The Bell System Technical Journal

Vol. XVIII *October, 1939* *No. 4*

Experience in Applying Carrier Telephone Systems to Toll Cables

By W. B. BEDELL, G. B. RANSOM and W. A. STEVENS

THE application of carrier telephone systems to toll cable conductors, particularly those conductors in existing cables, is expected to become an important means of providing additional long distance telephone circuits. Eight hundred and thirty-seven route miles have been equipped in the United States to date, and 17 twelve-channel systems have been placed in service, providing a total of about 58,000 circuit miles. Late in 1939, 200 additional route miles are expected to be completed which, together with additional systems on existing routes, will add nine systems and about 48,000 circuit miles to the above figures.

The type of carrier system which has been installed is that described by Messrs. C. W. Green and E. I. Green before the American Institute of Electrical Engineers in 1938, and which is now designated as type K.[1] The problems incident to the application of this system to toll cable conductors may be of general interest and it is the purpose of this paper to describe some of these. This description will start at the point where traffic needs have indicated that additional circuits should be provided along a given route and economic and other considerations have shown that they should be provided by means of type K cable carrier telephone systems. For specific examples, reference will be made to the New York-Charlotte and Detroit-South Bend projects, in which sections the application of the initial type K carrier systems has been completed. Figure 1 shows the geographical location of these installations, as well as some of those sections where type K carrier systems will probably be installed in the future.

[1] For references see end of paper.

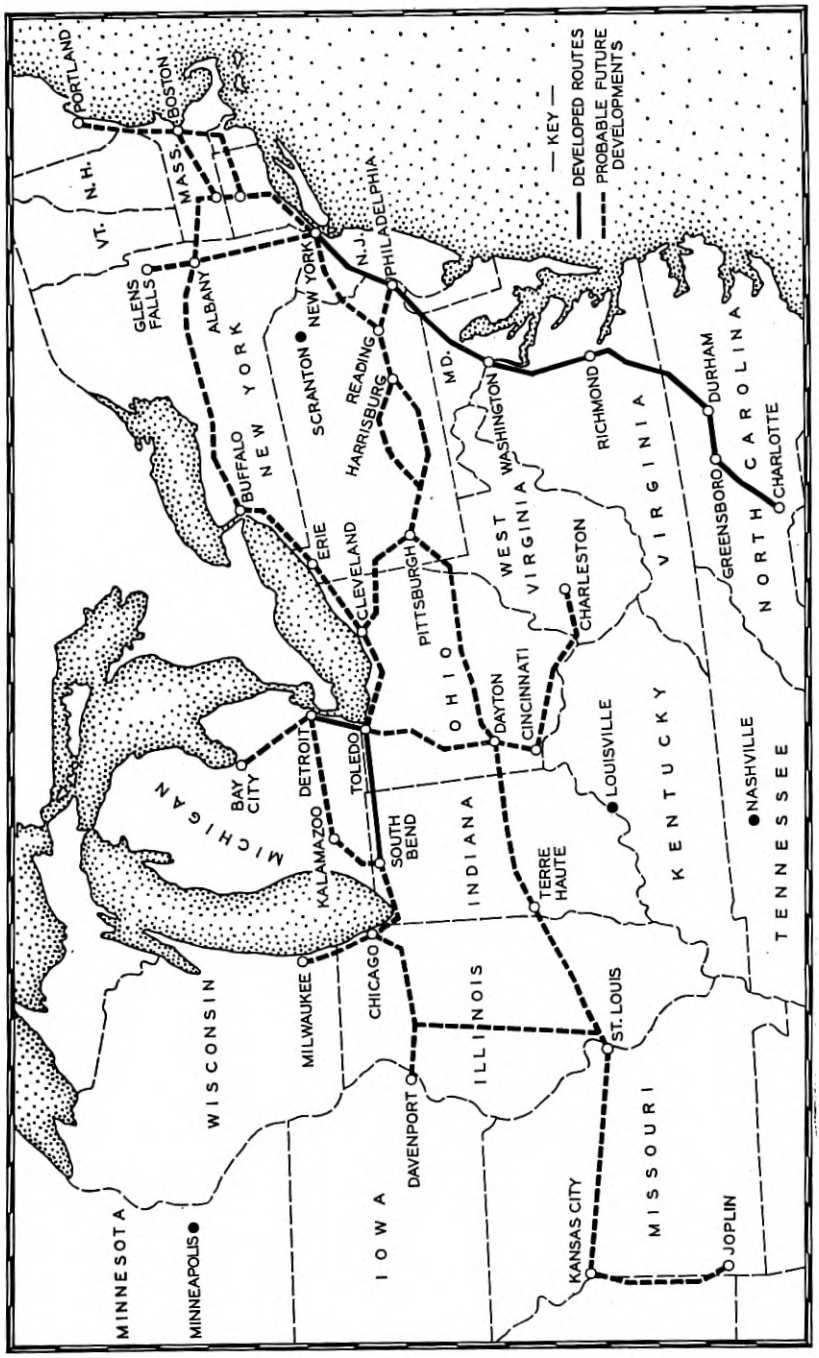

Fig. 1—Routes of existing and probable future type K cable carrier developments.

Selection of Cables to be Equipped

Where more than two cables existed on a route selected for carrier operation, a number of factors influenced the selection of cables to be equipped initially. Among these were the ages of the cables, their makeup, specific route, number of branch cables and open wire junctions, and lengths of underground cable involved. Between Detroit and South Bend there were but two cables on the route selected and hence no selection was necessary. Between New York and Washington on the New York-Charlotte route all cables are underground, and since there were from three to six cables in each repeater section, the two cables which it was decided to employ were selected because they were relatively new and had the smallest number of branches. From Washington to Petersburg, Va., there were two cables, while between Petersburg and Charlotte there was but one cable and it was necessary to install a small second cable chiefly for carrier operation.

The Petersburg-Greensboro section of this second cable was installed one year ahead of the carrier application in order to make use of part of its conductors which were loaded for voice frequency operation. This cable is made up in most sections of 32 quads of 19-gauge conductors, of which 20 are loaded with H-88-50 loading units, leaving 10 quads non-loaded for carrier use and two for maintenance purposes.

The second cable in the Greensboro-Charlotte section was installed coincidentally with the installation of the initial carrier systems. This cable contains 61 non-quadded pairs throughout, except in certain sections where it also contains some loaded conductors for short voice frequency circuits. Paired construction was used because it was expected to be slightly more economical and temporary voice usage of the conductors was not planned.

One additional factor which, in special cases, influences the selection of cables is that of carrier repeater spacing. This is brought about by the fact that on multi-cable routes all of the cables may not follow exactly the same route. For example, one cable may be aerial and the other underground, and the two may be separated in some sections; or underground cables, for conduit reasons, may follow different routes. It is desirable that the two cables used be near each other at repeater points.[2]

One interesting feature in the construction of the Greensboro-Charlotte Cable is the method by which the cable was attached to the messenger. The cable was lashed to the messenger by means of a galvanized steel wire continuous between poles, as shown in Fig. 2. This method of installation is expected to reduce buckling, ring cuts,

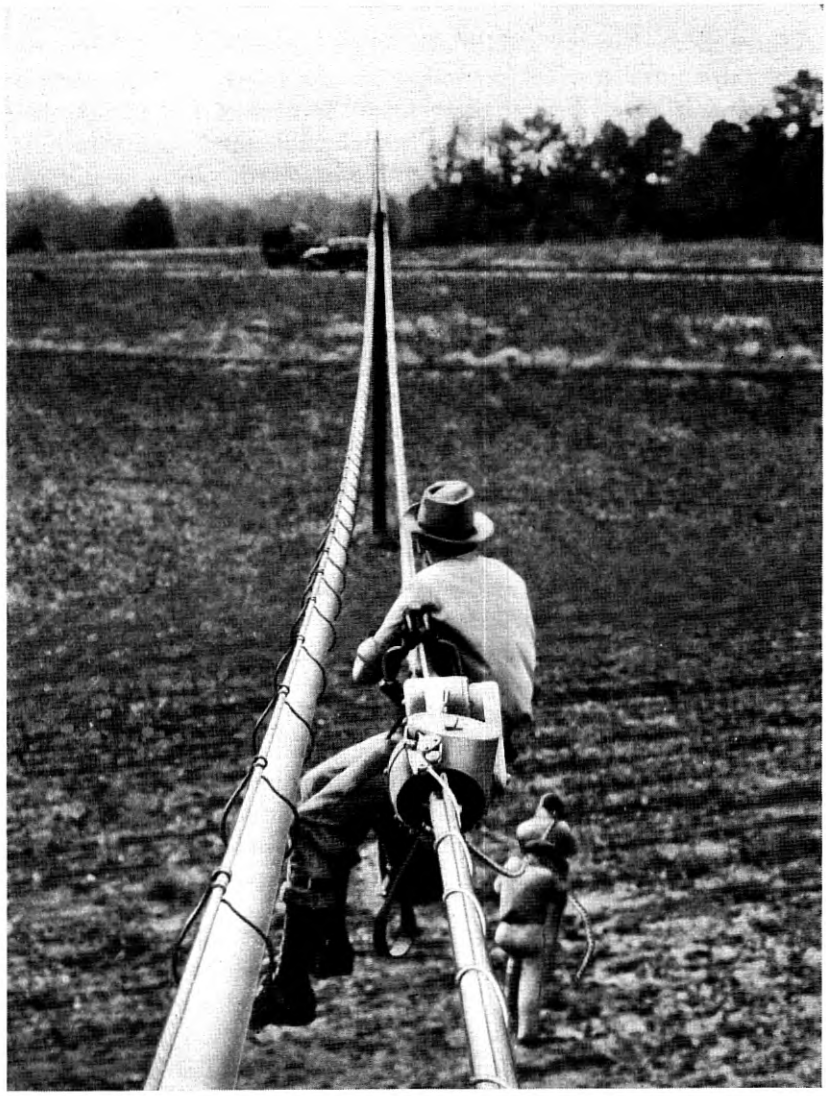

Fig. 2—Method of lashing cable to messenger with galvanized steel wire shown in progress.

and jumping, and avoids the necessity of splicing the cable under tension.

Selection of Conductors for Carrier Use

In general, non-loaded cable pairs are not available in existing cables and it is necessary to remove voice frequency loading from pairs

intended for carrier operation. As has been described previously in a paper in this *Journal*[2] the crosstalk mitigation plans in connection with type K carrier are designed on the assumption that cable pairs will be developed in units of 20 pairs for each direction of transmission. Further, the design of this carrier system contemplates the use of 19-gauge pairs.

Ten quads (20 pairs) were, therefore, selected in each cable in which carrier operation was planned. These quads were selected, for crosstalk reasons, from a large voice complement. Two-wire facilities may be used for carrier where a sufficiently large complement exists. This results, however, in the loss of twice as many voice circuits as compared to unloading four-wire quads. In the sections unloaded to date it has been impracticable to unload two-wire facilities.

Where four-wire facilities were used, five quads from the groups designed for each direction of voice frequency transmission in each cable were selected. Over 20,000 circuit miles of four-wire facilities have been unloaded for carrier use. Of this total, H-174-63 loading units were removed from 2,280 circuit miles, and H-44-25 units were removed from the remainder. The H-44-25 loading units removed from 2,475 miles of four-wire circuits were transferred to two-wire 16-gauge quads loaded with H-174-63 units in the same cable, and these latter units released, thus providing at small cost transmission improvement on a total of 4,950 circuit miles.

Preparation of Cable Conductors

Coincidentally with the removal of the loading from the quads selected for carrier operation, special splicing work was performed for crosstalk and transmission reasons. The exact method of making these splices depended upon the layup of the cable involved. For example, if the cable involved concentric segregation, the five former east bound quads were spliced at random to the five former west bound quads and vice versa at each loading point; in cables involving split layer segregation, the ten quads were spliced at each loading point in a planned random manner.

The removal of the loading at the point nearest the center of each carrier repeater section was left until last, so that a special splice, called a poling splice, based on measurements of within-quad admittance unbalances, might be made at each such point.[5] These measurements could not be made until all loading coils on the carrier pairs in the repeater section had been removed. Using these measurements as a guide the quads in one half-section were connected to quads in

the other half-section so that the unbalances in the two sections tended to compensate. Table 1 shows for a typical type K repeater section

TABLE 1

MEASUREMENTS OF MUTUAL INDUCTANCE UNBALANCE (G) AND CAPACITANCE UNBALANCE (C) BEFORE AND AFTER POLING ON QUADS IN THE PHILADELPHIA-PHILADELPHIA KN SECTION OF THE NEW YORK-PHILADELPHIA E CABLE

Pairs	Before Poling		After Poling	
	G Micromho	C Mmf.	G Micromho	C Mmf.
1–2	.16	110	.12	50
3–4	.12	40	.01	30
5–6	.12	80	.01	45
7–8	.01	40	.02	0
9–10	.07	85	.05	5
11–12	.06	0	0	0
13–14	.10	10	.04	25
15–16	.13	65	0	45
17–18	.11	30	0	20
19–20	.08	25	0	10

the unbalance measurements before poling and the final results after the poling splice was made. These measurements were made at voice frequencies, since as discussed in a previous paper [2] satisfactory results were obtained at these frequencies. It will be noted that poling reduced markedly the unbalances in the quads in the section shown. This is particularly true of the mutual inductance unbalances, indicated in the table by G, the reduction of which is important in reducing crosstalk at carrier frequencies.

After carrying out this and the balancing operations described later for the reduction of crosstalk, it was still necessary to take other steps to reduce the within-quad crosstalk. The recurrence of within-quad coupling between carrier systems which may be assigned to two of the pairs in a 20-pair carrier complement, has been reduced by means of a splicing plan so worked out, that two given carrier systems will operate on the same quad as infrequently as possible. This was accomplished by splitting the quads at the ends of each carrier repeater section on a planned basis. The plan is shown in Table 2. Nineteen types of splices are shown. This table is used as a guide in performing the splice between the balancing bays and the input sealed test terminal. For example, in performing splice type 8, sealed test terminal jacks K-1 are connected to the pair designated 8 at the balancing bay cable terminal, jacks K-2 to the pair designated 16, jacks K-3 to

TABLE 2

DETAILED PLAN FOR SPLITTING QUADS IN NINETEEN CONSECUTIVE CARRIER REPEATER SECTIONS FOR A TEN-QUAD GROUP ARRANGED FOR TYPE K CABLE CARRIER OPERATION

Pair Designation at Sealed Test Terminal Jacks	Planned Splice Type Number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
K-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
K-2	2	3	5	7	9	12	14	16	18	20	1	2	4	6	8	10	13	15	17
K-3	3	4	7	10	14	17	20	3	6	9	19	13	16	19	2	5	8	12	15
K-4	4	5	9	14	18	2	6	10	15	19	3	3	7	12	16	20	4	8	13
K-5	5	6	12	17	2	7	13	18	3	8	17	14	19	4	9	15	20	5	10
K-6	6	7	14	20	6	13	19	5	12	18	5	4	10	17	3	9	16	2	8
K-7	7	8	16	3	10	18	5	13	20	7	15	15	2	9	17	4	12	19	6
K-8	8	9	18	6	15	3	12	20	8	17	7	5	14	2	10	19	7	16	4
K-9	9	10	20	9	19	8	18	7	17	6	13	16	5	15	4	14	3	13	2
K-10	10	11	11	11	11	11	11	11	11	11	9	11	11	11	11	11	11	11	11
K-11	11	12	2	13	3	14	4	15	5	16	14	6	17	7	18	8	19	9	20
K-12	12	13	4	16	7	19	10	2	14	5	8	17	8	20	12	3	15	6	18
K-13	13	14	6	19	12	4	17	9	2	15	16	7	20	13	5	18	10	3	16
K-14	14	15	8	2	16	9	3	17	10	4	6	18	12	5	19	13	6	20	14
K-15	15	16	10	5	20	15	9	4	19	14	18	8	3	18	13	7	2	17	12
K-16	16	17	13	8	4	20	16	12	7	3	4	19	15	10	6	2	18	14	9
K-17	17	18	15	12	8	5	2	19	16	13	20	9	6	3	20	17	14	10	7
K-18	18	19	17	15	13	10	8	6	4	2	2	20	18	16	14	12	9	7	5
K-19	19	20	19	18	17	16	15	14	13	12	12	10	9	8	7	6	5	4	3
K-20	20	1	1	1	1	1	1	1	1	1	10	1	1	1	1	1	1	1	1

Note: Above numbers are pair number designations at balancing bay cable terminals of pairs which are connected to Sealed Test terminal jacks as indicated.

pair 3, etc. Using this plan two systems are exposed to each other on the same quad only once in 19 carrier repeater sections. Beginning with the 20th section the plan is repeated so that between New York and Charlotte two given systems operate together on the same quad in only two repeater sections. Planned splices were made at each end of each carrier repeater section so that after the quads had been split in a definite way at one end of a section, a complementary splice was made at the other end, to rearrange the pairs into a given order as they go through each repeater office. Each carrier pair has been made to appear in the same position at each carrier testboard throughout the 19 types of planned splice sections. This is for convenience in maintaining and identifying them, because carrier systems must be assigned to the same carrier pair in a series throughout the 19 types of planned splice sections if the quad splitting plan is to be completely effective.

The poling splice and the planned splices were, of course, not re-

quired in the paired cable placed for carrier operation between Greensboro and Charlotte.

Far-end crosstalk was still further reduced by means of balancing coils installed at the end of each repeater section connected to the repeater inputs.[2, 8] After the splicing operations just discussed were completed and the balancing coils were installed and connected to the carrier pairs, the coupling between each pair and each other pair of the carrier complement was reduced to the lowest value practicable by adjustment of these coils. This was accomplished by sending a disturbing testing tone on one pair, receiving on each other pair in turn, and adjusting the coil which couples each combination of two pairs until a minimum amount of the testing tone was measured on the disturbed pair. Figure 3 shows these adjustments in progress while Table 3 shows a summary of the final crosstalk measurements made after the adjustments. About 19,000 balancing coils have been installed on the carrier routes equipped to date.

