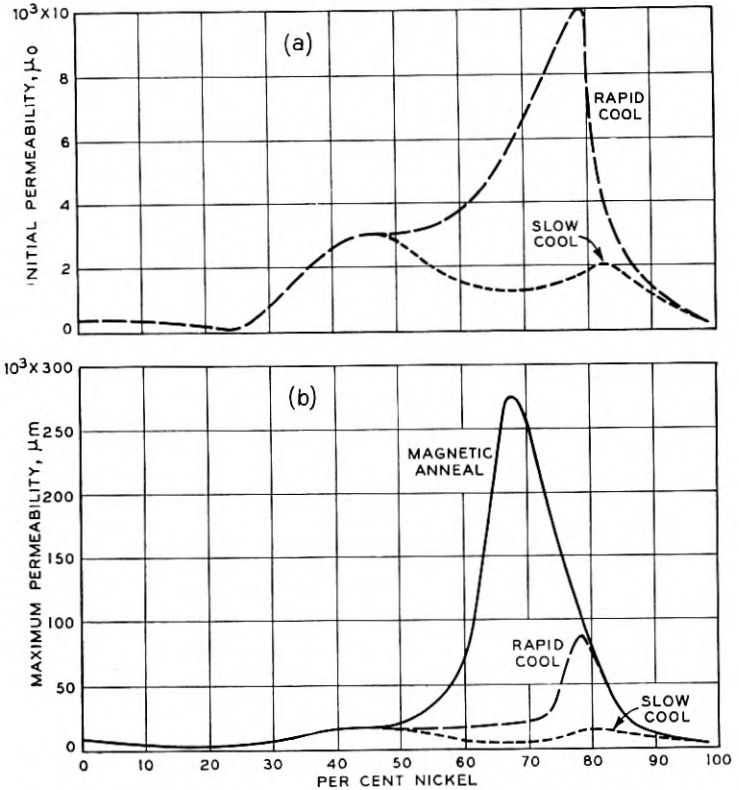


Fig. 20—Dependence of the initial and maximum permeabilities (μ_0, μ_m) of iron-nickel alloys on the heat treatment.

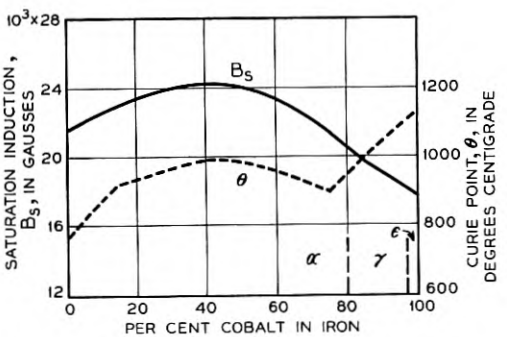


Fig. 21—Variation of B_s and θ of iron-cobalt alloys with composition.

den changes in the Curie point curve are associated with α, γ phase boundaries, as mentioned earlier in this chapter. The peak of the permeability curve (Fig. 22) occurs at the composition for which atomic ordering is stable at the highest temperature (see also Fig. 2). The sharp decline near 95 per cent cobalt coincides with the phase change γ,ϵ at this composition. Additions of vanadium, chromium and other elements are used in making commercial ternary alloys.

Some useful alloys based on the binary iron-silicon, iron-nickel and iron-cobalt alloys are described in Table III.

The *hardening* of material resulting from the precipitation of one phase in another is often used to advantage when magnetic hardness (as in permanent magnets), or mechanical hardness, is desired. To illustrate this

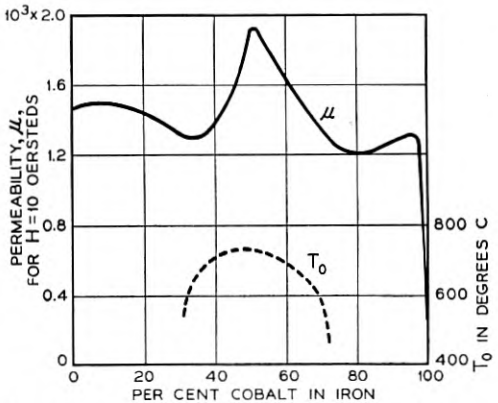


Fig. 22—Variation of permeability at $H = 10$ oersteds, and of the critical temperature of ordering, with the composition of iron-cobalt alloys.

process consider the binary iron molybdenum alloys, a partial phase diagram of which is given in Fig. 23. The effect of the boundary between the α and $\alpha + \epsilon$ fields is shown by the variation of the properties with composition (Fig. 24a). Saturation magnetization and Curie point are affected but little, the principle change in the former being a slight change in the slope of the curve at the composition at which the phase boundary crosses 500°C, the temperature below which diffusion is very slow. The Curie point curve has an almost imperceptible break at the composition at which the phase boundary lies at the Curie temperature. The changes of maximum permeability and coercive force are more drastic; μ_m drops rapidly as the amount of the second phase, ϵ, increases and produces more and more internal strain (Fig. 24b), and H_c increases at the same time. The experimental points correspond to a moderate rate of cooling of the alloy after annealing.

TABLE III

Some Properties of Some Useful Alloys Based on the Fe-Si, Fe-Co and Fe-Ni Binary Systems

Name	Composition (Per cent)	Heat Treatment	Initial Permeability, μ_0	Maximum Permeability, μ_m	Coercive Force, H_c (oersteds)	Saturation Induction, B_s (gauss)	Curie Point, θ (°C)
Hot rolled Silicon Iron	4Si, 96Fe	800°C	500*	7000	0.5	19700	690
Grain Oriented Silicon Iron	3Si, 97Fe	1200°C	1500*	40000	0.15	20000	700
Sendust	9Si, 85Fe, 5Al	Cast	30000	120000	0.05	10000	500
45 Permalloy**	45Ni, 55Fe	1200°C, H_2	3500	50000	0.07	16000	440
4-79 Permalloy	79Ni, 17Fe, 4Mo	1100°C	20000	100000	0.05	8700	420
Mumetal	75Ni, 18Fe, 2Cr, 5Cu	1175°C, H_2	20000	100000	0.05	6500	430
Supermalloy	79Ni, 16Fe, 5Mo	1300°C, H_2	100000	1000000	0.002	8000	400
Permendur	50Co, 50Fe	800°C	800	5000	2.0	24500	980
2V-Permendur	49Co, 49Fe, 2V	800°C	800	4500	2.0	24000	980
Hiperco	34Co, 64Fe, 1Cr	850°C	650	10000	1.0	24200	—

* Measured at B = 20 instead of B = 0.

** Similar alloys: Hipenik, Nicaloi, 4750 alloy, and others.

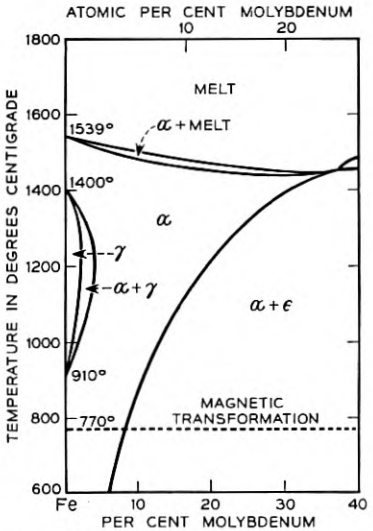


Fig. 23—Phase diagram of iron-rich iron-molybdenum alloys, showing solid solubility curve important in the precipitation-hardening process.

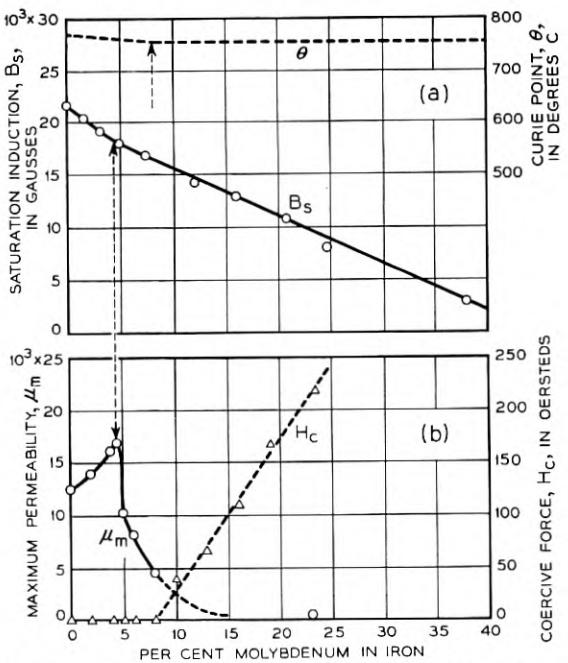


Fig. 24—Change of structure-insensitive properties (θ and B_s) and structure-sensitive properties (μ_m and H_c) with the composition when precipitation-hardening occurs.

When the amount of the second phase is considerable (as in the 15% Mo alloy) it is common practice to quench the alloy from a temperature at which it is a single phase (e.g. 1100 or 1200°C) and so maintain it temporarily as such, and then to heat it to a temperature (e.g., 600°C) at which diffusion proceeds at a more practical rate. During the latter step the second phase separates slowly enough so that it can easily be stopped at the optimum point, after a sufficient amount has been precipitated but before diffusion has been permitted to relieve the strains caused by the precipitation. A conventional heat treatment for precipitation-hardening of this kind, used on many permanent magnet materials, has already been given in Fig. 16.

In some respects the development of atomic order in a structure is like the precipitation of a second phase. When small portions of the material become ordered and neighboring regions are still disordered, severe local strains may be set up in the same way that they are during the precipitation hardening described above. The treatment used to establish high strains is the same as in the more conventional precipitation hardening. The decomposition of an ordered structure in the iron-nickel-aluminum system has been held responsible, by Bradley and Taylor,[10] for the good permanent magnet qualities of these alloys.

Some of the common permanent magnets, heat treated to develop internal strains by precipitation of a second phase, or by the development of atomic ordering, are described in Table IV.

The changes in properties to be expected when the composition varies across a phase boundary of a binary system are shown schematically by the curves of Fig. 25.

Impurities

The principle of precipitation hardening, as just described, applies also to the lowering of permeability by the presence of accidental impurities. For example, the solubilities of carbon, oxygen and nitrogen in iron, described by the curves of Fig. 26, are quite similar in form to the curve separating the α and $\alpha + \epsilon$ areas of the iron-molybdenum system of Fig. 23; the chief difference is that the scale of composition now corresponds to concentrations usually described as impurities. One expects, then, that the presence of more than 0.04 per cent of carbon in iron will cause the permeability of an annealed specimen to be considerably below that of pure iron. The amount of carbon present in solid solution will also affect the magnetic properties.

Because the amounts of material involved are small, it is difficult to carry out well defined experiments on the effects of each impurity, especially in

[10] A. J. Bradley and A. Taylor, *Proc. Roy. Soc.* (London) 166, 353–75 (1938).

TABLE IV

Some Useful Permanent Magnets and Their Properties

Name	Composition (Per cent)*	Fabrication	Heat Treatment	H_c	B_r	Mechanical Properties
Carbon Steel	1Mn, 0.9C	HR, PM	Q800	50	10000	H, S
Tungsten Steel	5W, 0.3Mn, 0.7C	HR, PM	Q850	70	10300	H, S
Chromium Steel	3.5Cr, 0.3Mn, 0.9C	HR, PM	Q830	65	9700	H, S
Cobalt Steel	36Co, 4Cr, 5W, 0.7C	HR, PM	Q950	240	9500	H, S
Remalloy (Comol)	17Mo, 12Co	HR, PM	Q1200, B700	250	10500	H
Alnico 2	12Co, 17Ni, 10Al, 6Cu	C, G	A1200, B600	550	7200	H, B
Alnico 5	24Co, 14Ni, 8Al, 3Cu	C, G	A1300, **B600	550	12500	H, B
Alnico 12	35Co, 18Ni, 6Al, 8Ti	C, G	Cast, B650	950	5800	H, B
Alcomax	25Co, 11Ni, 8Al, 6Cu	C, G	A1300, **B600	550	12500	H, B
Vicalloy	52Co, 10V	C, Cr, PM	B600	300	8800	D
Cunife	20Ni, 60Cu	C, Cr, PM	B600	550	5400	D
Platinum-Cobalt	77Pt, 23Co	C, Cr, PM	Q1200, B650	2600	4500	D
Silmanal	87Ag, 9Mn, 4Al	C, Cr, PM		6000†	550	D

* Remainder iron
Q—quenched from indicated centigrade temperature in oil
A—cooled in air from indicated temperature
B—baked at indicated temperature
HR—hot rolled
CR—cold rolled
PM—punched or machined
C—cast
G—ground
** Cooled in magnetic field
† Coercive force for I = 0
H—hard
B—brittle
D—ductile or malleable
S—strong

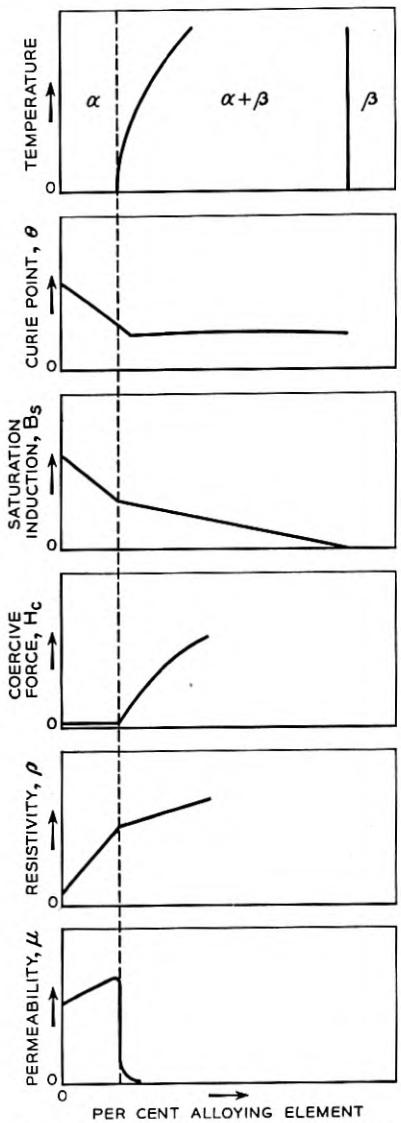


Fig. 25—Diagrams illustrating the changes in various properties that occur when a second phase precipitates.

the absence of disturbing amounts of other impurities. Two examples of the effect of impurities will be given, in addition to Fig. 8. In Fig. 27 Yensen and Ziegler[11] have plotted the hysteresis loss as dependent on carbon content,

[11] T. D. Yensen and N. A. Ziegler, *Trans. Am. Soc. Metals* 24, 337–58 (1936).

the curve giving the mean values of many determinations. The hysteresis decreases rapidly at small carbon contents, when these are of the order of magnitude of the solid solubility at room temperature.