At two points on the New York-Charlotte route, retardation coils which are described later were used to increase the attenuation in potential crosstalk paths. At Petersburg, Va., and Burlington, N. C., 60 and 4 voice quads, respectively, connected direct from one carrier cable to the other. These quads, through secondary induction, were likely to serve as crosstalk paths at carrier frequencies between the two cables. Retardation coils were installed in each quad to limit crosstalk currents. Two other situations, where quads connected between the carrier cables, were eliminated by cable rearrangements.

Lateral Cables

At each carrier repeater station four lateral cables were installed to bring the carrier pairs into the repeater building. One of these was required for each direction of transmission for each cable; that is, two input cables and two output cables were required. As a result of the transposition of directions of transmission at each repeater point, the two input cables connect to one toll cable and the two output cables to the other. Figure 4 shows these cables installed at an aerial cable repeater point. All lateral cables were installed for the probable ultimate capacity for carrier systems of the cables being developed; that is, 100 systems on the Detroit-South Bend route, 100 systems north and 60 systems south of Richmond, Va., on the New York-Charlotte route.

The Location of Carrier Repeater Station Sites

The next important step in the development of a route for type K carrier operation is the selection of points at which intermediate

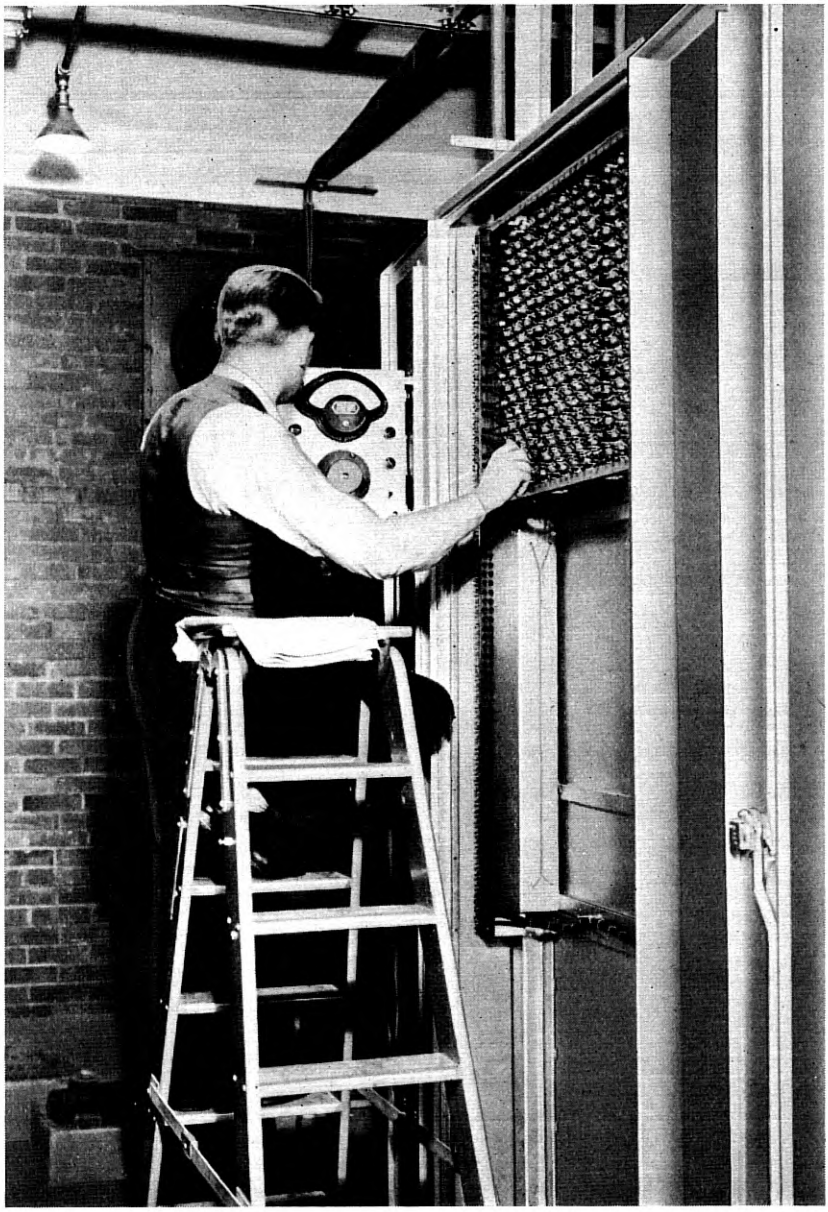

Fig. 3—Adjustment of coils in crosstalk balancing panel.

TABLE 3

FINAL FAR-END CARRIER CROSSTALK MEASUREMENTS AFTER BALANCING COIL ADJUSTMENTS

Section	New York Toward Charlotte					Charlotte Toward New York				
	Ca.	Crosstalk Units 39.85 kc		Crosstalk Units 28.15 kc		Ca.	Crosstalk Units 39.85 kc		Crosstalk Units 28.15 kc	
		R.M.S.	Max.	R.M.S.	Max.		R.M.S.	Max.	R.M.S.	Max.
New York-N. Y. KS	E	42	283	36	122	F	33	184	31	100
N. Y. KS-Prin. KN	F	37	148	36	122	E	29	95	27	92
Prin. KN-Prin.	E	38	130	36	114	F	40	212	37	155
Prin.-Prin. KS	G	37	212	33	130	E	35	132	36	130
Prin. KS-Phila. KN	E	35	148	32	164	G	30	100	28	112
Phila. KN-Phila.	G	35	250	31	100	E	29	114	29	127
Phila.-Phila. KS	D	45	243	44	226	F	37	127	33	122
Phila. KS-Elk. KN	F	29	116	28	112	D	45	138	45	122
Elk. KN-Elk.	D	45	161	45	145	F	36	145	36	130
Elk.-Elk. KS	F	36	204	36	164	D	42	217	42	176
Elk. KS-Balt. KN	D	38	126	39	129	F	35	207	33	145
Balt. KN-Balt.	F	35	107	34	114	D	45	224	45	170
Balt.-Wash. KN	D	44	241	42	179	E	43	219	38	148
Wash. KN-Wash.	E	39	179	38	195	D	34	100	36	126
Wash.-Wash. KS	A	43	184	43	148	B	44	182	48	158
Wash. KS-Fred. KN	B	43	224	40	167	A	47	141	46	155
Fred. KN-Fred.	A	54	179	50	163	B	44	170	46	208
Fred.-Fred. KS	B	45	173	43	152	A	49	219	48	219
Fred. KS-Rich. KN	A	47	167	46	164	B	39	167	39	127
Rich. KN-Rich.	B	43	167	46	176	A	39	200	38	125
Rich.-Rich. KS	A	39	179	40	127	B	30	190	39	158
Rich. KS-McK. KN	B	48	138	53	155	A	37	265	35	200
McK. KN-McK.	A	36	130	34	100	B	56	148	64	167
McK.-McK. KS	B	57	190	62	174	A	38	152	38	155
McK. KS-Norl. KN	A	36	145	35	145	B	56	161	62	187
Norl. KN-Norl.	B	53	173	56	170	A	41	198	39	190
Norl.-Norl. KS	A	34	110	38	134	B	59	167	59	170
Norl. KS-Dur. KN	B	56	179	56	195	A	39	142	40	142
Dur. KN-Dur.	A	42	190	38	190	B	65	200	62	187
Dur.-Dur. KS	B	61	155	58	145	A	38	114	43	134
Dur. KS-Gbo. KN	A	42	195	42	118	B	59	205	59	187
Gbo. KN-Gbo.	B	57	219	58	195	A	37	118	36	107
Gbo.-Gbo. KS	A	39	145	37	141	B	55	173	50	155
Gbo. KS-Sal. KN	B	57	338	54	346	A	40	224	38	161
Sal. KN-Sal.	A	34	161	35	155	B	43	245	50	300
Sal.-Sal. KS	B	54	265	52	245	A	35	126	32	107
Sal. KS-Chlot. KN	A	34	141	32	155	B	53	245	53	286
Chlot. KN-Chlot.	B	43	200	42	155	A	33	167	31	129

Note: These measurements were made in connection with balancing work where relative values are important and do not necessarily represent the absolute magnitude of the crosstalk.

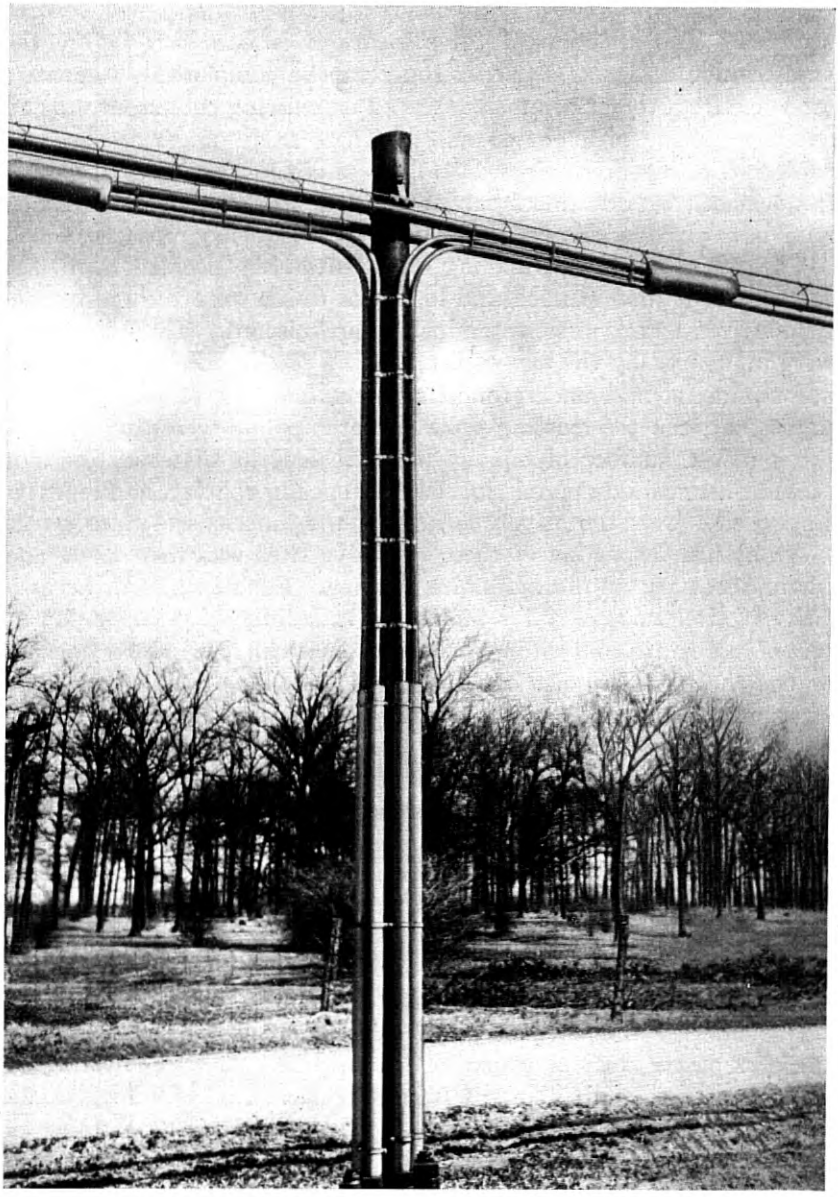

Fig. 4—The four lateral cables containing carrier pairs shown as installed at an auxiliary repeater point on aerial cables.

repeater stations, known as auxiliary repeater stations, will be constructed. The selection of these locations is necessary before the transmission design of a carrier route can be completed. Repeaters for voice frequency circuits are located on existing cable routes at an average spacing of about 45 miles. These same offices are used as cable carrier repeater points, but, because of the high losses at carrier frequencies, two additional carrier repeater stations, on the average, have been provided between each two voice frequency repeater offices.

The cables and routes having been tentatively decided upon, the route records were studied and locations which were the most practicable from a transmission standpoint were selected. These selections were influenced by the expected maximum section losses, taking into account aerial and underground construction. In general, the distances between the existing voice repeater points were divided into the smallest number of equal parts, the lengths of which did not exceed the maximum permissible carrier repeater spacing, and repeater station sites were tentatively located at the junctions of these parts.

A physical inspection of these tentative sites was then made and where necessary an alternate site selected. Such factors as accessibility of site, suitability for building, availability of primary power, cost of real estate, and willingness of owners to sell determined whether or not the tentative site could be used and, if not, what alternate location might be used. Where a suitable existing telephone building happened to be located near a proposed repeater station location, the possibility of using such building was studied. It has been practicable, however, in only one instance to use an existing building in installing the 34 auxiliary stations provided to date. The sites which were considered satisfactory after the physical inspection were then examined to check their suitability from a carrier transmission standpoint. In cases where transmission limits had been exceeded, the sites were reinspected and compromise locations finally agreed upon. In most cases it has not been difficult to find sites which are suitable both from a transmission and a construction standpoint. Of the 34 type K carrier repeater stations which have been built, 20 were constructed at the sites originally selected from a transmission standpoint. In most of the other cases the final sites were within a short distance of the originally selected locations. In a few cases, however, where ideal sites fell in populous centers or comparatively inaccessible wilds, it was necessary to take unusual steps.

In one case the site which had been selected from a transmission standpoint fell at a location where the two cables which had been selected followed different conduit runs and were separated by more

than $2\frac{1}{2}$ miles. Two lateral cables, each of that length, would have been required in order to make use of this site. Moving the location back to the point where the cables came together would have resulted in an excessively long carrier repeater section and would have located the station within the business section of Wilmington, Del. The problem was to find a compromise site between these two points where the lengths of lateral cables would not be excessive and the repeater section could be kept within desirable limits. A tide water stream between these two points complicated the problem.

Nine sites were inspected and four of them studied in detail. Flood and fire hazards, as well as high prices of real estate, were added to the other factors governing the choice. None of the locations was entirely desirable, but a compromise choice was finally made of a location which resulted in the longest repeater spacing in the New York-Charlotte project, but the lengths of the lateral cables were reduced to between eleven and twelve hundred feet.

Table 4 shows the theoretical spacings which, considering the fixed

TABLE 4

COMPARISON OF THE ORIGINALLY SPECIFIED CARRIER REPEATER SPACINGS AND ACTUAL SPACINGS ON THE NEW YORK-CHARLOTTE AND DETROIT-SOUTH BEND CABLE ROUTES

Project	No. of Repeater Sections	Theoretically Best Spacings			Actual Spacings Used		
		Min.	Ave.	Max.	Min.	Ave.	Max.
New York-Charlotte...	38	13.5	16.5	18.8	13.2	16.5	20.2
Detroit-South Bend...	13	14.9	16.1	17.6	13.6	16.1	18.4

location of existing repeater points, were selected as best from a transmission standpoint, and the actual spacings which it was found necessary to use. The theoretically best spacings for the two routes differed because of the difference in spacings of the existing voice frequency repeater points of which use has been made as carrier repeater points. Figure 5 shows the repeater office locations as they are distributed on the New York-Charlotte project. The various types of repeater offices shown are discussed later in this paper.

DIRECTION OF TRANSMISSION

The circuits used for the two directions of transmission of type K carrier systems operate in separate cables on the projects so far completed and, for crosstalk reasons, these two directions of transmission have been transposed between the two cables at each carrier repeater point.[2] The location of branch cables and taps to open wire have an

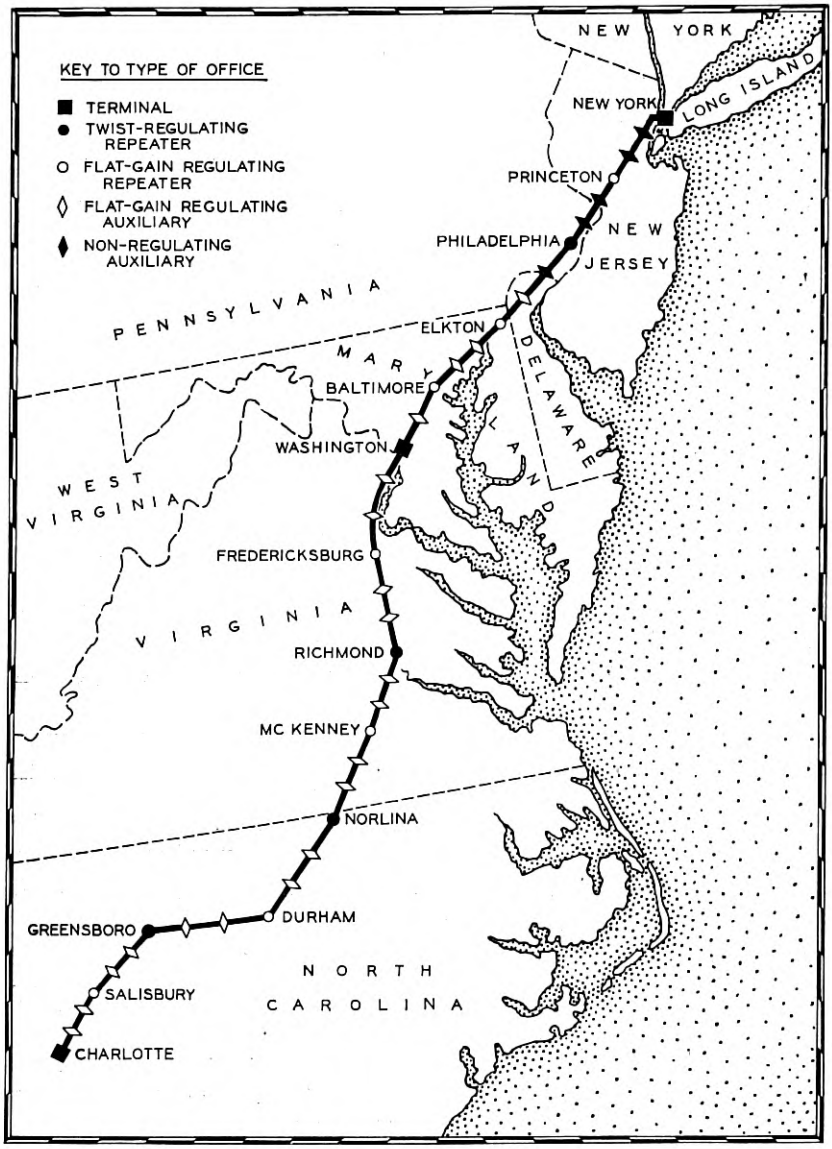

Fig. 5—Route of New York-Charlotte cable carrier development.

important bearing on the selection of the direction of transmission, since it is desirable to assign directions of transmission which will result in a minimum number of taps to open wire and branch cables occurring near carrier repeater inputs. Where the lengths of under-

ground construction adjacent to a repeater station differ on the two cables, it is preferable to have pairs in the cable having the longer section of underground connected to repeater inputs. With these and other factors in mind, tentative directions of transmission were assigned to the cable conductors and computations of expected noise currents made and checked with computations which assumed the directions reversed. The total overall noise currents computed to be 1.25 db better when assuming the directions of transmission finally selected for the New York-Charlotte project than when assuming these directions reversed. This was largely due to the fact that south of Petersburg, where but one cable had existed, a small cable was added to permit carrier operation. Pairs in this cable, because of its small size, are more susceptible to static induction directly into the cable than pairs in larger cables. The effect, therefore, of the greater contribution to overall noise currents, which this small cable tends to cause, was reduced by selecting the directions of transmission so as to take advantage of the increased shielding resulting from underground construction adjacent to repeaters. Tables 5 and 6 show the final noise level computations for the 1000-cycle point of channel 12 (57 kc on the line) of a New York-Charlotte system. It will be noted that the longer repeater sections contribute a great deal of noise as compared to average or shorter sections. Noise measurements which have been made indicate that noise conditions compare favorably with those which it was calculated might be expected.

Non-Regulating Repeater Points

Examination of the carrier repeater sections on the New York-Charlotte route showed seven to be unusually short and involving all underground cable construction. The usual plan would have been to provide flat gain regulation at each carrier repeater point, but since these seven sections averaged but 14.8 miles in length and the theoretical transmission variation might be but $\pm$ 1.42 db, it was obvious that the regulators having a normal range of $\pm$ 7.15 db would be required to operate only over a small part of their range. However, the real limitation is not the regulating mechanism but the lower levels to which the line currents without regulated gain would drop during periods of high cable temperature with corresponding impairments in noise levels. In this layout, omission of regulation at one station increases the noise level about the same as lengthening the following repeater section about $\frac{3}{4}$ of a mile. Omission of two successive regulators is approximately equivalent to increasing the second repeater section about $\frac{3}{4}$ of a mile and the third about $1\frac{1}{2}$ miles. Repeater

TABLE 5

NEW YORK-CHARLOTTE TYPE K CARRIER NOISE COMPUTATIONS

Section	Cable	Miles			57 kc Loss Max. Temp.[1]	Estimated 57 kc Noise Level[2] at Rept. Output
		U.G.	Aerial	Total		
New York-New York KS	E	15.45	—	15.45	58.09	− 8.91
New York KS-Princeton KN	F	14.45	—	14.45	54.33	− 9.71
Princeton KN-Princeton	E	16.59	—	16.59	62.38	+ 1.10
Princeton-Princeton KS	G	13.22	—	13.22	49.71	−13.29
Princeton KS-Philadelphia KN	E	13.92	—	13.92	52.34	−10.75
Philadelphia KN-Philadelphia	G	14.79	—	14.79	55.61	− 2.19
Philadelphia-Philadelphia KS	D	15.14	—	15.14	56.93	−10.07
Philadelphia KS-Elkton KN	F	13.49	—	13.49	50.72	− 9.38
Elkton KN-Elkton	D	19.70	—	19.70	74.07	+ 8.57
Elkton-Elkton KS	F	17.28	—	17.28	64.97	− 2.03
Elkton KS-Baltimore KN	D	16.89	—	16.89	63.51	− 3.49
Baltimore KN-Baltimore	F	16.91	—	16.91	63.58	− 3.42
Baltimore-Washington KN	D	18.61	—	18.61	69.97	+ 2.97
Washington KN-Washington	E	18.95	—	18.95	71.25	+ 4.25
Washington-Washington KS	A	7.31	11.28	18.59	71.93	+ 5.93
Washington KS-Fredericksburg KN	B	—	18.45	18.45	72.69	+ 6.69
Fredericksburg KN-Fredericksburg	A	.10	16.54	16.64	65.55	+ 2.95
Fredericksburg-Fredericksburg KS	B	.11	18.23	18.34	72.24	+ 6.24
Fredericksburg KS-Richmond KN	A	—	18.26	18.26	71.94	+ 5.94
Richmond KN-Richmond	B	7.64	10.99	18.63	72.03	+ 5.03
Richmond-Richmond KS	A	9.82	7.09	16.91	64.85	− 1.15
Richmond KS-McKenney KN	B	10.13	5.85	15.98	61.14	+ 8.54
McKenney KN-McKenney	A	.08	15.43	15.51	61.09	− 1.51
McKenney-McKenney KS	B	.11	16.87	16.98	66.88	+14.08
McKenney KS-Norlina KN	A	—	15.23	15.23	60.01	− 2.59
Norlina KN-Norlina	B	.06	14.28	14.34	56.49	+ 3.89
Norlina-Norlina KS	A	—	16.36	16.36	64.46	+ 3.66
Norlina KS-Durham KN	B	—	16.68	16.68	65.72	+13.12
Durham KN-Durham	A	.12	17.62	17.74	69.87	+ 3.87
Durham-Durham KS	B	.05	18.05	18.10	71.31	+18.51
Durham KS-Greensboro KN	A	—	17.79	17.79	70.09	+ 4.09
Greensboro KN-Greensboro	B	2.38	16.25	18.63	72.98	+12.18
Greensboro-Greensboro KS	A	5.64	12.18	17.82	69.20	+ 3.20
Greensboro KS-Salisbury KN	B	—	16.41	16.41	64.66	+12.06
Salisbury KN-Salisbury	A	1.85	14.84	16.69	65.43	− 2.53
Salisbury-Salisbury KS	B	1.82	11.92	13.74	53.80	+ 1.2
Salisbury KS-Charlotte KN	A	—	14.47	14.47	57.01	− 4.09
Charlotte KN-Charlotte	B	3.78	9.96	13.74	53.45	− 8.35
New York-Charlotte Totals				627.42		+23.27[3]

[1] Attenuation figures used: 3.76 for U.G. at 73° and 3.94 for Aerial at 110°.
[2] For top channel, referred to − 9 db switchboard level.
[3] Computed on a root-sum-square basis.

sections adjacent to New York and Philadelphia are short and it was decided, therefore, to omit regulation from the two auxiliary stations between New York and Princeton, N. J., two between Princeton and Philadelphia, and one between Philadelphia and Wilmington, Del.

TABLE 6

CHARLOTTE-NEW YORK TYPE K CARRIER NOISE COMPUTATIONS

Section	Cable	Miles			57 kc Loss Max. Temp.[1]	Estimated 57 kc Noise Level[2] at Rept. Output
		U.G.	Aerial	Total		
Charlotte-Charlotte KN	A	3.78	9.96	13.74	53.45	− 9.15
Charlotte KN-Salisbury KS	B	—	14.47	14.47	57.01	+ 5.91
Salisbury KS-Salisbury	A	1.82	11.92	13.74	53.80	− 7.60
Salisbury-Salisbury KN	B	1.85	14.84	16.69	65.43	+14.13
Salisbury KN-Greensboro KS	A	—	16.41	16.41	64.66	+ 2.06
Greensboro KS-Greensboro	B	5.64	12.18	17.82	69.20	+ 4.70
Greensboro-Greensboro KN	A	1.66	16.97	18.63	73.10	+ 7.10
Greensboro KN-Durham KS	B	—	17.79	17.79	70.09	+17.29
Durham KS-Durham	A	.05	18.05	18.10	71.31	+ 5.31
Durham-Durham KN	B	.12	17.62	17.74	69.87	+17.07
Durham KN-Norlina KS	A	—	16.68	16.68	65.72	+ 3.12
Norlina KS-Norlina	B	—	16.36	16.36	64.46	+11.86
Norlina-Norlina KN	A	.06	14.28	14.34	56.49	− 6.11
Norlina KN-McKenney KS	B	—	15.23	15.23	60.01	+ 8.91
McKenney KS-McKenney	A	.11	16.87	16.98	66.88	+ .78
McKenney-McKenney KN	B	.08	15.43	15.51	61.09	+ 8.49
McKenney KN-Richmond KS	A	6.18	9.81	15.99	61.89	− .71
Richmond KS-Richmond	B	16.84	.07	16.91	63.60	+ .60
Richmond-Richmond KN	A	7.52	11.08	18.60	71.94	+ 5.94
Richmond KN-Fredericksburg KS	B	—	18.26	18.26	71.94	+ 5.94
Fredericksburg KS-Fredericksburg	A	.10	18.23	18.33	72.21	+ 6.11
Fredericksburg-Fredericksburg KN	B	.10	16.54	16.64	65.55	+ 2.95
Fredericksburg KN-Washington KS	A	—	18.45	18.45	72.69	+ 6.69
Washington KS-Washington	B	8.69	9.91	18.60	71.72	+ 4.72
Washington-Washington KN	D	18.90	—	18.90	71.06	+ 4.06
Washington KN-Baltimore	E	18.64	—	18.64	70.09	+ 3.09
Baltimore-Baltimore KN	D	16.89	—	16.89	63.51	− 3.49
Baltimore KN-Elkton KS	F	16.86	—	16.86	63.39	− 3.61
Elkton KS-Elkton	D	17.33	—	17.33	65.16	− 1.84
Elkton-Elkton KN	F	20.21	—	20.21	75.99	+ 8.99
Elkton KN-Philadelphia KS	D	13.51	—	13.51	50.80	−12.20
Philadelphia KS-Philadelphia	F	16.00	—	16.00	60.16	− 1.34
Philadelphia-Philadelphia KN	E	13.36	—	13.36	50.23	−12.77
Philadelphia KN-Princeton KS	G	13.93	—	13.93	52.38	− 9.34
Princeton KS-Princeton	E	13.49	—	13.49	50.72	− 9.67
Princeton-Princeton KN	F	16.69	—	16.69	62.75	− 2.75
Princeton KN-New York KS	E	14.50	—	14.50	54.52	− 5.38
New York KS-New York	F	15.38	—	15.38	57.88	− .69
Charlotte-New York Totals				627.70		+23.46[3]

[1] Attenuation figures used: 3.76 for U.G. at 73° and 3.94 for Aerial at 110°.
[2] For top channel, referred to − 9 db switchboard level.
[3] Computed on a root-sum-square basis.

Figure 6 shows a comparison of the computed levels for maximum cable temperature conditions in the New York-Philadelphia Cable under conditions of both regulation and non-regulation. Omitting the regulation from a carrier repeater office where noise conditions and

gain requirements are favorable has the advantage of economy in saving the cost of the regulating apparatus and in an expected saving in maintenance.

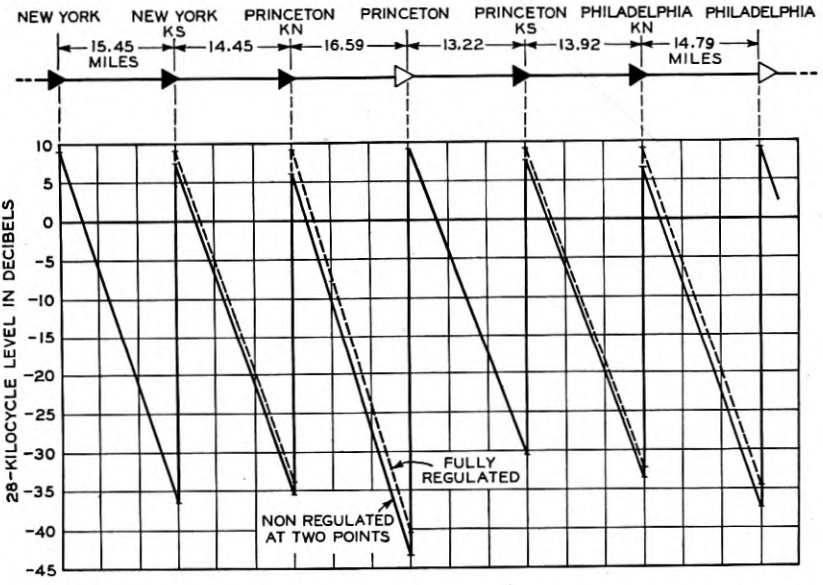

Fig. 6—Level diagram of New York-Philadelphia section, showing theoretical maximum effect on 28 kc levels of omitting regulation at the auxiliary repeater stations.

Noise Suppression Devices

Two types of noise suppression devices were used on voice frequency circuits to limit noise currents which might enter the cables used for carrier.[2] The first of these is a retardation coil designed to suppress longitudinal currents but to have a negligible effect on the metallic currents on the side and phantom circuits. These were installed at voice frequency repeater points, in all voice quads in cables which contained carrier pairs connected to carrier repeater inputs. These coils attenuate longitudinal noise currents at carrier frequencies generated in the voice repeater office which might enter the cable over the pairs used for voice circuits and be induced into the carrier pairs. A total of 3,000 retardation coils was installed along the New York-Charlotte route for this purpose. The coils are connected into the cable conductors on the office side of the point at which the carrier lateral cable is connected to the main cable. These coils were installed in the office cable vault or manhole. Figure 7-A shows a number of

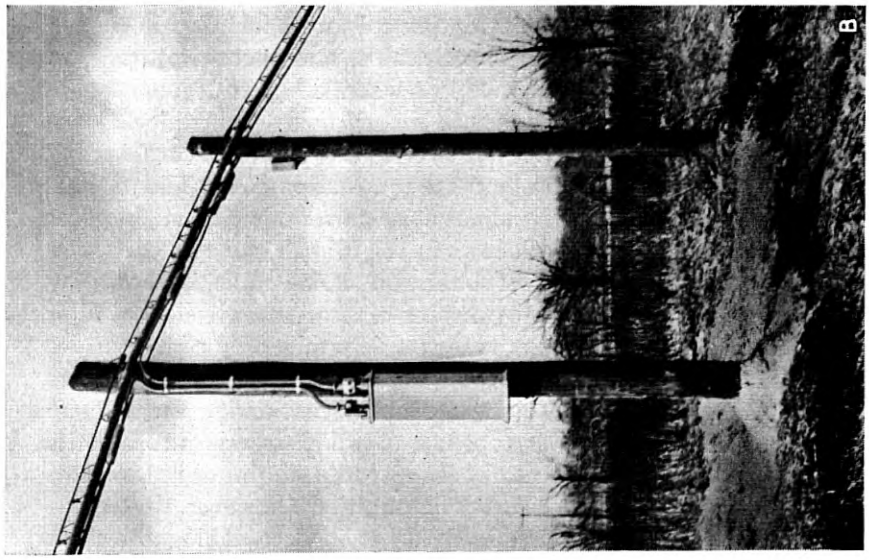

Fig. 7-B—Installation of filters on aerial cable.

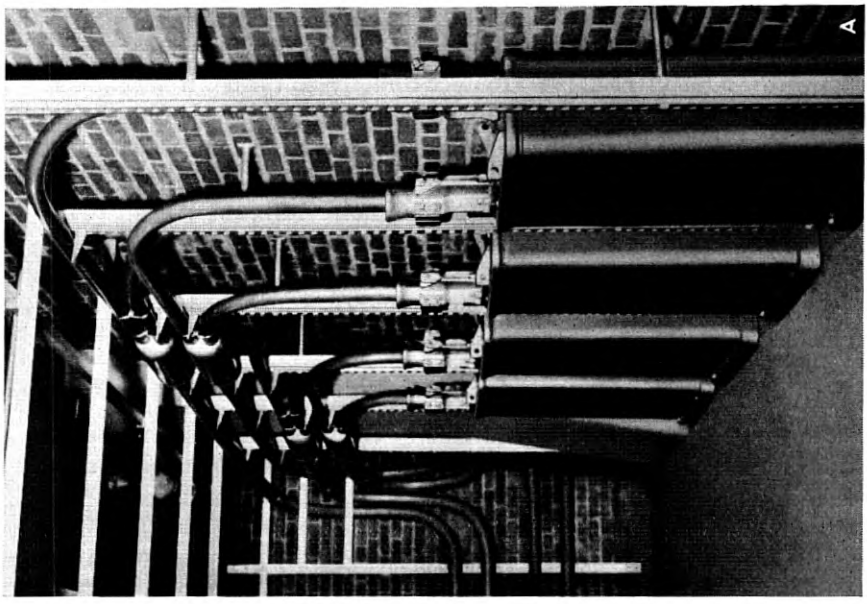

Fig. 7-A—Retardation coil cases installed in cable vault at West Unity, Ohio.

apparatus cases containing these retardation coils installed in the cable vault at West Unity, Ohio.

In addition to this usage, retardation coils were installed in certain instances in the conductors of branch cables and open wire taps. In other branch cables and open wire taps where a higher degree of noise current suppression was required the second noise suppression device was used.[6] This second device is a filter which provides a considerably greater degree of suppression than the retardation coil. The purpose of these is to attenuate longitudinal noise currents which might enter the main cable over the conductors in the branch cable or open wire tap. A typical installation of filters on aerial cable is shown by Fig. 7-B. The question as to whether a retardation coil or a filter was required in each particular case was determined by computations of expected noise which might be contributed by the conductors entering the main cable. This was done by considering the makeup and length of the branch or tap, use to which it was put, and its location with respect to the nearest carrier repeater input in the cable to which it was connected. These computations, however, were made coincidentally with those described earlier in determining the most desirable directions of transmission.

Five hundred fifty-one retardation coils and 132 filters were installed in the 26 branch cables and open wire taps along the New York-Charlotte route, 11 of which connect directly to open wire. On the Detroit-South Bend project, 12 branch cables and open wire taps were equipped with 44 retardation coils and 124 filters. Nine of the 12 are taps connected directly to open wire.