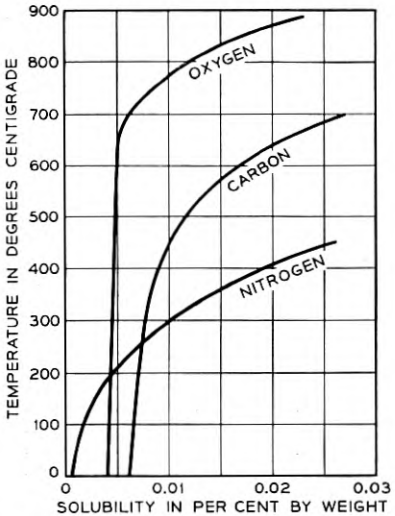


Fig. 26—Approximate solubility curves of carbon, oxygen and nitrogen in iron.

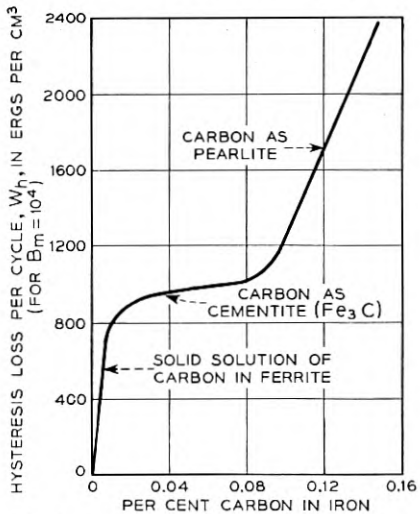


Fig. 27—Effect of carbon content on hysteresis in iron. *Yensen and Ziegler.*

Cioffi[12] has purified iron from carbon, oxygen, nitrogen and sulfur by heating in pure hydrogen at 1475°C, and has measured the permeability

[12] P. P. Cioffi, *Phys. Rev. 39*, 363–7 (1932).

at different stages of purification. Table V shows that impurities of a few thousandths of a per cent are quite effective in depressing the maximum permeability of iron.

Carbon and nitrogen, present as impurities, are known to cause "aging" in iron—that is, the permeability and coercive force of iron containing these elements as impurities will change gradually with time when maintained somewhat above room temperature. As an example, a specimen of iron was maintained for 100 hours first at 100°C, then 150°C, then 100°C, and so on.

TABLE V

Maximum permeability of Armco iron with different degrees of purification, effected by heat treatment in pure hydrogen at 1475°C for the times indicated (P. P. Cioffi).
Analyses from R. F. Mehl (private communication to P. P. Cioffi).

Time of Treatment in Hours	μ_m	Composition in Per Cent					
		C	S	O	N	Mn	P
0	7000	0.012	0.018	0.030	0.0018	0.030	0.004
1	16000	.005	.010	.003	.0004	—	—
3	30000	.005	.006	.003	.0003	—	—
7	70000	.003	—	.003	.0001	—	—
18	227000	.005	<.003	.003	.0001	.028	.004
Precision of analysis.....		.001	.002	.002	.0001		

The corresponding changes in coercive force are given in the diagram of Fig. 28. A change of about 2-fold is observed.

SOME IMPORTANT PHYSICAL PROPERTIES

There are many physical characteristics that are important in the study of ferromagnetism from both the practical and the theoretical point of view. These include the resistivity, density, atomic diameter, specific heat, expansion, hardness, elastic limit, plasticity, toughness, mechanical damping, specimen dimensions, and numerous others. In a different category may be mentioned corrosion, homogeneity and porosity. Most of these properties are best discussed in connection with specific materials or properties; only the most important characteristics will be mentioned here. A table of the atomic weights and numbers, densities, melting points, resistivities and coefficients of thermal expansion of the metallic elements, is readily available in the Metals Handbook.

Dissolving a small amount of one element in another increases the *resistivity* of the latter. To show the relative effects of various elements, the common binary alloys of iron and of nickel are shown in Figs. 29 and 30. From a theoretical standpoint it is desirable to understand (1) the relatively

high resistivity of the ferromagnetic elements compared to their neighbors in the periodic table and (2) the relative amounts by which the resistivity of iron (or cobalt or nickel) is raised by a given atomic percentage of various other elements. From a practical standpoint, a high resistivity is usually

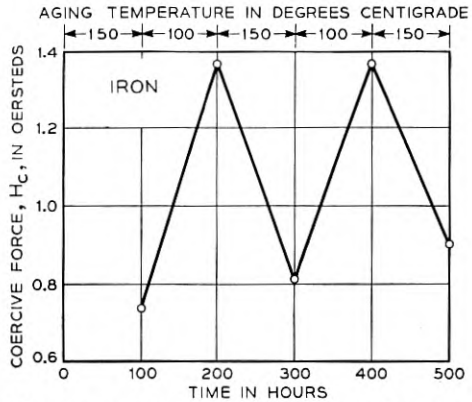


Fig. 28—Effect of nitrogen impurity on the coercive force of iron annealed successively at 100 and 150°C.

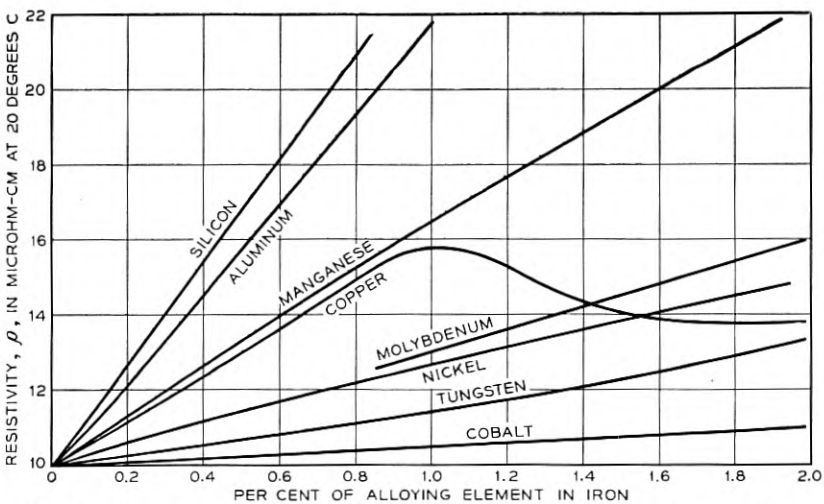


Fig. 29—Dependence of resistivity on the addition of small amounts of various elements to iron.

desirable in order to decrease the eddy-current losses in the material, and so decrease the power wasted and the lag in time between the cause and effect, for example, the time lag of operation of a relay.