As a further step toward prevention of noise currents in the carrier pairs, the shielding furnished by the lead sheaths of the cables has been kept effective by maintaining continuous the electrical path through these sheaths by means of shunts consisting of large condensers placed across each insulating joint.[4]

Auxiliary Repeater Station Buildings

Small buildings to house the auxiliary repeaters have been erected at the sites determined to be acceptable from transmission and construction standpoints. These structures are of fire resistive construction with concrete foundations, brick walls, and slate roofs. Since these buildings house equipment which is expected to operate for long periods of time without attention, no openings have been provided in the walls except for an entrance door and ventilating units. Thermal insulation has been provided over the ceiling. Two sizes of buildings have been used. The larger one which is 24 ft. × 24 ft., inside dimen-

sions, is used on routes where an ultimate of 100 systems is expected, while the smaller one, 21 ft. × 24 ft., is used on a route to be developed for a maximum of 60 systems. The ceiling height in these buildings is sufficient to care for 11′-6″ relay racks. Eleven of the small buildings and 22 of the larger ones have been built on the carrier projects so far completed.

The architectural treatment of the exterior of these buildings varies somewhat, depending upon the location of the site selected and the character of the buildings in the immediate neighborhood. The present designs may be classified as three types; i.e., plain brick with no trim, plain brick with limestone trim, and plain brick with limestone trim and artificial windows. In the latter type the window arrangements are obtained by the use of a wooden frame and sash with rough wire glass, backed by the interior brick wall. The brick portion behind the window is painted buff on the upper half facing the window, and black on the lower half, to simulate a true window with the shade half drawn. Typical examples of these types may be seen in Fig. 8. The type of building selected for each station depended upon the locality.

Arrangements have been provided in these buildings for automatically controlling the heating and ventilation by the use of thermostatically controlled electric heater and fan units. Although experience was generally lacking on the heating and ventilating problem for these stations, tentative requirements were set up. A minimum temperature of 40° F. has been considered satisfactory for the operation of the equipment in these stations and the thermostat has been set to turn on the heater unit if the inside temperature drops below that point.

Ventilating equipment consisting of intake and exhaust ventilators, exhaust fan and control equipment has been provided so that advantage may be taken of the effect of cooler outside air when the temperature inside the buildings rises to about 90° F. Consideration was given to the direction of the prevailing winds in locating these units in the building walls. The room side of the intake ventilator unit is equipped with a spun glass filter. These ventilators are equipped with rigid and movable louvers. The movable louvers are actuated by solenoids which are connected to the exhaust fan control which functions by means of a thermostat and a differential temperature control. The latter includes outside and inside temperature compensating elements. The thermostat is set at 90° which, with the differential feature of the control, will cause the louvers to open and the exhaust fan to start only when the inside temperature is more than 10° above that prevailing outside the building. When the inside temperature has been reduced to within 10° of that outside, the control circuit is opened to shut off the fan and close the louvers.

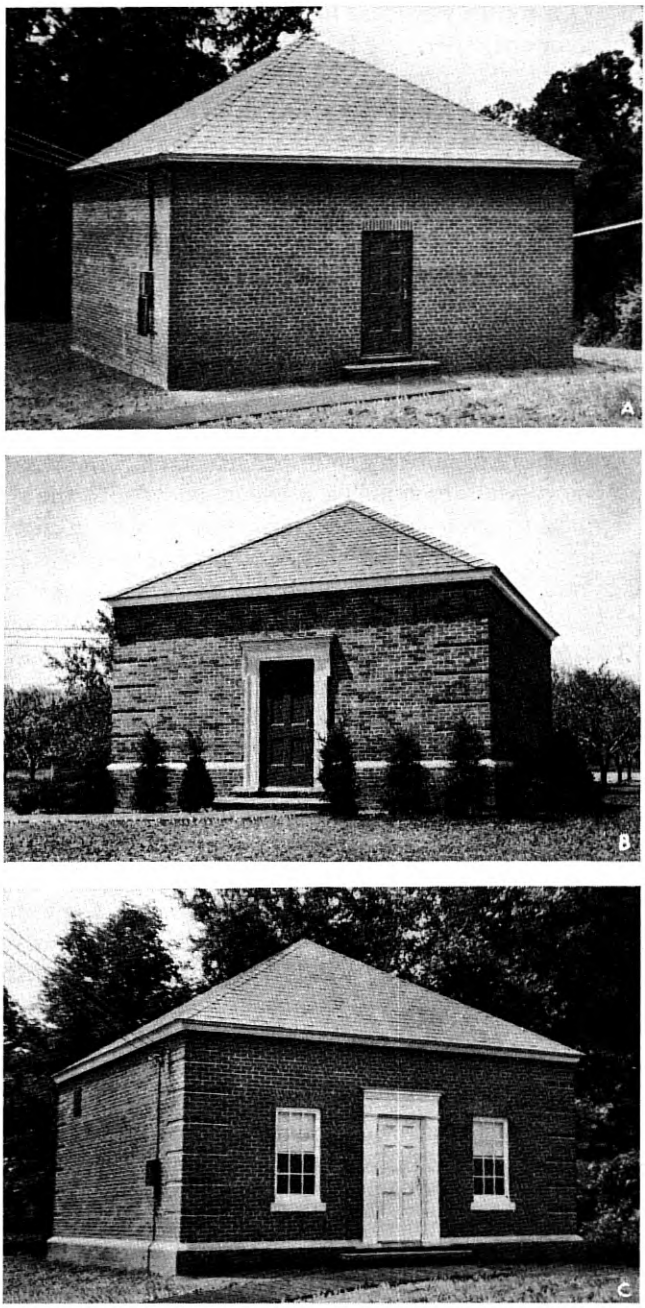

Fig. 8—Auxiliary repeater station buildings: A, without trim; B, with limestone trim; C, with limestone trim and artificial windows.

Terminal Office Equipment

The various major items of equipment which are provided at a terminal office are shown schematically by Fig. 9. Two bays of sealed test terminals—one input and one output—are associated with the pair of cables which bring the carrier pairs into the office and are equipped initially to terminate 40 pairs. The input high-frequency jacks are mounted in high-frequency patching bays adjacent to the input sealed test terminal bays and the output jacks in bays adjacent to the output sealed test terminal bays. At points where more than 50 terminals are expected to be required at some future date plans have been made for one input and two output high-frequency patching bays. The input and output high-frequency patching and sealed test terminal bays for two cable routes, together with a high-frequency transmission measuring bay, are grouped so as to form a desirable arrangement for testing and maintenance purposes. This group of bays serves somewhat the same purpose as a primary testboard on voice frequency facilities. The arrangement of these bays as installed at New York is shown in Figs. 10-A and 10-B.

A portable transmission measuring set has been provided and may be placed on a writing shelf mounted in the high-frequency transmission measuring bay. This set may be connected to the various circuits by means of patching cords.

Line and twist amplifiers with associated flat and twist gain master controller equipment and crosstalk balancing bays also are installed at each terminal office. The arrangement of the amplifier equipment and controllers as installed at New York is shown in Fig. 11.

The terminal equipment for one system consists of six channel modem (modulator plus demodulator) panels, each of which mounts the sending and receiving apparatus for two channels.[3] Channel modem equipment for three systems is mounted in two adjacent bays with the first two systems occupying separate bays. One bay of carrier supply equipment for each ten systems provides both regular and emergency units for generating carrier frequencies for the operation of the channel and group modem units and pilot channel equipment. The group modem units, one of which is required for each system, are mounted nine in one bay. Figures 12 and 13, respectively, show the method in which these bays are installed at New York.

The d-c power supply for the carrier equipment at terminal and main repeater offices is obtained from the existing 24-volt and 130-volt office power plants. Two sets of main distributing leads have been provided for each filament and plate power supply. Odd numbered circuits are

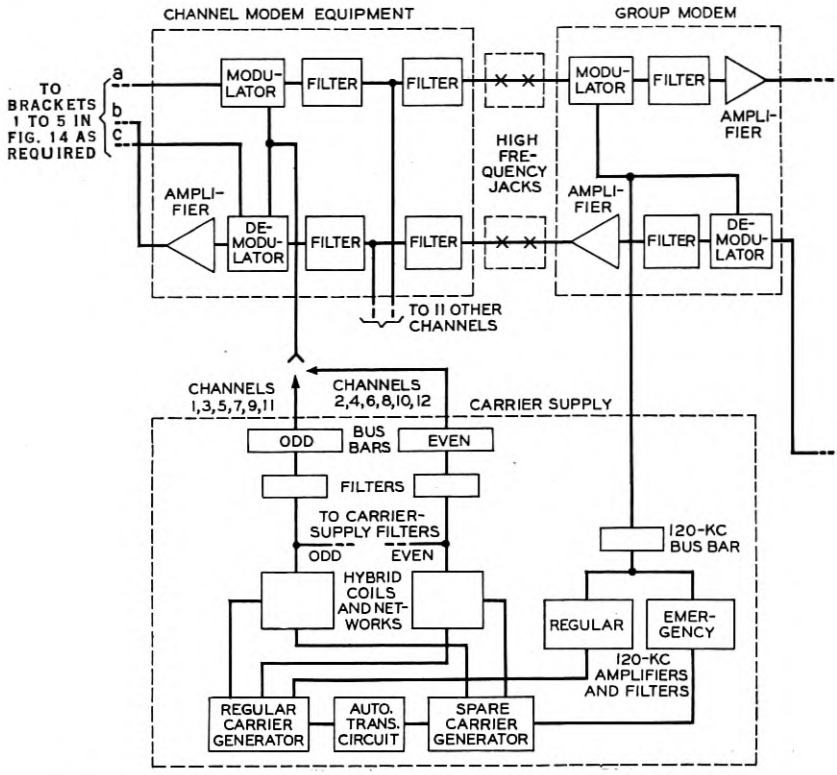

Fig. 9—Schematic showing the order of equipment at a terminal office.

connected to one set of leads for each type of power, and even circuits to another set.

The voice frequency sides of the channel modem units have been terminated in jacks which are located in a four-wire jack field mounted in a voice frequency patching bay. From this point the four-wire jack circuits are connected to the distributing frame for interconnection on a four-wire basis with other channel equipment, voice frequency repeater equipment, or terminating apparatus, as shown schematically in Fig. 14. Voice frequency patching bays shown in Fig. 10-B may be considered the equivalent of a secondary testboard.

Voice frequency transmission measuring apparatus has been mounted with the voice frequency patching bays.

Main Repeater Office Equipment

There are two types of installations at main repeater offices: (*a*) flat gain regulation only, and (*b*) both flat and twist gain regulation

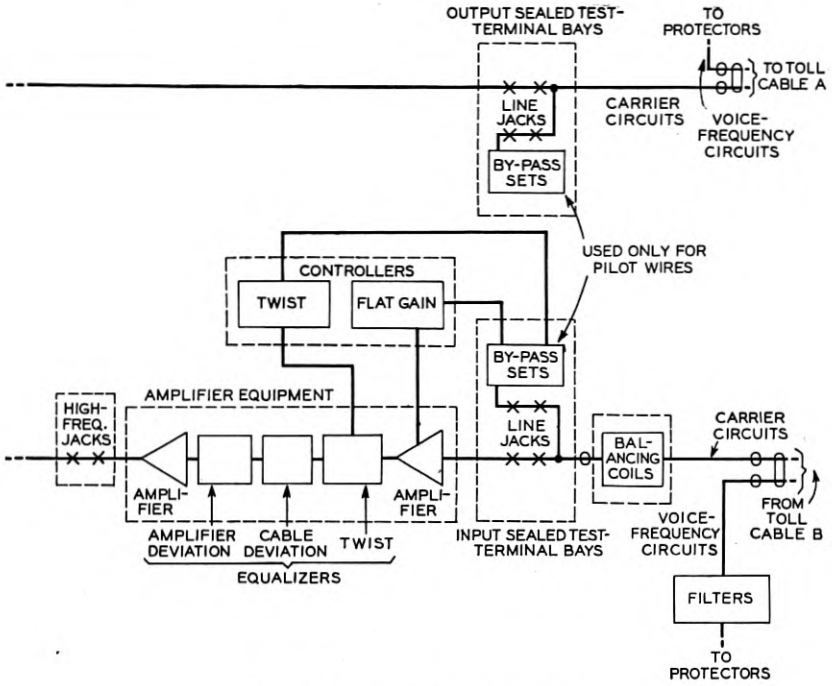

Fig. 9—Continued from page 570.

In general, these offices are attended regularly by maintenance forces. They serve in some cases as control or supervisory points for the auxiliary repeater stations. Since the installations at main stations have been made in existing voice frequency repeater offices, the available power plant is used to furnish filament and plate current for the amplifier equipment.

The input and output sealed test terminal bays with a high-frequency transmission measuring bay have been installed adjacent to each other to form a five-bay unit for testing and patching purposes.

Line amplifiers for each direction of transmission have been grouped together in adjacent relay rack bays. Each bay has a capacity of 20 amplifiers, except the first, in which is mounted the test amplifier associated with the high-frequency measuring system and 19 line amplifiers. Associated with the line amplifiers are the flat gain master controllers and power supply and cable balancing equipment. A schematic arrangement of circuits in a flat gain repeater office is shown in Fig. 15.

Repeater offices giving both flat and twist gain regulation have been

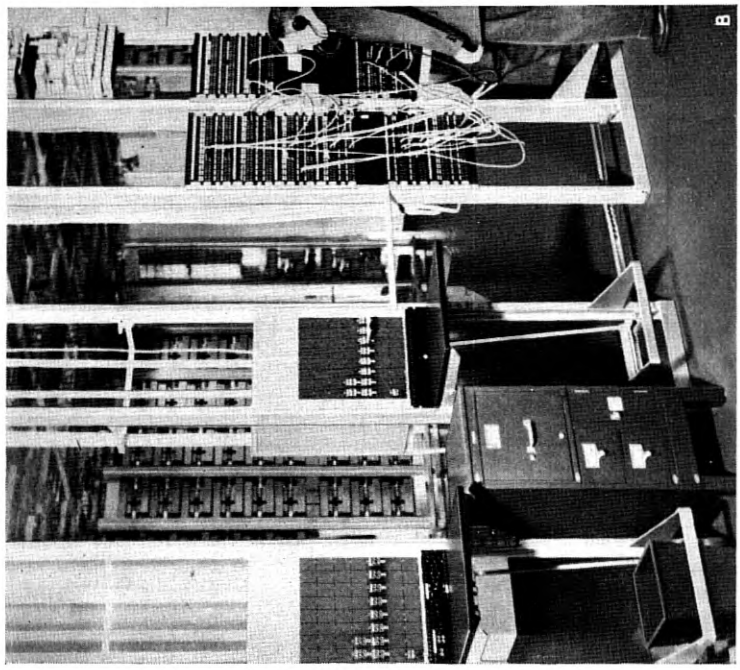

Fig. 10B—High frequency and voice frequency patching bays at New York, N. Y.

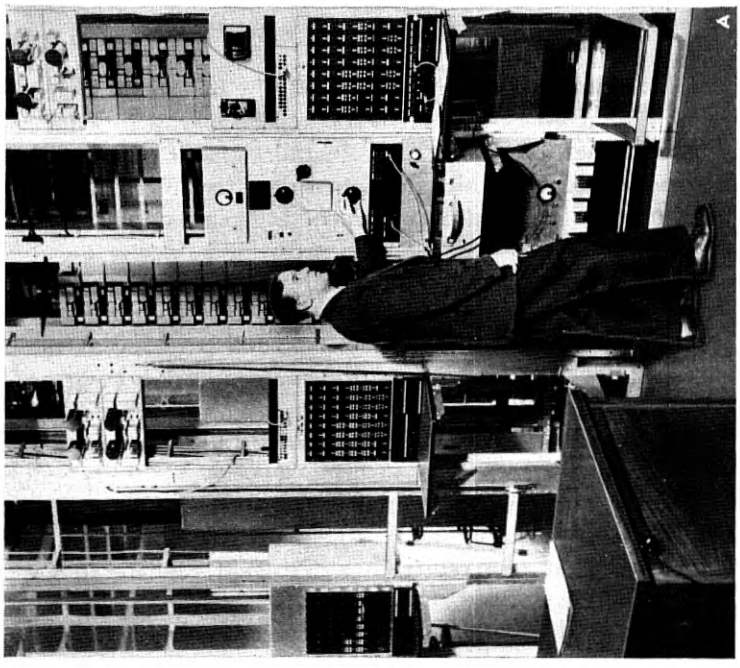

Fig. 10A—High frequency test bays at New York, N. Y.

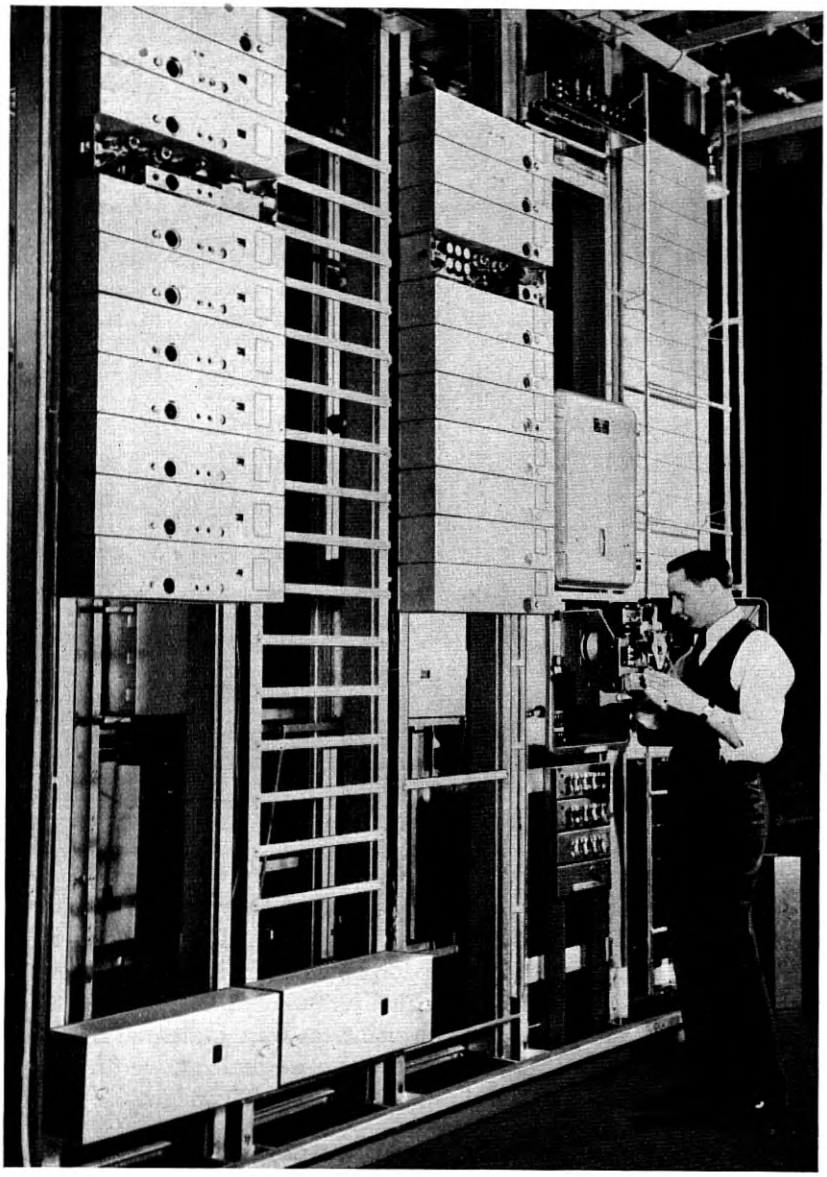

Fig. 11—Line and twist amplifier and controller equipment bays as installed at New York.

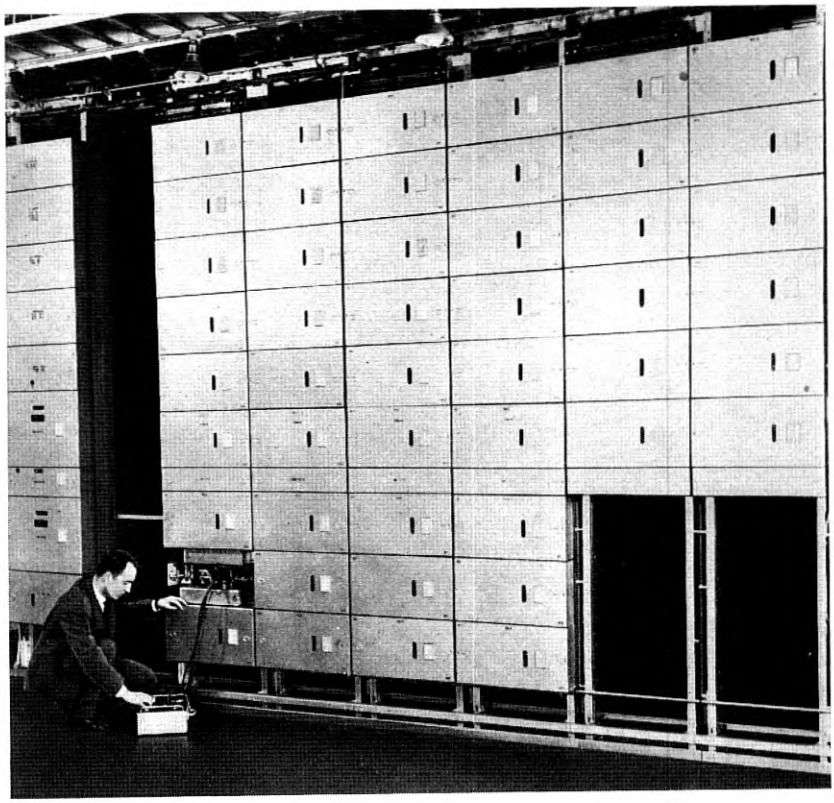

Fig. 12—Arrangement of carrier supply and channel modem equipment at New York, N. Y.

provided at intervals of about 100 miles. They differ from the flat gain regulating repeater office chiefly in that they include twist correction regulation [1] and its associated amplifiers. Provision was made at each twist gain regulating office for the installation of amplifier and cable deviation equalizers. The equalizers were actually connected to the circuits, however, only at such points as were indicated by lineup tests. To permit lineup without delay, spare equalizers were available at points where computations had indicated they might be needed. Amplifier deviation equalizers were required at South Bend, Toledo, Philadelphia, Richmond, and Greensboro on completed projects. It was not found necessary to install cable deviation equalizers at any point. Equalizer, flat gain, and twist gain amplifiers were installed in three-bay groups with the equalizer equipment occupying the center bay, and controller equipment in the fourth bay, similar to the ar-

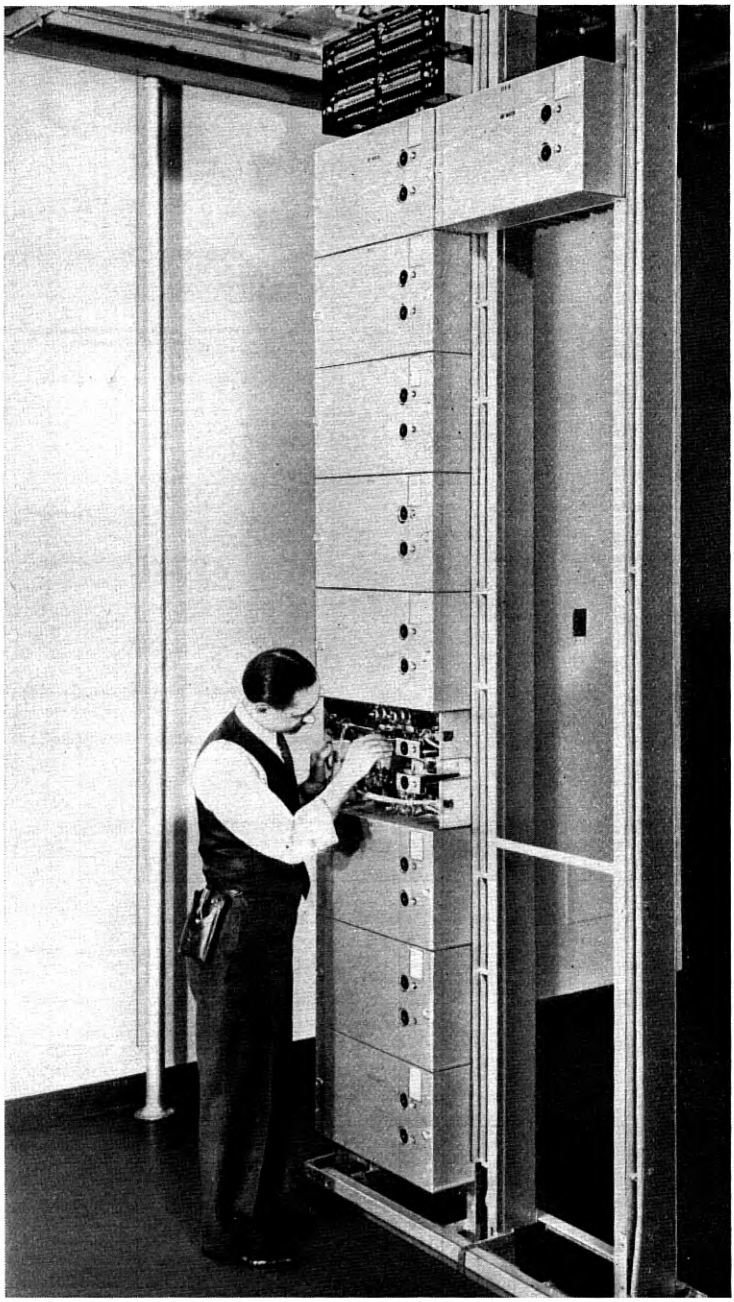

Fig. 13—Bay arrangement for group modem equipment as installed at New York.

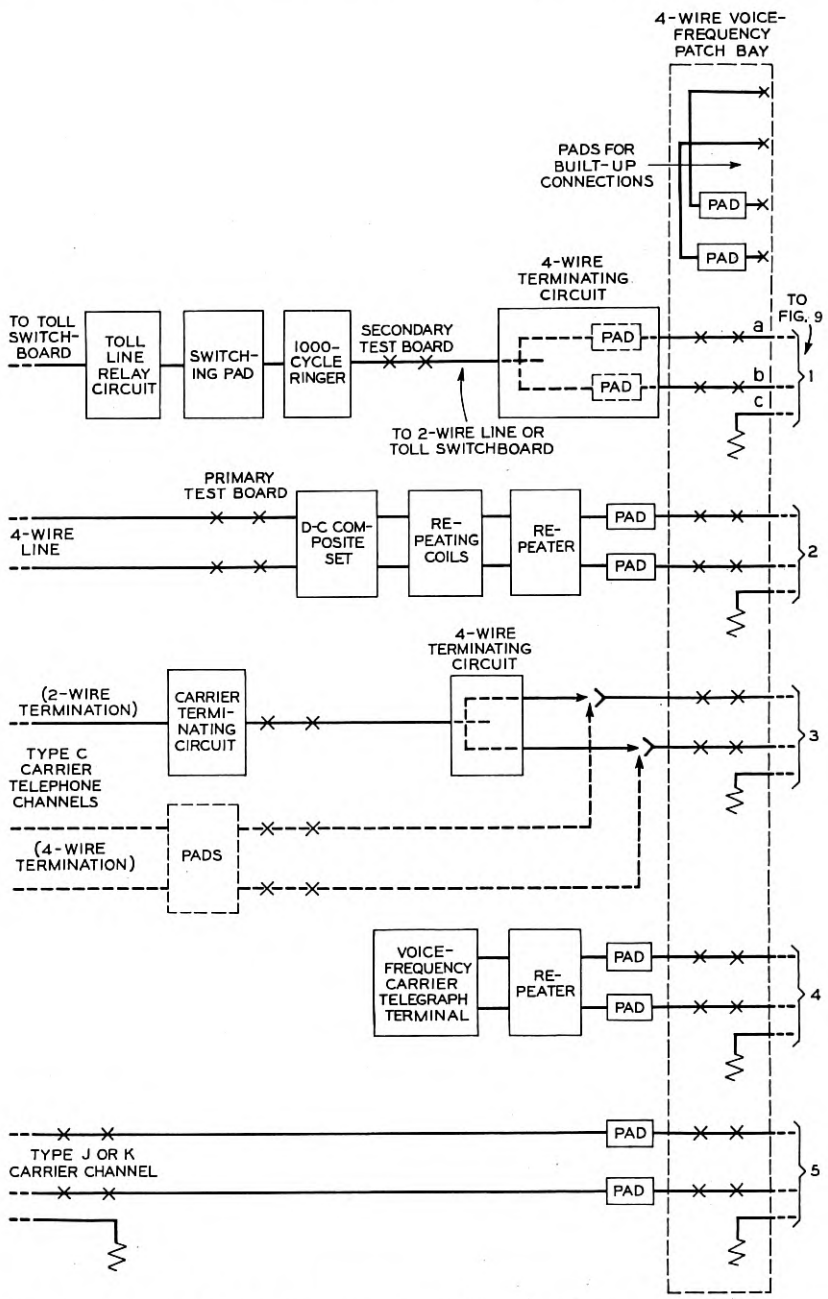

Fig. 14—Schematic showing the various circuit arrangements on the office side voice frequency patching bay.

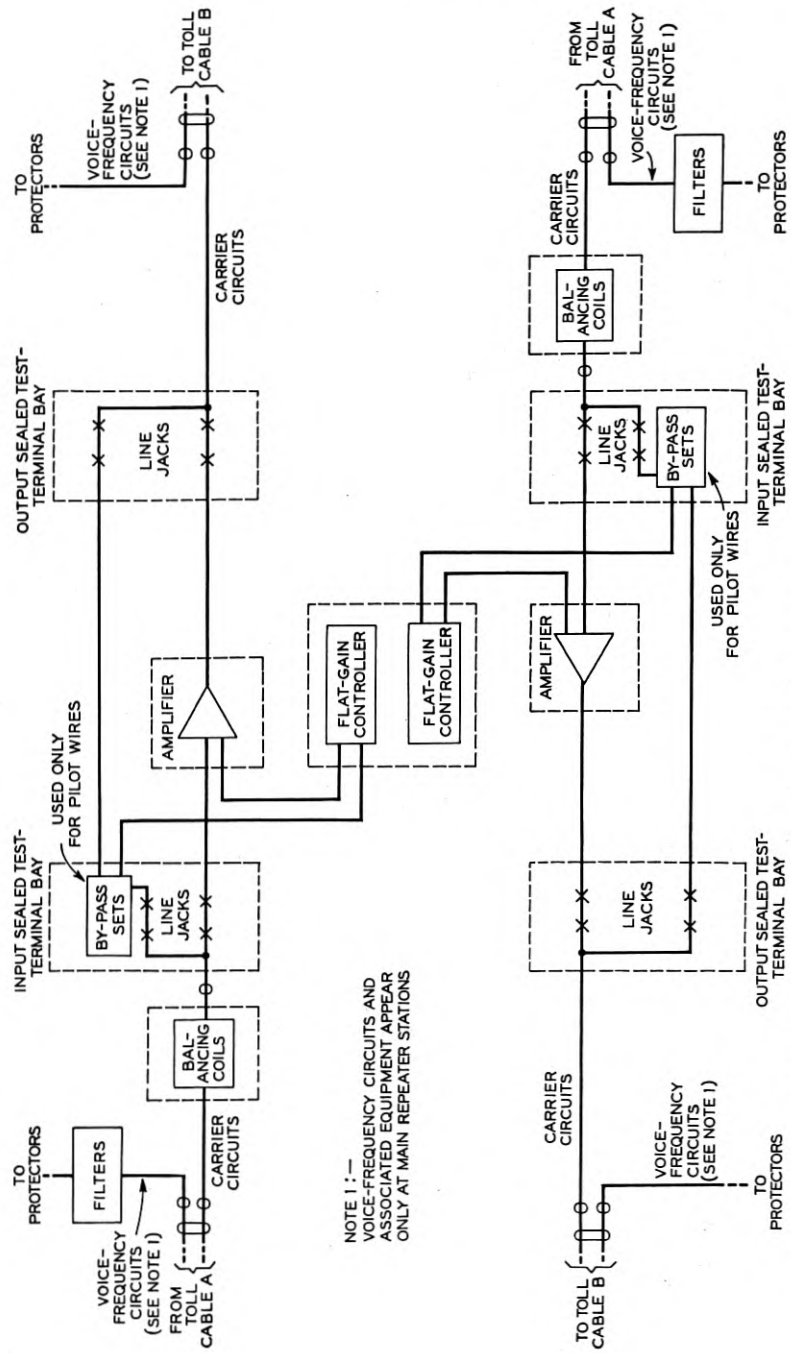

Fig. 15—Schematic showing the order of equipment at a flat gain repeater office.

rangement shown in Fig. 11. Each group of three bays mounts equipment for one direction of transmission for 17 systems. A schematic arrangement of equipment in a twist and flat gain regulating repeater office is shown in Fig. 16.

Auxiliary Repeater Station Equipment

Each auxiliary repeater station houses crosstalk balancing equipment, sealed test terminals, line amplifiers, pilot wire regulators, and a power plant. A typical floor plan arrangement of the equipment required in one of these stations for a maximum of 100 systems is shown in Fig. 17. The equipment arrangement for a 60-system route is practically the same, except that provision has been made for a smaller number of amplifiers and crosstalk balancing bays. A schematic arrangement of equipment circuits is shown in Fig. 15.

Four bays of sealed test terminals have been installed, one input and one output for each direction of transmission. Initially each unit contains carrier line and equipment jacks for testing or patching purposes for 40 carrier and eight miscellaneous circuits. In addition, miscellaneous auxiliary equipment is mounted in these bays.

Twenty-one amplifier panels for one direction of transmission may be mounted in a bay. Two bays are required for the flat gain master controllers and the associated controller power supply equipment. Figure 18-B shows amplifier, controller, and testing bays.

High-frequency testing apparatus consisting of a variable test oscillator and a portable transmission measuring set mounted in a mobile relay rack bay has been provided at each auxiliary station. This unit may be connected to the jacks in the sealed test terminals as required by means of patch cords.