Knowledge of the *atomic diameter* is important in considering the effects

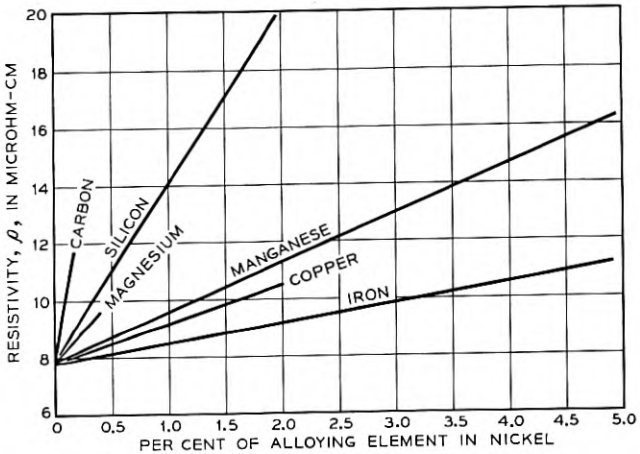


Fig. 30—Resistivity of various alloys of nickel.

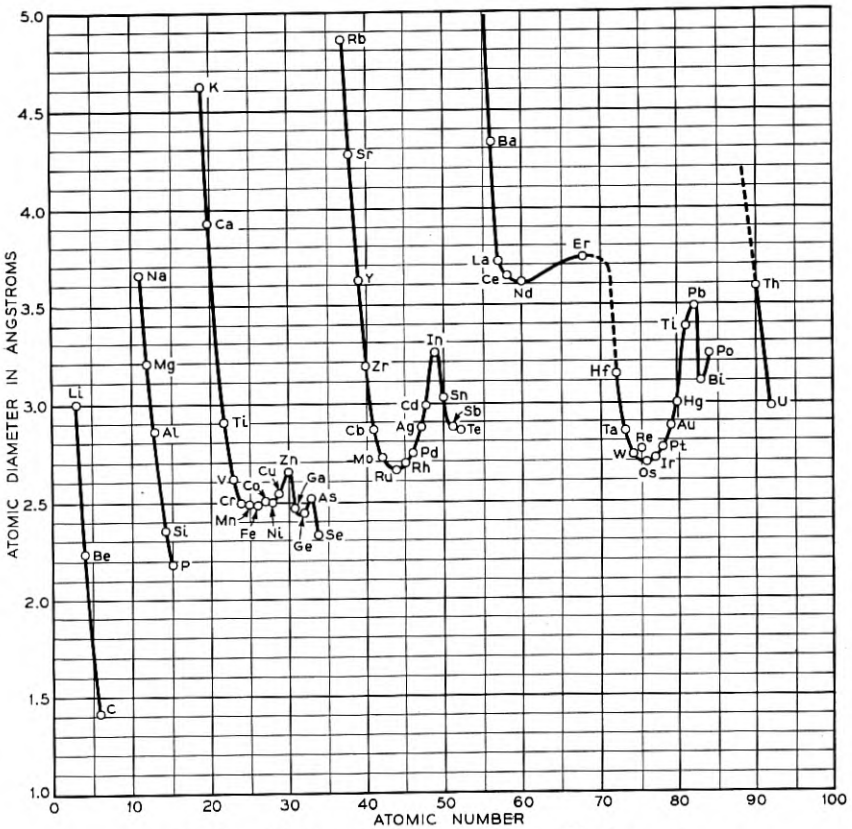


Fig. 31—Atomic diameter of various metallic elements.

of alloying elements, and values for the metallic and borderline elements are shown in Fig. 31. Most of the values are simply the distances of nearest approach of atoms in the element as it exists in the structure stable at room temperature. Atomic diameter is especially important in theory because the very existence of ferromagnetism is dependent in a critical way on the distance between adjacent atoms. This has been discussed more fully in a previous paper.[13]

Even when no phase change occurs in a metal, important *changes in structure* occur during fabrication and heat treatment, and these are complicated and imperfectly understood. When a single crystal is elongated by tension, slip occurs on a limited number of crystal planes that in general are inclined to the axis of tension. As elongation proceeds, the planes on which slip is taking place tend to turn so that they are less inclined to the axis. In this way a definite crystallographic direction approaches parallelism

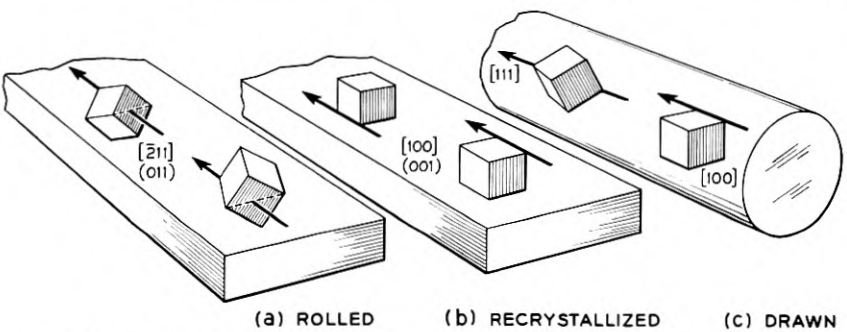


Fig. 32—The preferred orientations of crystals in nickel sheet and wire after fabrication and after recrystallization.

with the length of the specimen. In a similar but more complicated way, any of the usual methods of fabrication cause the many crystals of which it is composed to assume a non-random distribution of orientations, often referred to as preferred or *special orientations*, or *textures*. Some of the textures reported for cold rolled and cold drawn magnetic materials are given in Table VI, taken from the compilation by Barrett.[14] The orientations of the cubes which are the crystallographic units are shown in Fig. 32 (a) and (c) for cold rolled sheets and cold drawn wires of nickel.

Since the magnetic properties of single crystals depend on crystallographic direction (anisotropy), the properties of polycrystalline materials in which there is special orientation will also be direction-dependent. In fact it is difficult to achieve isotropy in any fabricated material, even if fabrication involves no more than solidifying from the melt. The relief of the internal

[13] R. M. Bozorth, *Bell Sys. Tech. Jl.* 19, 1–39 (1940).
[14] C. S. Barrett, Structure of Metals, McGraw Hill, New York (1943).

strains in a fabricated metal by annealing proceeds only slowly at low temperatures (up to 600°C for most ferrous metals) without noticeable grain growth or change in grain orientation, and is designated *recovery*. The principle change is a reduction in the amplitude of internal strains, and this can be followed quantitatively by X-ray measurements. Near the point of complete relief distinct changes occur in both grain size and grain orientation, and the material is said *to recrystallize*. At higher temperatures grain growth increases more rapidly. The specific temperatures necessary for both recovery and recrystallization depend on the amount of previous deformation, as shown in Fig. 33. Special orientations are also present in fabricated materials after recrystallization, and some of these are listed in Table VI and illustrated for nickel in Fig. 32 (b).

As an example of the dependence of various magnetic properties on direction, Fig. 34 gives data of Dahl and Pawlek[15] for a 40 per cent nickel iron

TABLE VI

Preferred Orientations in Drawn Wires and Rolled Sheets, Before and After Recrystallization, and in Castings (Barrett[14])

The rolling plane and rolling direction, or wire axis, or direction of growth, are designated

Metal	Crystal Structure	Drawn wires		Rolled Sheets		As Cast
		As Drawn	Recrystallized	As Rolled	Recrystallized	
Iron.....	BCC	[110]	[110]	(001), [110] and others	(001), 15° to [110]	[100]
Cobalt.....	HCP	—	—	(001)	—	—
Nickel.....	FCC	[111] and [100]	—	(110), [112] and others	(100), [001]	

alloy reduced 98.5 per cent in area by cold rolling and then annealed at 1100°C. After further cold rolling (50 per cent reduction) the properties are as described in Fig. 35.

The mechanical properties ordinarily desirable in practical materials are those which facilitate fabrication. Mild steel is often considered as the nearest approach to an ideal material in this respect. Silicon iron is limited by its brittleness, which becomes of major importance at about 5 per cent silicon; this is shown by the curve of Fig. 36. Permalloy is "tougher" than iron or mild steel and requires more power in rolling and more frequent annealing between passes when cold-rolled, but can be cold-worked to smaller dimensions. If materials have insufficient stiffness or hardness, parts of apparatus made from them must be handled with care to avoid bending and consequent lowering of the permeability. If the hardness is too great the material must be ground to size. This is the case with some permanent magnets.

[15] O. Dahl and F. Pawlek, *Zeits. f. Metallkunde* 28, 230–3 (1936).

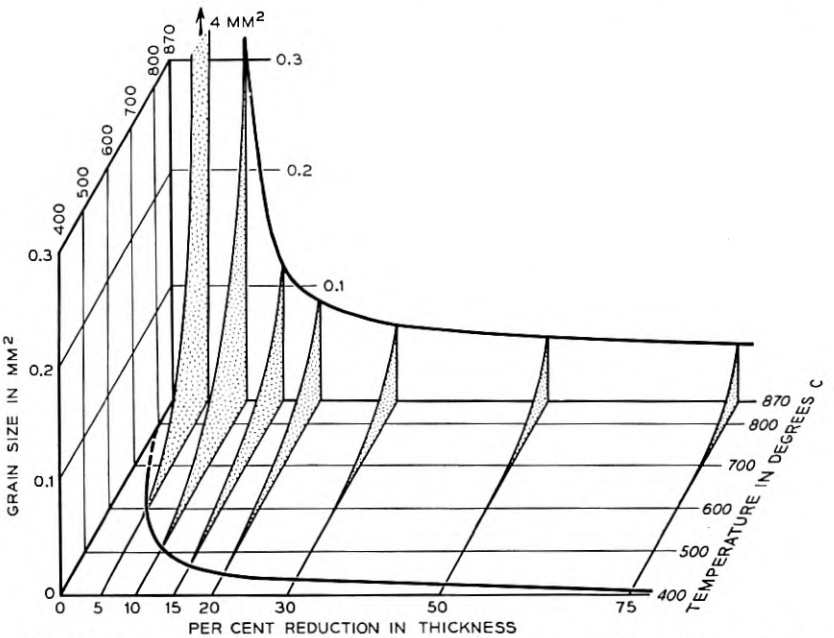


Fig. 33—Dependence of the grain size of iron on the amount of deformation and on the temperature of anneal. *Kenyon.*

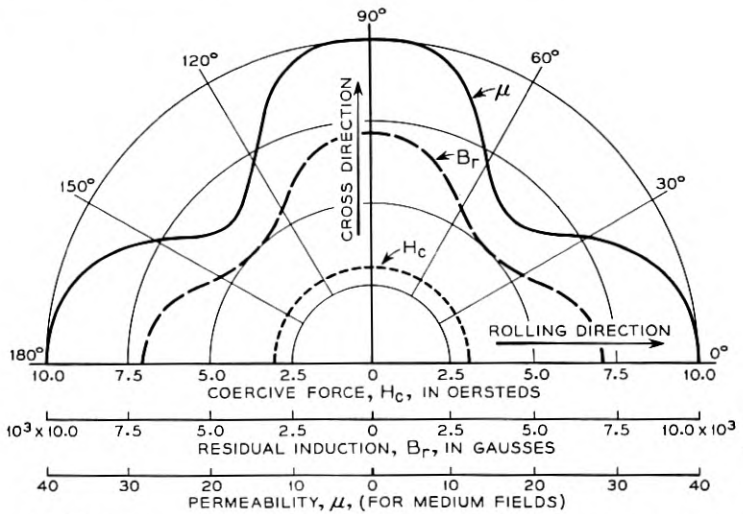


Fig. 34—Variation of magnetic properties with the direction of measurement in a sheet of iron-nickel alloy (40% Ni) severely rolled (98.5%) and annealed at 1100°C.

The effect of size of a magnetic specimen is often of importance. This is well known in the study of *thin films*, and *fine powders* in which the smallest

dimension is about 10^{-4} cm or less. Many studies have been made of thin electrodeposited and evaporated films. Generally it is found that the permeability is low and the coercive force high. The interpretation is uncertain

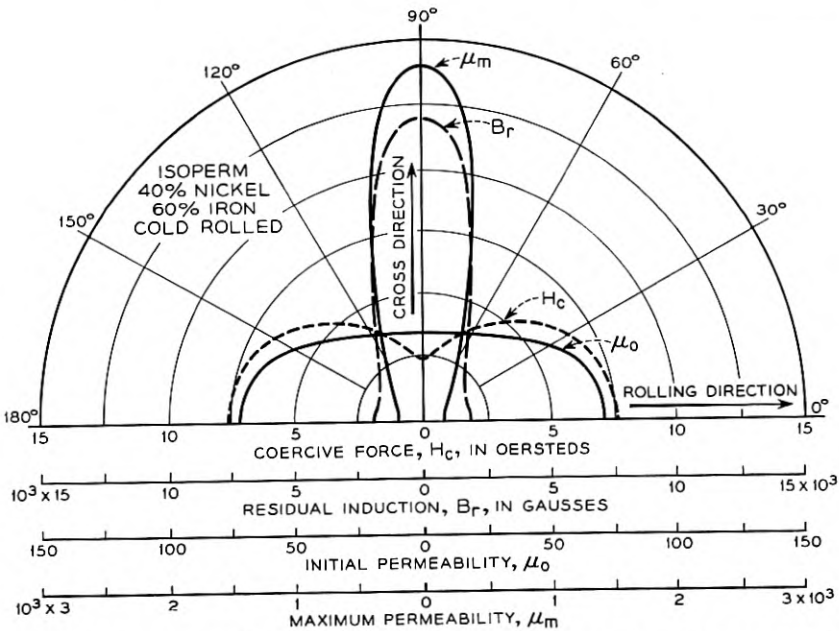


Fig. 35—Properties of the same material as that of Fig. 34, after it has been rolled, annealed, and again rolled.