Since auxiliary stations are designed to operate for considerable periods of time without attention, the power plant is of the automatic type.[7] It consists of a 70-cell, 152-volt storage battery which is continuously floated across regulated tube rectifiers fed from a commercial power supply. Figure 18-A shows a typical installation. Two rectifiers are provided initially, one which floats the battery, and the other which is connected automatically into the charging circuit in case of failure of the first unit or to increase the charging rate after a prolonged failure of the outside power. Arrangements are available in the power service cabinet to terminate leads from a portable emergency engine driven alternator set which may be set up outside the building. It is expected that the battery installed initially will be of sufficient size to provide a minimum of 24 hours reserve throughout its life, taking

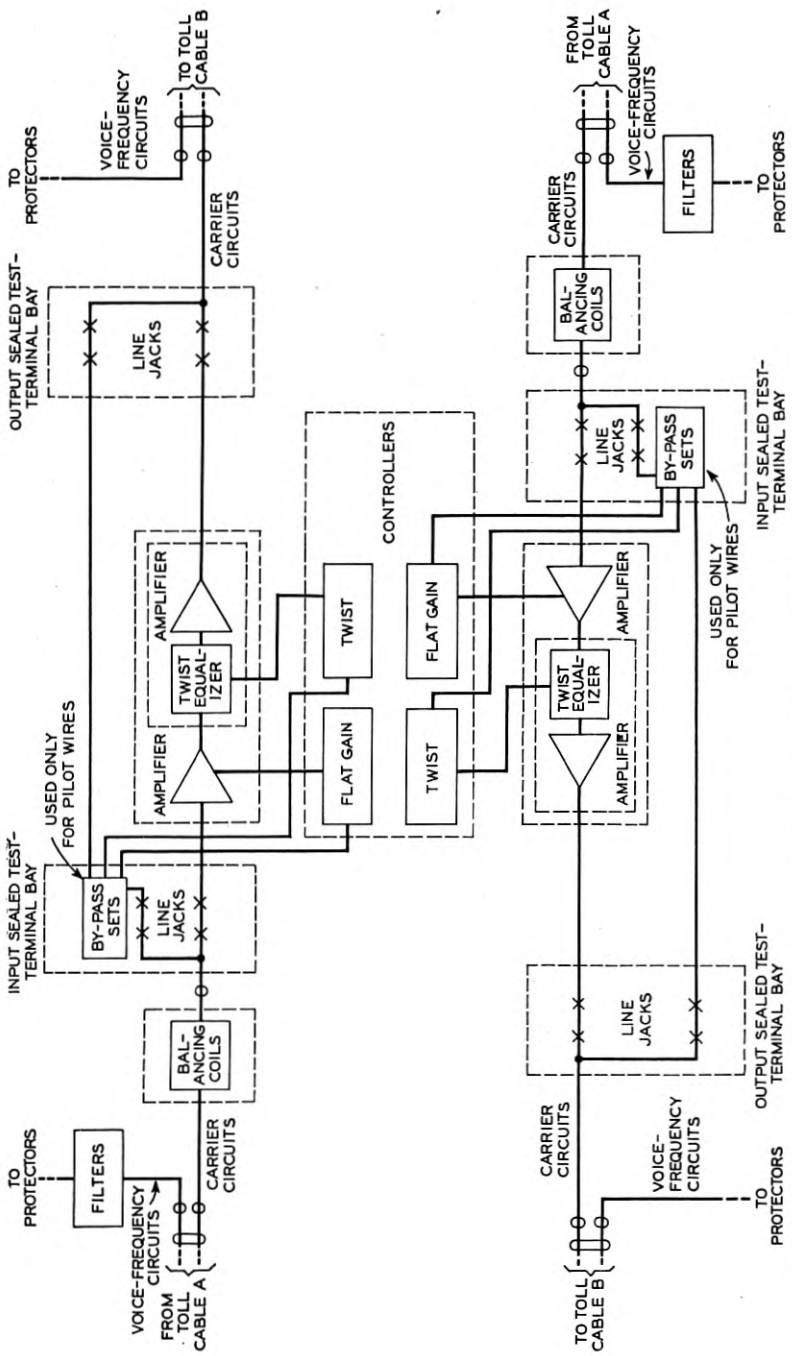

Fig. 16—Schematic showing the order of equipment at a twist and flat gain office.

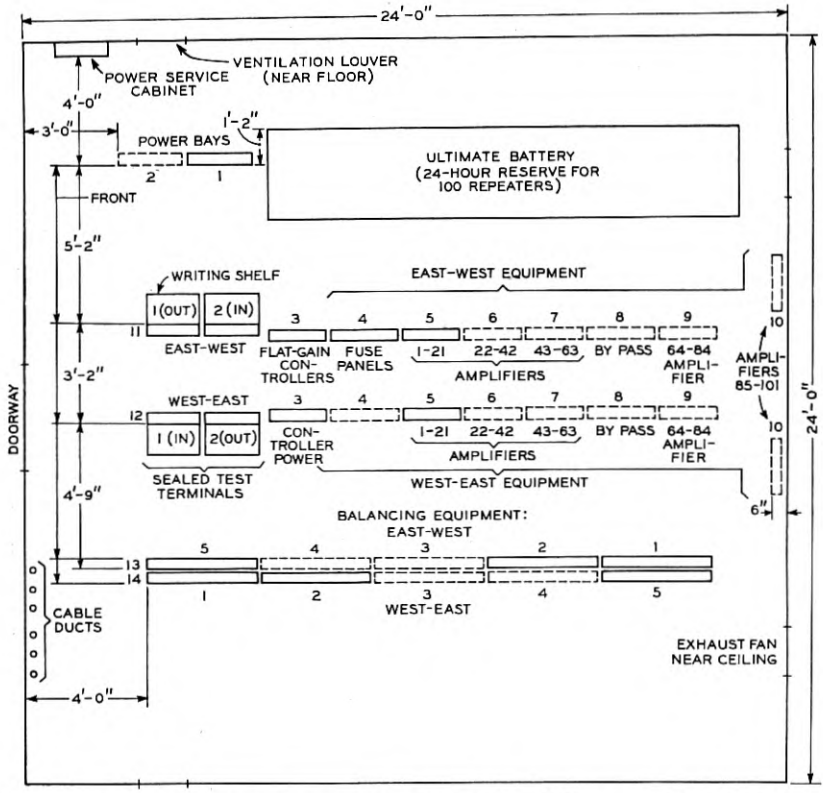

Fig. 17—Floor plan arrangement of equipment for 100 systems at an auxiliary repeater station.

into account the estimated growth of carrier amplifier equipment requirements for that period.

The entire voltage of the battery is used to furnish plate supply for the amplifier tubes. The 70-cell battery is arranged in seven groups of 10 cells each and taps are taken from each group to supply current for the heaters of the tubes of each amplifier. To prevent an uneven drain, amplifiers are connected across the battery in multiples of seven. In case the number of amplifiers installed is not an even multiple of seven, dummy load resistances are connected as required, in lieu of amplifiers to fill out the unequipped multiple.

The controller power supply bay contains apparatus for the 140-volt d-c pilot wire bridge supply and 55-volt, 60-cycle a-c supply. An emergency rotary converter, which automatically provides 110-volt, 60-cycle a-c supply during outside power failure, and operates intermittently from the battery, also is mounted in this bay.

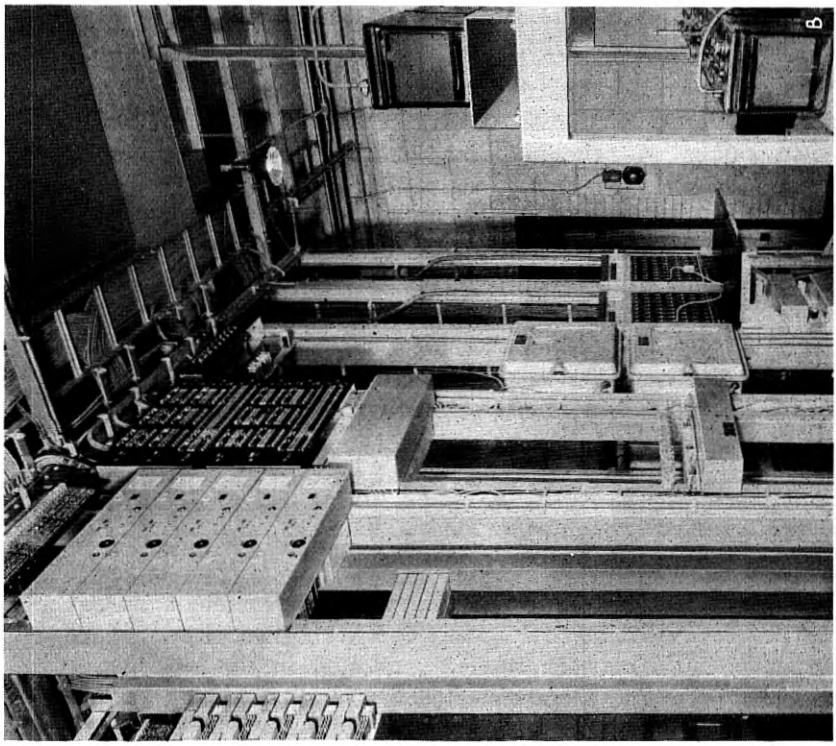

Fig. 18B—Amplifier, controller, and testboard equipment arrangement at an auxiliary repeater station.

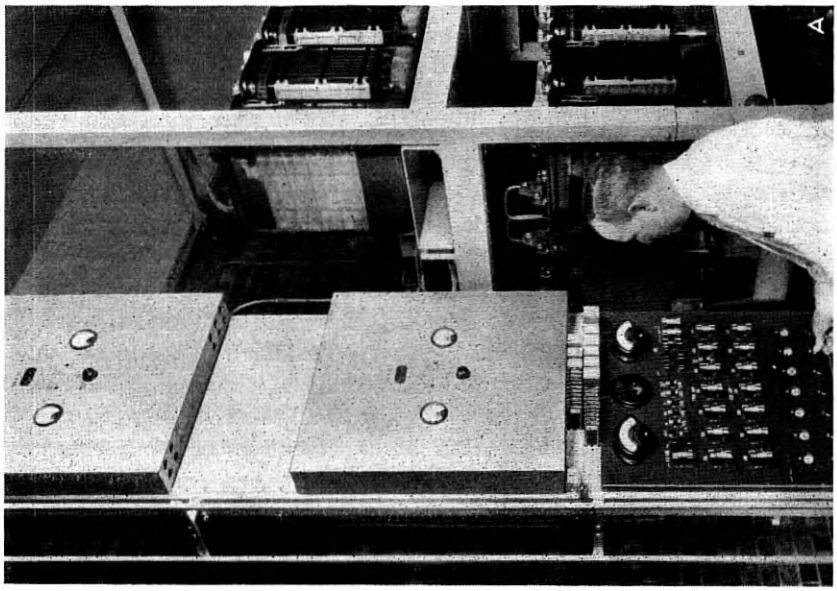

Fig. 18A—Auxiliary repeater station power plant.

Cabling Problems and Floor Plan Layouts

The floor plan arrangements for type K carrier equipment have been controlled to a considerable extent by transmission requirements, with consideration also being given to satisfactory operating and maintenance layouts both for the initial installation and for the future. The first consideration of any space which is to be used for carrier equipment is that the various units of equipment can be so located, with respect to each other, that the established maximum wiring lengths, as determined by transmission, operating, and economic requirements, will not be exceeded. For example, the length of shielded pair cable between the jacks in the input sealed test terminal bay and the input side of the line amplifier has been limited to about 50 feet and the potentiometer lead between the voice frequency patching bay and the channel modem units has been kept short in order not to limit the adjustment range of the potentiometer and a limit of 150 feet has been set. These limits were set in the design of the type K systems. Two types of cabling were installed in the transmission part of the carrier circuit; i.e., standard lead covered cable and shielded pair cable.

In cabling the carrier equipment the input leads were not run on the same cable racks with voice frequency cables. Due to the difference in transmission level between cabling connected to the input sides of carrier amplifiers and that connected to the output sides, these two groups of cabling have been segregated by running them over separate cable racks. The input leads were spaced not closer than two feet to any leads carrying interrupted direct current or power supply leads which might possibly carry high-frequency noise currents. Output leads were run on the same cable racks with voice frequency cabling where necessary, but were kept six or more inches away from possible disturbing leads such as those just mentioned. Cabling from the voice frequency side of the channel modem equipment was installed without greater precautions than are used when installing other voice frequency cabling. Crosstalk balancing bays were installed in any convenient location without special limitations in the lengths of lead covered cables between these and the input sealed test terminal bays.

All cable racks carrying the rubber covered shielded cables from the amplifier and group modem bays to the sealed test terminal and high-frequency patching bays were arranged so that these leads were run loosely and without sewing. This arrangement provides a ready means for switching cables for circuit layout purposes, particularly at terminal offices or at junctions of carrier cables.

In existing offices the high-frequency jack and testing equipment has been located as close as practicable to the existing toll testboard posi-

tions. Separate testboard lines have been established in several offices opening off, or convenient to, the main operating aisle in front of the toll testboards. The voice frequency patching jack bays at carrier terminals have been located, where practicable, near the secondary toll testboard equipment. These arrangements have been made in order to facilitate operating and maintenance, particularly during light load periods when a small force is on duty.

The amplifiers and group modems have been closely associated with the high-frequency patching bays and sealed test terminals in order to limit cabling lengths. Channel modems and carrier supply have been located convenient to the other equipment but within wiring limitations to the voice frequency patching bays.

The adequacy of all floor plan layouts, in providing for ultimate requirements, particularly at large terminal offices, was studied. This problem was given special consideration where it is expected that routes in addition to the initial one may be developed later for carrier operation. For example; at New York it was necessary to plan for the development of K carrier and other broad band facilities on four separate routes requiring a considerable amount of space for the necessary terminal equipment. The installation of carrier equipment at this office, therefore, has been made in space separate from the existing voice equipment. A floor plan arrangement of the equipment layout at New York is shown in Fig. 19.

Order Wires and Alarm Circuits

One or more auxiliary stations have been associated with an adjacent main or terminal office for maintenance control. Interoffice trunk and alarm equipment provide talking and signaling facilities between each auxiliary and main repeater station over a loaded cable pair. The various alarm signals are terminated in lamps which are mounted in the sealed test terminal bay at the controlling main office. These alarms are arranged to indicate such happenings as fire, open door, high-low battery voltage, main discharge fuse operation, a-c power failures, etc. When an alarm signal is received at the main station, it is rechecked and upon its reappearance an attendant may be dispatched at once or later to the auxiliary station involved, depending upon whether the signal is of major or minor importance.

The auxiliary stations are not always controlled by the nearest attended station. For example, in several cases it has been thought better to have them controlled by a station located in a small town rather than a city, because in cases of necessity an attendant should be able to drive to the auxiliary station in less time than required to drive

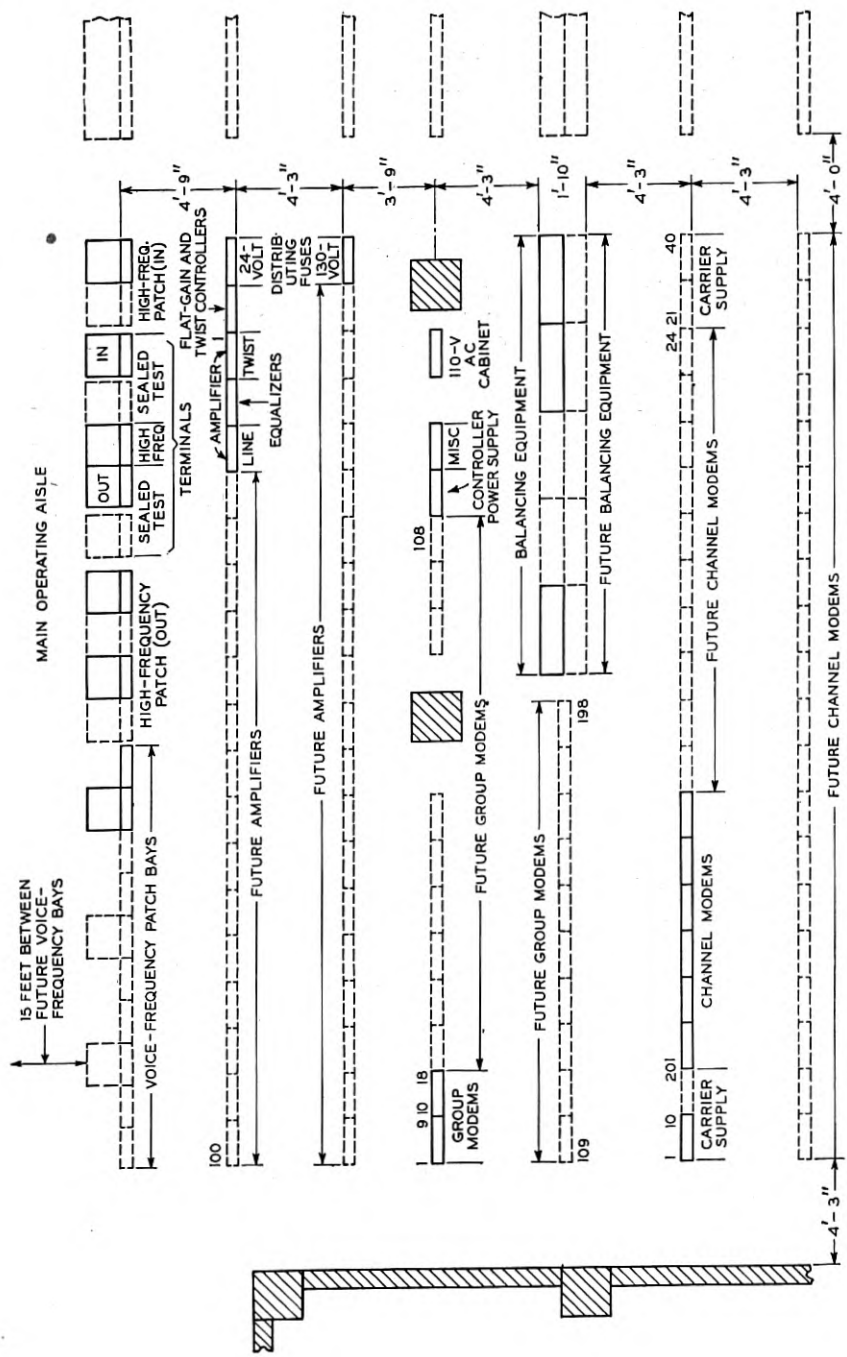

Fig. 19—Floor plan arrangement of equipment installed at New York, N. Y.

through a city area. In other cases certain main stations are not manned 24 hours per day and control and alarm circuits have not been terminated at such points. In one case a main repeater station has terminated in it the control leads from eight auxiliary stations. Four of these are connected through the partially attended main stations on either side.

Completion Tests and Overall System Adjustments

The usual completion tests were made on each unit of equipment after it was installed and on each cable pair between repeater stations after it was unloaded, in order to insure readiness of each item to be connected to form the overall carrier system. The gains of the repeaters were given a final adjustment by connecting each repeater input to the cable pair with which it was installed to work, then sending a predetermined amount of power at 28 kc into the cable pair at the adjacent office and adjusting the gain of the repeater until it delivered the desired output level. The flat gain master regulator was adjusted with respect to its pilot wire so that it would adjust the gain of the amplifier to maintain the desired output level at 28 kc at all cable temperatures.

The repeater sections beginning at one end of each twist regulating section were measured progressively at the output of each repeater. In each case transmission was checked at ten frequencies throughout the range from 12 to 60 kc. In this way a check was obtained to determine whether proper equalization was being provided at each line amplifier. It was necessary in some cases to change the type of equalizer provided because the transmission characteristics of specific cable pairs differed from the average which had been assumed in providing equalizers. Measurements were also made on the overall twist regulating section and transmission checked at ten frequencies throughout the 12 to 60 kc range, since the output of a perfectly corrected twist section would be the same at all frequencies within this band.