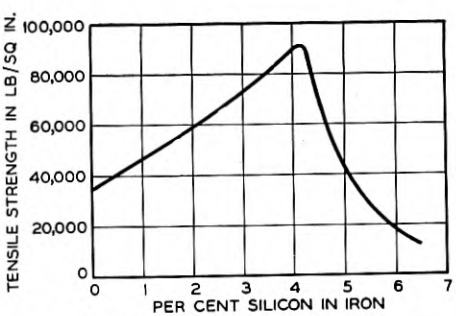


Fig. 36—Variation of the breaking strength of iron-silicon alloys, showing the onset of brittleness near 4 per cent silicon.

because it is difficult to separate the effects of strains and air gaps from the intrinsic effect of thickness, though it is known that each one of these variables has a definite effect. As one example of the many experiments, we

will show here the effect of the thickness of electrodeposited films of cobalt. Magnetization curves are shown in Fig. 37 according to previously unpublished work of the author.

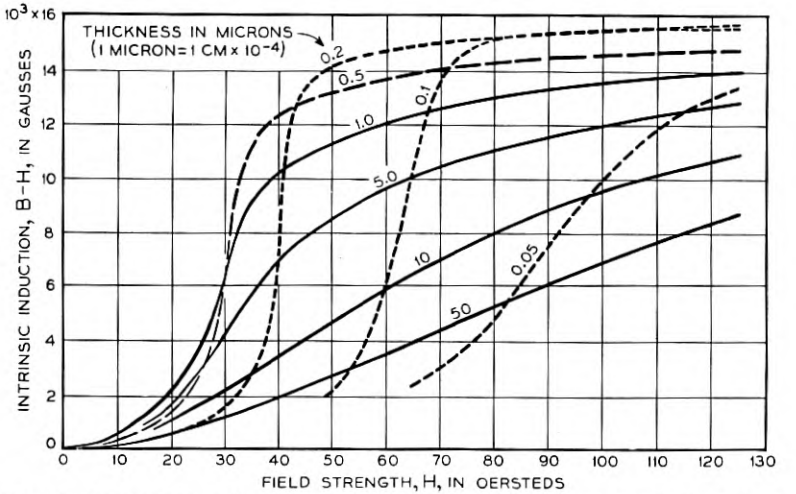


Fig. 37—Dependence of the magnetization curves of pure electrodeposited cobalt films on the thickness.

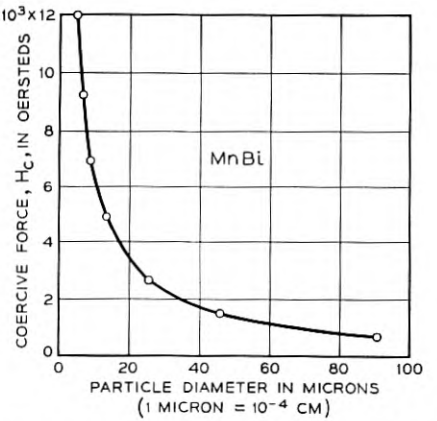


Fig. 38—Dependence of coercive force on the particle size of MnBi powder. *Guillaud.*

The high coercive force obtained in fine powders by Guillaud[16] is one of the most clear cut examples of the intrinsic effect of particle size. The coercive force increases by a factor of 15 as the size decreases to 5×10^{-4} cm (Fig. 38).

[16] C. Guillaud, Thesis, Strasbourg (1943).

Properties Affected by Magnetization

In addition to the magnetization, other properties are changed by the direct application of a magnetic field. Some of these, and the amounts by which they may be changed, are as follows:

Length and volume (magnetostriction) (0.01%)
Electrical resistivity (5%)
Temperature (magnetocaloric effect; heat of hysteresis) (1°C)
Elastic constants (20 per cent)
Rotation of plane of polarization of light (Kerr and Faraday effects) (one degree of arc)

In addition to these properties there are others that change with temperature because the magnetization itself changes. Thus there is "anomalous" temperature-dependence of:

Specific heat
Thermal expansion
Electrical resistivity
Elastic constants
Thermoelectric force

and of other properties below the Curie point of a ferromagnetic material, even when no magnetic field is applied.

Also associated with ferromagnetism are galvanomagnetic, chemical and other effects.

Technical Articles by Bell System Authors Not Appearing in the Bell System Technical Journal

Measurement Method for Picture Tubes. M. W. BALDWIN.[1] *Electronics*, V. 22, pp. 104–105, Nov., 1949.

Diffusion in Binary Alloys.† J. BARDEEN.[1] *Phys. Rev.*, V. 76, pp. 1403–1405, Nov. 1, 1949.

ABSTRACT—Darken has given a phenomenological theory of diffusion in binary alloys based on the assumption that each constituent diffuses independently relative to a fixed reference frame. It is shown that diffusion via vacant lattice sites leads to Darken's equations if it is assumed that the concentration of vacant sites is in thermal equilibrium. Grain boundaries and dislocations may act as sources and sinks for vacant sites and act to maintain equilibrium. The modifications required in the equations if the vacant sites are not in equilibrium are discussed.

Variable Phase-Shift Frequency-Modulated Oscillator. O. E. DE LANGE.[1] *I.R.E., Proc.*, V. 37, pp. 1328–1331, Nov., 1949.

ABSTRACT—The theory of operation of a phase-shift type of oscillator is discussed briefly. This oscillator consists of a broad-band amplifier, the output of which is fed back to the input through an electronic phase-shifting circuit. The instantaneous frequency is controlled by the phase shift through this latter circuit. True FM is obtained in that frequency deviation is directly proportional to the instantaneous amplitude of the modulating signal and substantially independent of modulation frequency.

A practical oscillator using this circuit at 65 mc is described.

Erosion of Electrical Contacts on Make.† L. H. GERMER[1] and F. E. HAWORTH.[1] *Jl. Applied Phys.*, V. 20, pp. 1085–1108, Nov., 1949.

ABSTRACT—When an electric current is established by bringing two electrodes together, they necessarily discharge a capacity. Unless the current which is set up is above 1 ampere, the erosion which is produced in a low voltage circuit is appreciable only when the capacity is of appreciable size and when it is discharged very rapidly by an arc. When the arc occurs, its energy is dissipated almost entirely upon the positive electrode and, when the circuit inductance is sufficiently low, melts out a crater intermediate in volume between the volume of metal which can be melted by the energy

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

and that which can be boiled. Some of the melted metal lands on the negative electrode and, with repetition of the phenomenon, results in a mound of metal transferred from the anode to the cathode. This transfer, which is about 4×10^{-14} cc of metal per erg, is the erosion which occurs on the make of electrical contacts.

The arc voltage is of the order of 15. If the initial circuit potential is more than about 50 volts, there may be more than one arc discharge, successive discharges being in opposite directions and resulting in the transfer of metal in opposite directions—always to the electrode which is negative.

The occurrence of an arc is dependent upon the condition of the electrode surfaces and upon the circuit inductance. For "inactive" surfaces an arc does not occur for inductances greater than about 3 microhenries. Platinum surfaces can be "activated" by various organic vapors, and in the active condition they give arcs even when the circuit inductance is greater than this limiting value by a factor of 10^3.

The Conductivity of Silicon and Germanium as Affected by Chemically Introduced Impurities. G. L. PEARSON.[1] Paper presented at A. I. E. E., Swampscott, Mass., June 20–24, 1949. Included in compilation on semiconductors. *Elec. Engg.*, V. 68, pp. 1047–1056, Dec. 1949.

ABSTRACT—Silicon and germanium are semiconductors whose electrical properties are highly dependent upon the amount of impurities present. For example, the intrinsic conductivity of pure silicon at room temperature is 4×10^{-6} (ohm cm)$^{-1}$ and the addition of one boron atom for each million silicon atoms increases this to 0.8 (ohm cm)$^{-1}$, a factor of 2×10^5.

Although such impurity concentrations are too weak to be detected by standard chemical analysis, the use of radioactive tracers and the Hall effect has made it possible to make quantitative measurements at impurity concentrations as small as one part in 5×10^8.

Silicon and germanium are elements of the fourth group of the periodic table with the same crystal structure as diamonds and they have respectively 5.2×10^{22} and 4.5×10^{22} atoms per cubic centimeter. The addition of impurity elements of the third group such as boron or aluminum gives defect or p-type conductivity. Elements from the fifth group such as phosphorous, antimony or arsenic give excess or n-type conductivity.

The conductivity at room temperature, where it has been shown that each impurity atom contributes one conduction charge, is given by equation (1) where N is the number of solute atoms per cubic centimeter.

$$\sigma = A + BN. \tag{1}$$

[1] B.T.L.

The constants A and B for the various alloys investigated are given in the following table:

Alloy	A	B
Si + B	4×10^{-6}	1.6×10^{-17}
Si + P	4×10^{-6}	4.8×10^{-17}
Ge + Sb	1.7×10^{-2}	4.2×10^{-16}

Equation (1) applies to solute atom concentrations as high as 5×10^{19} per cc. At higher concentrations the mobilities are lowered due to increased impurity scattering so that the computed conduction is higher than the measured.

Microstructures of Silicon Ingots.† W. G. PFANN[1] and J. H. SCAFF.[1] *Metals Trans.*, V. 185 (*Jl. Metals*, V. 1) pp. 389–392, June, 1949.

Increasing Space-Charge Waves.† J. R. PIERCE.[1] *Jl. Applied Phys.*, V. 20, pp. 1060–1066, Nov. 1949.

ABSTRACT—An earlier paper presented equations for increasing waves in the presence of two streams of charged particles having different velocities, and solved the equations assuming the velocity of one group of particles to be zero or small. Numerical solutions giving the rate of increase and the phase velocity of the increasing wave for a wide range of parameters, covering cases of ion oscillation and double-stream amplification, are presented here.

Traveling-Wave Oscilloscope. J. R. PIERCE.[1] *Electronics*, V. 22, pp. 97–99, Nov., 1949.

ABSTRACT—This paper describes a 1,000 volt oscilloscope tube with a traveling-wave deflecting system. The tube is suitable for viewing periodic signals with frequencies up to 500 mc. A signal of 0.037 volt into 75 ohms deflects the spot one spot diameter. A few milliwatts input gives a good pattern, so that the tube can be used without an amplifier. The pattern is viewed through a sixty power microscope.

P-type and N-type Silicon and the Formation of Photovoltaic Barrier in Silicon Ingots.† J. H. SCAFF,[1] H. C. THEURERER[1] and E. E. SCHUMACHER.[1] *Metals Trans.*, V. 185 (*Jl. Metals*, V. 1) pp. 383–388, Jan., 1949.

Longitudinal Noise in Audio Circuits. H. W. AUGUSTADT[1] and W. F. KANNENBERG.[1] *Audio Engg.*, V. 34, pp. 22–24, 45, Jan., 1950.

Transistors. J. A. BECKER.[1] Compilation of three papers presented at A. I. E. E. meeting Swampscott, Mass., June 20–24, 1949. *Elec. Engg.*, V. 69, pp. 58–64, Jan., 1950.

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Application of Thermistors to Control Networks.† J. H. BOLLMAN[1] and J. G. KREER.[1] *I. R. E., Proc.*, V. 38, pp. 20–26, Jan., 1950.

ABSTRACT—In connection with the application of thermistors to regulating and indicating systems, there have been derived several relations between current, voltage, resistance, and power which determine the electrical behavior of the thermistor from its various thermal and physical constants. The complete differential equation describing the time behavior of a directly heated thermistor has been developed in a form which may be solved by methods appropriate to the problem.

Sensitive Magnetometer for Very Small Areas.† D. M. CHAPIN.[1] *Rev. Sci. Instruments*, V. 20, pp. 945–946, Dec., 1949.

ABSTRACT—A vibrating wire system for measuring weak magnetic fields is described for use in very small spaces. Quartz crystals are used for drivers to get sufficient velocity with very small displacements. To adjust the driving voltage to correspond exactly to the natural crystal frequency, the crystal is also used to regulate the oscillator.

Method of Calculating Hearing Loss for Speech from an Audiogram.† H. FLETCHER.[1] *Acoustical Soc. Am., Jl.*, V. 22, pp. 1–5, Jan., 1950.

ABSTRACT—The question frequently arises, Can one compute the hearing loss of speech from the audiogram and thus make it unnecessary to make a speech test after the hearing loss for several frequencies has been recorded. This paper shows that this can be done by taking a weighted average of the exponentials of the hearing loss at each frequency. Or if β_s is the hearing loss for speech and β_f the hearing loss at each frequency,

$$10^{(\beta_s/10)} = \int G\, 10^{(\beta_f/10)} df$$

The weighting factor G was determined by Fletcher and Galt from threshold measurements of speech coming from filter systems. As specifically applied to the case of hearing loss at the five frequencies 250, 500, 1000, 2000 and 4000 cps, the above equation is approximately equivalent to

$$\beta_s = -10 \log [.01 \times 10^{-(\beta_1/10)} + .13 \times 10^{-(\beta_2/10)} + .40 \times 10^{-(\beta_3/10)} + .38 \times 10^{-(\beta_4/10)} + .08 \times 10^{-(\beta_5/10)}]$$

where β_1 is hearing loss at 250 cps
β_2 is hearing loss at 500 cps
β_3 is hearing loss at 1000 cps
β_4 is hearing loss at 2000 cps
β_5 is hearing loss at 4000 cps

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Designing for Air Purity. A. M. HANFMANN.[2] *Heating & Ventilating,* V. 47, pp. 59–64, Jan., 1950.