Overall measurements similar to these were made on the high-frequency line between terminal points and Fig. 20 shows the results for typical New York-Charlotte and New York-Washington systems. The most desirable characteristic would be a straight line and it will be noted that this curve differs materially from such an ideal. This difference is due to inadequate compensation by means of equalizers for small deviations from linearity in the individual line amplifiers. This lack of linearity in the overall high-frequency line has not materially affected the systems being operated at present but might be-

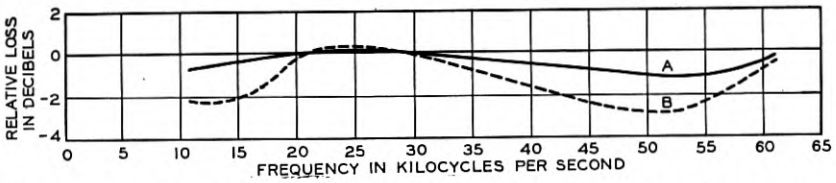

Fig. 20—Typical overall transmission frequency characteristic of high frequency line between New York and Charlotte, N. C. (B), and New York-Washington (A).

come objectionable on future long systems and it is planned to improve this characteristic by means of different equalizers.

The overall transmission frequency characteristic of a channel on a New York-Charlotte type K system is shown by Fig. 21. Measure-

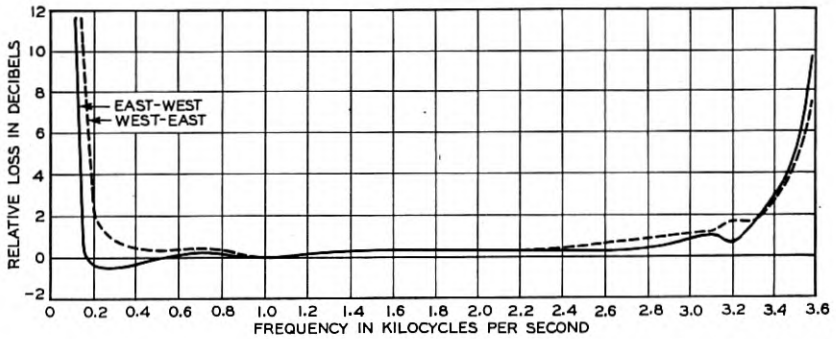

Fig. 21—Overall transmission frequency characteristic of a type K carrier channel between New York and Charlotte, N. C., as measured between two-wire voice frequency lines.

ments for this characteristic were made between the two-wire sides of hybrid coils connected to the two directions of carrier transmission of the channel concerned.

Use of Initial Systems

Telephone message circuits are being operated over most of the type K channels now available for use. In most cases the channels are used as parts of circuits which are longer than the carrier systems. For example, most of the 60 channels between New York and Charlotte, N. C., are used for circuits between New York and southern cities beyond Charlotte. Some of these circuits are obtained by connecting type K carrier channels at Charlotte to channels of type J open wire carrier systems which operate between Charlotte and West Palm Beach, Fla. Of the 204 channels available for use, only 21 are used as all

carrier message circuits between the system terminals. This is not necessarily typical of what the usage of type K channels will be, but is brought about by their relatively limited application to date and demonstrates the flexibility of the type K carrier circuit in fitting in with the other types of circuit facilities.

Conclusions

Although experience with type K systems in service is rather limited, they are providing circuits of excellent quality and performance. The band width of the individual channels slightly exceeds the original estimates. Such instabilities of transmission as have been experienced have, in most cases, been corrected before service was interrupted and have been caused largely by non-recurring troubles inherent to most installations of new equipment. In brief, the operating experience with type K cable carrier systems confirms the view that they are expected to become an important means of providing additional long distance telephone circuits over cable facilities.

References

1. C. W. Green and E. I. Green, "A Carrier Telephone System for Toll Cables," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
2. M. A. Weaver, R. S. Tucker and P. S. Darnell, "Crosstalk and Noise Features of Cable Carrier Telephone System," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
3. R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, "Cable Carrier Telephone Terminals," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
4. A. J. Aikens, "Suppressing Noise and Crosstalk on the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 7, March, 1939.
5. F. W. Amberg, "Crosstalk Poling for Cable Carrier System," *Bell Laboratories Record*, Vol. 17, No. 6, February, 1939.
6. P. W. Rounds, "A Longitudinal Noise Filter for the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 9, May, 1939.
7. H. H. Spencer, "Power Plant for Broadband Repeater Stations," *Bell Laboratories Record*, Vol. 17, No. 8, April, 1939.
8. L. Hochgraf, "Crosstalk Balancing for the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 6, February, 1939.

The Toronto-Barrie Toll Cable *

By M. J. AYKROYD and D. G. GEIGER

General

DURING 1937 a 60-mile toll cable was completed between Toronto and Barrie which, in several respects, is unique. Among the interesting features in the design and construction of this toll cable were the use of non-quadded exchange cable and loading, a 60-mile repeater spacing, planning for future carrier operation, and extended pole spacings.

Prior to the installation of this toll cable, the territory to the north and northwest of Toronto was served by three open-wire pole lines. Figure 1 shows these lines and the territory served by them. The Toronto-Owen Sound lead entering Toronto through a 7-mile entrance cable was poorly located in towns and on highways, and was paralleled by power lines which caused considerable noise on the longer circuits. The Toronto-Collingwood and Toronto-Barrie lines, which were common for some distance north of Toronto on a 6- and 5-arm lead, entered Toronto through an 11-mile entrance cable which had been in place about four years, and contained a number of spare conductors due to its having been designed for two additional lines.

It was realized that, if open wire were to be continued, circuit growth would require a new line arranged for carrier operation and a general rebuilding, rerouting and retransposing for carrier operation of the existing lines. In addition, carrier operation would necessitate expensive carrier loading of the entrance cables at Toronto, and the length of these entrance cables would limit the length of the carrier circuits for operation without intermediate repeater stations.

Studies Prior to Construction

With large expenditures foreseen for the continuance of open wire, it was only natural that a study should be made of the possibility of the use of a toll cable on a basic route and the use of as much as possible of the existing lines as feeders to the cable. Cost studies on an annual charge basis for a twenty-year period of open wire with superimposed 3-channel carrier systems, and for a 2-wire 19-gauge quadded cable

* The unusual solution of a difficult toll cable problem which is described in this paper will be of interest because of its novelty rather than because of any expected general application of this type of construction to toll cable routes.

with H88-50 loading with open wire or cable feeders, depending on the length and numbers of feeder circuits, indicated the cable plan to be best. In addition to the indicated money savings of the cable plan over the period of the cost studies, other indicated advantages in the toll cable plan were improved service continuity (the southerly section of the territory under study is one of heavy sleet conditions) and reduced noise from power induction.

The quadded cable plan, however, had one disadvantage in that it required a repeater station approximately 45 miles north of Toronto, in a territory remote from any town or village with unfavorable living conditions and subject to isolation during winter snow storms. The nearest feasible location to the ideal, at Cookstown, involved such an increase in length of cable and added expenditure that the cost advantage changed to the open-wire plan. Also, the use of B88-50 loading with a repeater spacing of 50 miles appeared to offer no advantage in that the additional cost of loading became an important factor.

These difficulties in the use of the standardized type of toll cable led to a review of the possibility of employing some combination of conductor and loading which would permit a 60-mile repeater spacing, thus eliminating any need for an intermediate repeater station between Toronto and Barrie. If such a cable were to have the same unit attenuation as 19-gauge H88-50 cable, then it must have considerably improved crosstalk and return loss characteristics. On the other hand, if a cable could be obtained with crosstalk and return loss characteristics about equal to that of 19H88-50 cable, it must have an attentuation of about ¾ that of 19H88-50 cable.

Of the standard types of cable and loading, 19-gauge non-quadded exchange cable having a capacity of about 0.083 mf. per mile with B-135 loading appeared to have an attenuation of about the value required to meet the second of the two requirements noted above. It was estimated that such a cable would have the following transmission characteristics:

1000-cycle attenuation at 68° F.	0.26 db per mile
Passive singing point at repeater exceeded by 72% of circuits	25 db
Maximum crosstalk gain	14 db
Overall active balance [1]	6.0 db
Overall circuit loss 8 db (PO-TC) with 4 db pad at PO.	

These assumed limits required a 72 per cent return loss of 26 db or better at the critical frequency which was expected to be about 2600 cycles, and a 1 per cent maximum near-end crosstalk of 74.5 db. Based

[1] Computed by summation of the 72 per cent singing points at individual repeaters with a 5 db end path.

on these values and limits, and assuming that Toronto would be the only gain switching center directly involved, a study was made of the transmission possibilities for each group of circuits that was expected to be routed through the cable. This study indicated that, provided the return loss and crosstalk values required of the cable by the assumed singing points and crosstalk gain could be met, all circuits could be 2-wire between Toronto and Barrie with some transmission margin and also that this type of cable could be extended at least another 20 miles to Orillia.

A cost study, assuming a 101-pair cable, of this type, indicated that while, due to the additional loading costs of the closer loading spacings, the cable costs were very nearly the same as for a quadded 19-gauge H88-50 cable, the considerably reduced repeater and repeater station costs made this plan appreciably less costly than any open wire plan. The elimination of any intermediate repeater station removed the repeater station difficulties of the quadded cable plans.

As no installation of such a length of this type of cable had been made, some confirmation of the estimated values for the transmission study, and particularly of the return loss and crosstalk, was considered necessary. An 8-mile H-44 loaded 19-gauge exchange cable, which had just been erected near Toronto, was chosen for study. Near-end crosstalk measured on 286 combinations of pairs indicated 99 per cent better than 81 db with an average of 91.7 db which, when modified for impedance and length differences, indicated 99 per cent better than 72.5 db, and an average of 83.2 db for the proposed cable. While these values were somewhat poorer than required, the size of the sample and one or two other factors indicated that the proposed cable could be erected to meet the crosstalk requirements. However, to obtain as much crosstalk margin as possible, arrangements were made for the manufacturer to use 6 lengths of twist, alternating 3 in each layer, rather than the 4 twists which had previously been used for this type of cable. It is felt that the excellent crosstalk results obtained as outlined in more detail later are in large part due to this feature.

Singing measurements on 10 pairs averaged 19.6 db. It was evident from impedance frequency measurements that these singing points could be raised to the desired value of 25 db by some modification in the networks. Accordingly an adjustable precision type network was developed.

Also, four 1500-foot lengths of the proposed type of cable were obtained from the manufacturer and tested for mutual capacitance of pairs and capacitance unbalance between pairs. On statistical analysis, these tests indicated a probable average near-end crosstalk of 79 db

and that for return loss 63 per cent of the circuits would be better than 27.0 db or 72 per cent would be better than 26 db at 2600 cycles, provided the following features were incorporated:

(*a*) Manufacture of complete length of cable in one continuous production with reasonably careful control of variables.
(*b*) Capacity equalization splicing at the mid-point of each 3000-foot loading section.
(*c*) Reel lengths be assigned as to location on the basis of average reel length capacity.

On the basis of these preliminary studies, it was decided to proceed with the cable plan, using the B-135 standard 19-gauge exchange cable. Figure 2 shows the plant layout finally adopted for the cable and its feeders.

At the Toronto end it was essential to use pairs in a recently placed 19- and 16-gauge quadded toll entrance cable (mutual capacity .062 mf. per mile) about eleven miles long in order to keep the cost of the cable to a minimum. This appeared feasible, using the same type loading coils as in the main cable, if the loading spacings were extended to provide the same loading section capacity as in the main non-quadded cable, and if the cable were sufficiently well respliced to break up the side-to-side (within-quad) adjacencies so that the crosstalk coupling would be comparable to that obtained in the main (non-quadded) cable.

Route

It was necessary to select the shortest practicable route passing as close as possible to the places to be served (see Fig. 2). The route selected is, for the most part, on a road which lies about midway between the main highways serving the territory north of Toronto. It is expected that the location chosen will be reasonably free from highway changes. Also, for the portion of the route south of Aurora, an existing open wire pole line was suitable for supporting the cable on long span construction.

At one point three miles of swamp covered with bush intervened on the direct route, the avoidance of which meant an increase in expenditure for right-of-way, as well as lengthening of the cable. It was decided to go straight through the swamp, using swamp fixtures, as shown in Fig. 3. An interesting sidelight on securing the route through the swamp was the fact that an original road right-of-way was shown on the map. On searching the records the surveyor found the original survey notes made in 1860 and eventually confirmed the location by finding some old pottery which, according to the records, had been

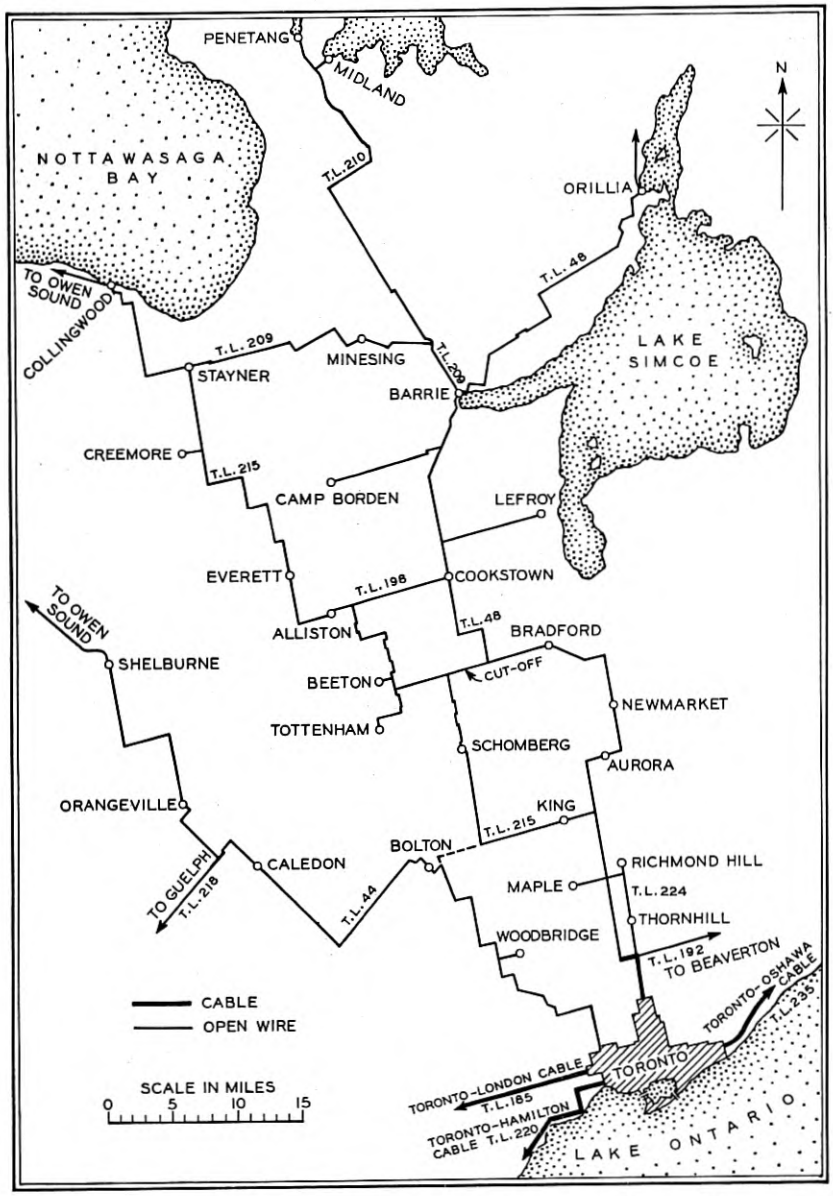

Fig. 1—Toll lines before construction of the Toronto-Barrie cable.

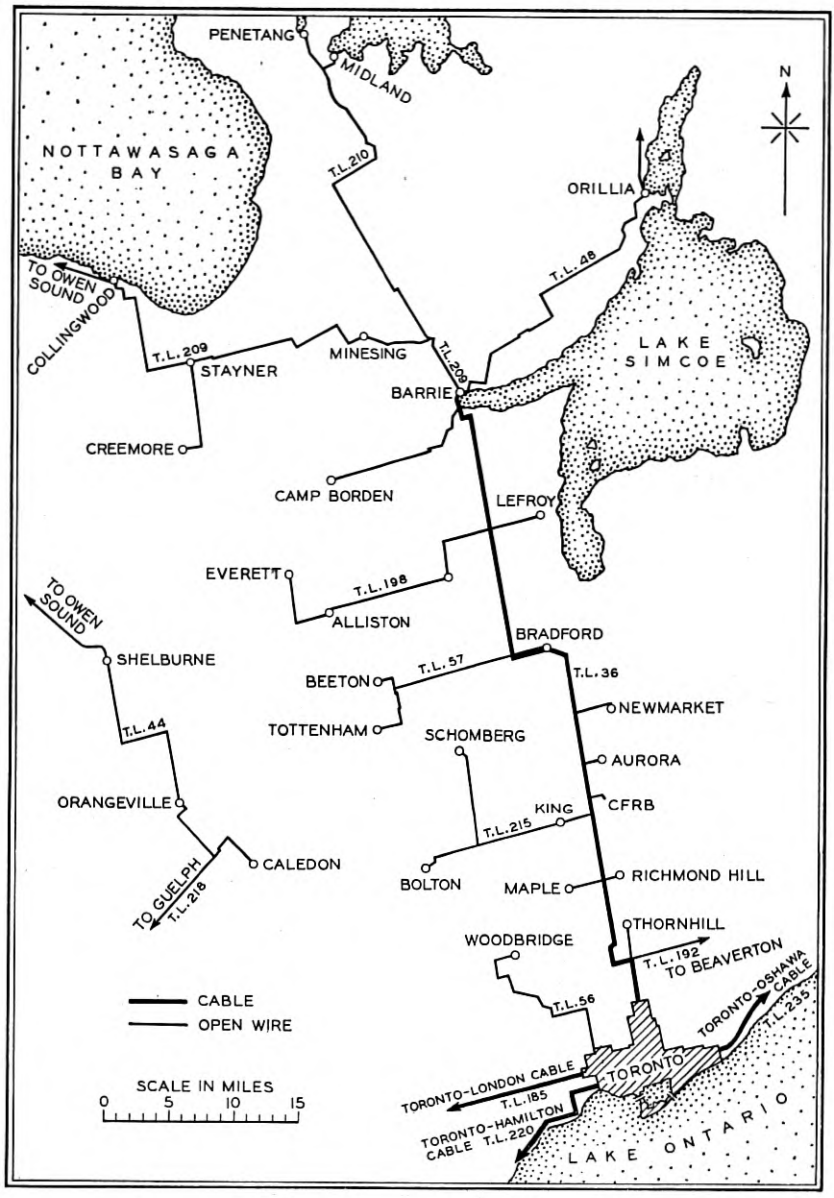

Fig. 2—Toronto-Barrie cable.