Reciprocity Pressure Response Formula Which Includes the Effect of the Chamber Load on the Motion of the Transducer Diaphragms.† M. S. HAWLEY.[1] *Acoustical Soc. Am., Jl.,* V. 22, pp. 56–58, Jan., 1950.

ABSTRACT—In order to reduce the effects of wave motion in the coupling chamber to permit reciprocity pressure response measurements to higher frequencies, only two of the three transducers involved are coupled at a time to the chamber. Given for these conditions is a derivation of the pressure response formula which includes the effect of the chamber load on the motion of the transducer diaphragms.

Theory of the "Forbidden" (222) Electron Reflection in the Diamond Structure.† R. D. HEIDENREICH.[1] *Phys. Rev.,* V. 77, pp. 271–283, Jan. 15, 1950.

ABSTRACT—The dynamical or wave mechanical theory of electron diffraction is extended to include several diffracted beams. In the Brillouin zone scheme this is equivalent to terminating the incident crystal wave vector at or near a zone edge or corner. The problem is then one of determining the energy levels and wave functions in the neighborhood of a corner. The solution of the Schrödinger equation near a zone corner is a linear combination of Bloch functions in which the wave vectors are determined by the boundary conditions and the requirement that the total energy be fixed. This leads to a multiplicity of wave vectors for each diffracted beam giving rise to interference phenomena and is an essential feature of the dynamical theory.

At a Brillouin zone edge formed by boundaries associated with reciprocal lattice points S and O the orthogonality of the unperturbed wave functions in conjunction with the periodic potential requires that another reciprocal lattice point λ be included in the calculation. The indices of λ must be such that $(\lambda_1\lambda_2\lambda_3) = (s_1s_2s_3) - (g_1g_2g_3)$. The perturbation at the zone edge results in non-zero amplitude coefficients Cg, Cs and Cj for the diffracted waves irrespective of whether or not the structure factor for λ , s or g vanishes. This is the basis of the explanation of the (222) reflection and since it arises through perturbation at a Brillouin zone edge or corner the term "perturbation reflection" is advanced to replace the commonly used "forbidden reflection."

The octahedron formed by the (222) Brillouin zone boundaries exhibits an array of lines due to intersections with other boundaries to form edges. This array of lines is called a "perturbation grid" and the condition for the occurrence of a (222) reflection is simply that the incident wave vector terminate on or near a grid line. Numerical intensity calculations are pre-

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[2] W. E. Co.

sented which show that a strong (222) can be accounted for by the dynamical theory.

An impedance network model is briefly discussed which may aid in qualitative considerations of the dynamical theory for the case of several diffracted waves.

Determination of g-Values in Paramagnetic Organic Compounds by Microwave Resonance. A. N. HOLDEN,[1] C. KITTEL,[1] F. R. MERRITT[1] and W. A. YAGER.[1] Letter to the Editor, *Phys. Rev.*, V. 77, pp. 146–147, Jan. 1, 1950.

Nonlinear Coil Generators of Short Pulses.† L. W. HUSSEY.[1] *I.R.E., Proc.*, V. 38, pp. 40–44, Jan., 1950.

ABSTRACT—Small permalloy coils and circuits have been developed which produce pulses well below a tenth of a microsecond in duration with repetition rates up to a few megacycles.

The construction of these coils is described. Low power circuits are dicussed suitable for different types of drive and different frequency ranges.

Subjective Effects in Binaural Hearing. W. KOENIG.[1] Letter to the Editor, *Acoustical Soc. Am., Jl.*, V. 22, pp. 61–62, Jan., 1950.

ABSTRACT—Experiments with a binaural telephone system disclosed some remarkable properties, notably its ability to "squelch" reverberation and background noises, as compared to a system having only one pickup. No explanation has been found for this subjective effect. It was also discovered that a well-known defect in the directional discrimination of binaural systems was remedied by a mechanical arrangement which rotated the pickup microphones as the listener turned his head.

Corrosion Testing of Buried Cables. T. J. MAITLAND.[3] *Corrosion*, V. 6, pp. 1–8, Jan., 1950.

40AC1 Carrier Telegraph System. A. L. MATTE.[1] *Tel. & Tel. Age*, No. 2, pp. 7–9, Feb., 1950.

Giving New Life to Old Equipment. P. H. MIELE.[3] *Bell Tel. Mag.*, V. 28, pp. 154–163, Autumn, 1949.

Thermionic Emission of Thin Films of Alkaline Earth Oxide Deposited by Evaporation.† G. E. MOORE[1] and H. W. ALLISON.[1] *Phys. Rev.*, V. 77, pp. 246–257, Jan. 15, 1950.

ABSTRACT—Monomolecular films of BaO or SrO were deposited by evaporation on clean tungsten or molybdenum surfaces with precautions to eliminate effects caused by excess metal of the oxide or by heating. Thermionic emissions of the same order of magnitude as from commercial oxide cathodes have been obtained from these systems. The results can be explained qualitatively by considering the adsorbed molecules as oriented dipoles. Although

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

the results may suggest a possible mechanism for a portion of the emission from thick oxide cathodes, there exist serious obstacles to such thin film phenomena as a complete explanation.

Long Distance Finds the Way. W. H. NUNN.[3] *Bell Tel. Mag.*, V. 28, pp. 137–147, Autumn, 1949.

Private Line Services for the Aviation Industry. H. V. ROUMFORT.[3] *Bell Tel. Mag.*, V. 28, pp. 165–174, Autumn, 1949.

Growing and Processing of Single Crystals of Magnetic Metals.† J. G. WALKER,[1] H. J. WILLIAMS[1] and R. M. BOZORTH.[1] *Rev. Sci. Instruments*, V. 20, pp. 947–950, Dec., 1949.

ABSTRACT—Single crystals of nickel, cobalt and various alloys are grown by slow cooling of the melt. They are oriented by optical means and by X-rays, and ground to the desired shape using the technique described.

A Look Around—and Ahead. L. A. WILSON.[3] *Bell Tel. Mag.*, V. 28, pp. 133–136, Autumn, 1949.

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

Contributors to this Issue

R. M. Bozorth, A.B., Reed College, 1917; U. S. Army, 1917–19; Ph.D. in Physical Chemistry, California Institute of Technology, 1922; Research Fellow in the Institute, 1922–23. Bell Telephone Laboratories, 1923–. As Research Physicist, Dr. Bozorth is engaged in research work in magnetics.

R. W. Hamming, B.S. in Mathematics, University of Chicago, 1937; M.A. in Mathematics, University of Nebraska, 1939; Ph.D. in Mathematics, University of Illinois, 1942. Dr. Hamming became interested in the use of large scale computing machines while at Los Alamos, New Mexico, and has continued in this field since joining the Bell Telephone Laboratories in 1946.

W. P. Mason, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals and in the study of ultrasonics.

J. R. Pierce, B.S. in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Dr. Pierce has been engaged in the study of vacuum tubes.