Fig. 3—Swamp construction.

buried at the road intersections. During the late summer and the autumn, when the swamp had dried out, a 60-foot right-of-way was cleared. As the soil was still moist and soft, the brush and small trees were uprooted by a tractor, a method of clearing which proved quick and economical. The swamp fixtures were placed before the ground froze, and the cable during the winter.

Pole Line

The cable was erected on the existing pole line between Toronto and Aurora. The size of the cable erected on 10 M. strand permitted the removal of every other pole in the old line, with a resultant 185-foot average spacing. As is shown in Fig. 4, where it was necessary to change an existing pole, a new pole was placed and fastened to the old pole by means of stub reinforcing bands, thus eliminating the expense of transferring the open wire. Upon the release of the open wire by transfer of circuits to the cable, the wire and old poles were removed.

The new section of pole line was erected on a 200-foot spacing, with occasional spans up to 250 feet, as shown in Fig. 5. This increased pole spacing was also expected to reduce ring cutting and bowing.

Construction Details and Tests

At a number of points open wire loops connected directly to the cable. At these junctions there were installed open space protectors having a lower breakdown than the cable pairs, and connected between the open wires and the cable sheath; also a few spans from the junction, 1000-volt protectors were connected between the open wires and driven grounds. This arrangement was more economical than the use of protection cable. The cable has gone through two complete lightning seasons without any failures or even permanent protector operations due to lightning.

In so far as manufacturing and storage facilities permitted, the reel lengths of the cable were assigned to their locations on the basis of obtaining as close an average loading section capacitance to the nominal value of 0.085 mfd per mile as was feasible. All reel lengths for the aerial sections were manufactured 1508 feet long, this length being sufficient to permit the assignment of a reel at any point in the line. Particular care was taken in this respect towards the ends of the cable where departures from the average would have the greatest effect on the return loss. To ensure proper assignment of reels, a route map was made up to scale with the manufacturer's reel number shown in its proper location.

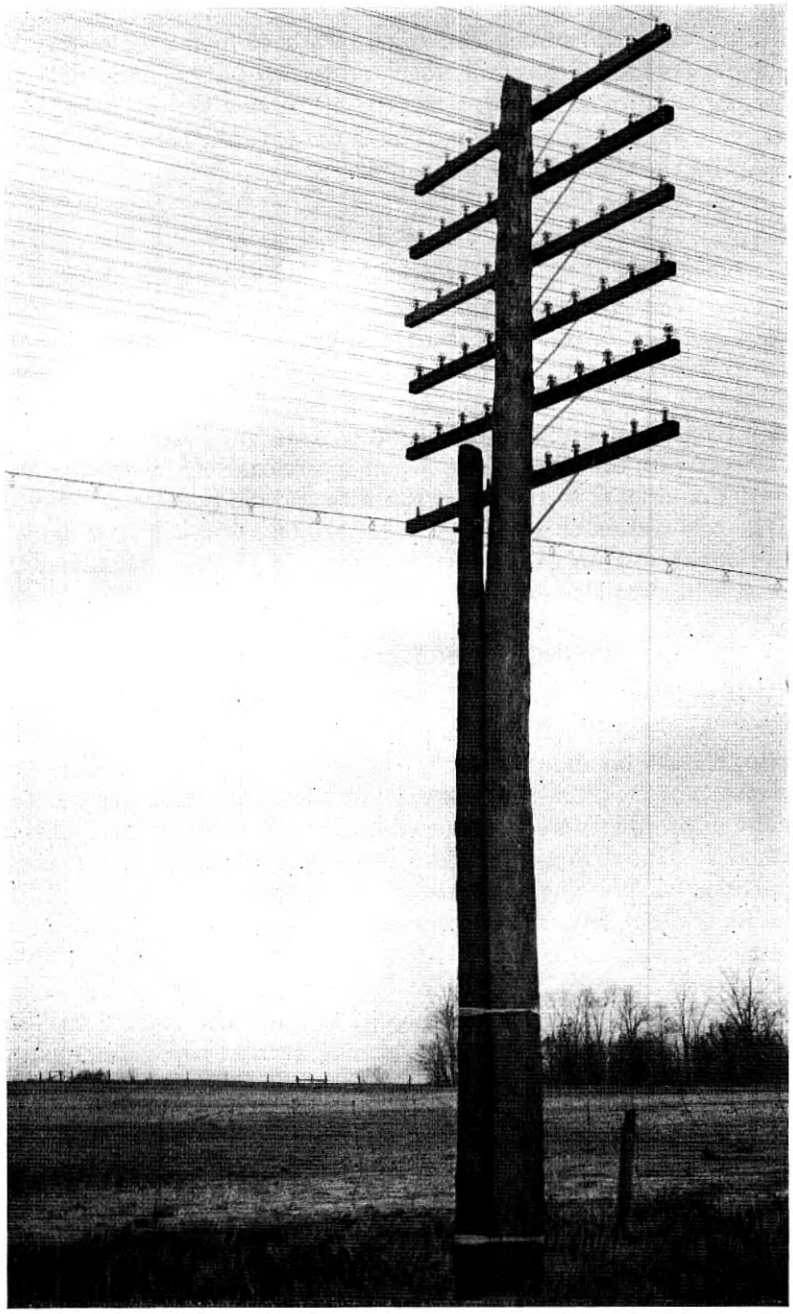

Fig. 4—Replacement of old pole with new creosoted pine pole.

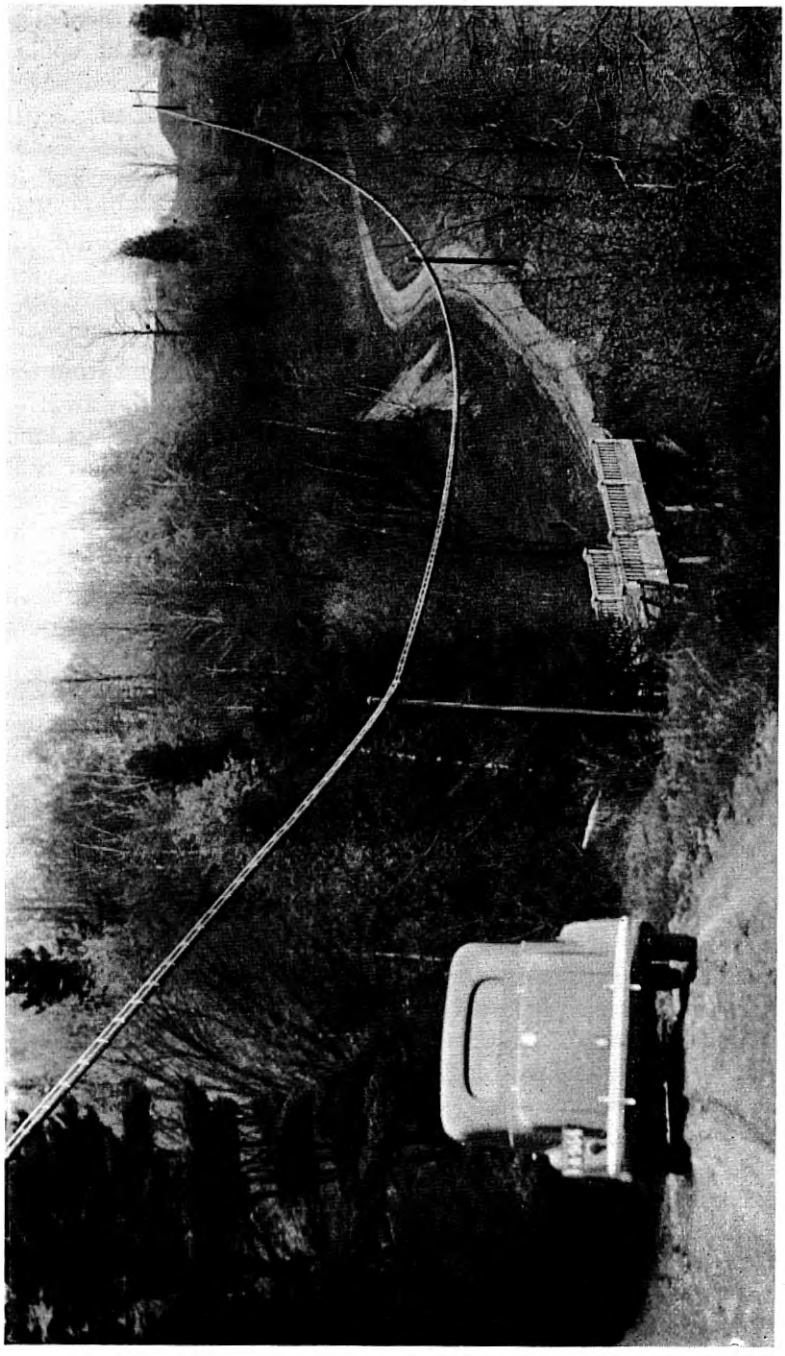

Fig. 5—Long span cable construction.

In addition, at the mid splice of each loading section, a test splice was made to equalize the capacity deviations. For these splices special linen boarding strips were used, each with 40 holes designated by a capacity ranging from about 15 per cent below to about 15 per cent above the expected average capacity of 1500 feet of cable. Small inexpensive capacity meters were used, and each pair was placed in the hole in the boarding strip corresponding to its capacity. The pairs were then spliced high to low capacity. This method did not require special testers, and substantially reduced splicing manhours.

Upon the completion of the splicing in each section, the section capacity of each pair was measured and recorded, and from this was determined the root mean square of the capacity deviations from the average capacity. These deviations combined with the deviation of the loading section average capacitances, loading coil spacings, and loading coil inductances, gave an irregularity function of 2 per cent which is almost identical with that for 19-gauge B88-50 cable. From this irregularity [2] function a 63 per cent return loss frequency curve was obtained, which is shown as curve '*A*' in Fig. 8.

When 15 miles of the cable had been completed south from Barrie, a 100-pair cross-connecting box was temporarily spliced in so that data could be obtained as a further check on the design estimates.

For crosstalk tests each pair was terminated at the box in a 1700 ohm resistance, and measurements were made of all pair combinations (approximately 5000). For these measurements a 15A oscillator and 2A Noise Measuring Sets were used, thereby very materially reducing the manhours required as compared with the labour that would have been required had crosstalk measuring sets been used. Analysis of these tests indicated 99 per cent of combinations better than 76.0 db, an average of 86.6 db and 99.5 per cent meeting the required 74.5 db of the preliminary studies.

For attenuation measurements, the pairs were looped back at the cross-connecting box. In order to obtain a value of the attenuation at a known temperature, a complete set of measurements was made at about 6 o'clock in the morning after the resistance of one of the pairs had been found to have ceased dropping due to temperature change and the outside temperature at the Barrie office had been very nearly constant for about one-half hour. During the time the attenuation measurements were being taken, air temperatures were measured at four places along the 15-mile length of cable. From these tests the average 1000 cycle attenuation at 62° F. was found to be 0.26 db per

[2] See "Irregularities in Loaded Telephone Circuits," George Crisson, *Bell System Technical Journal*, October 1925.

mile. In Fig. 6 is shown the mean of the attenuations of three pairs plotted against frequency. On one pair the measurements were made at frequencies up to 6500 cycles, from which cut-off was determined to take place at 4000 cycles. Assuming a 60-mile circuit, the frequency at which the attenuation is about 10 db greater than 1000 cycles, is about 3500 cycles.

Before the return loss tests were made, impedance-frequency measurements were taken on two of the balancing networks for each of the three adjustments provided, and on a representative number of cable pairs, to determine the optimum network adjustment. The resistance component of the impedance for one of the networks and one cable

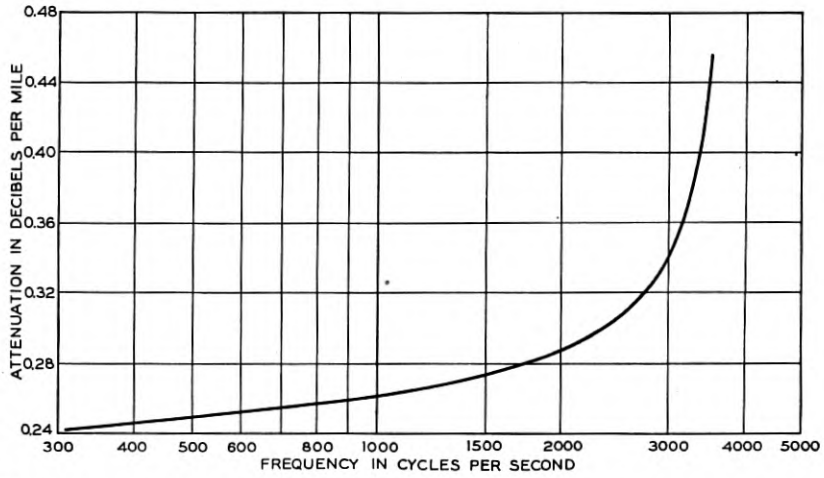

Fig. 6—Attenuation-frequency characteristic; mean of measurements on three pairs.

pair is shown in Fig. 7 (the two networks were found to be identical). From these tests the optimum network adjustment was determined to be that corresponding to a cable capacitance of 0.088 mfd per mile, which adjustment was then used for the return loss measurements.

As it was desired to obtain the singing point to be expected under operating conditions, the return loss measurements were made with the building-out condenser on the return loss set adjusted for optimum return loss at 2600, 2700 and 2800 cycles. The results of these measurements are given in Fig. 8 for comparison with the computed curve '*A*' mentioned previously. The improvement at the higher frequencies of the actual over the computed values is due almost entirely to the method employed in making the tests, and indicates the advantage to be derived from individual adjustment of each circuit.

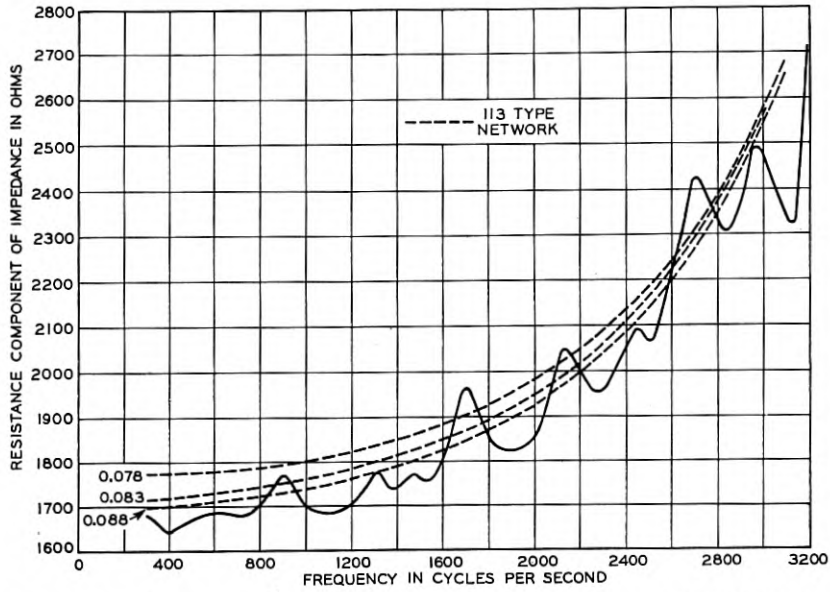

FIG. 7—Resistance component of impedance. 30.8 mile circuit, 19 CNB–B135. Termination, 113 type network.

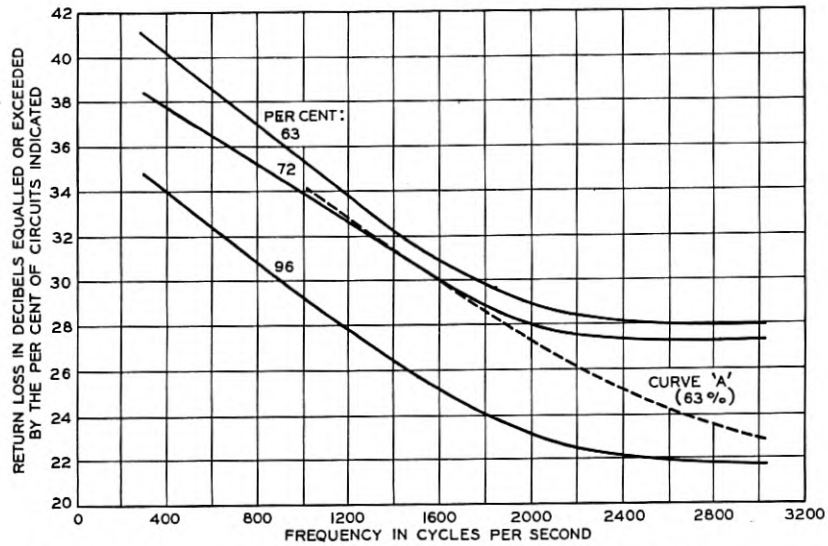

Fig. 8—Return loss—frequency characteristics. Measured on 61 circuits 30.8 miles long. Building-out condenser adjusted on each circuit for optimum return loss in the frequency range 2600, 2700 and 2800 cycles. Curve '*A*' is the 63 per cent return loss computed from attenuation and irregularity function measured on cable.

These return loss measurements were made on pairs looped back at the cross-connecting box and terminated at Barrie in one of the networks.

Completion Tests

Upon completion of the cable, further overall tests were made. Particular attention was paid to those tests made from the Toronto end, to determine the effects of the use of the reloaded and respliced toll entrance cable.

Attenuation measurements at 1000-cycles were found to agree closely but, due to the effects of the toll entrance cable at Toronto,

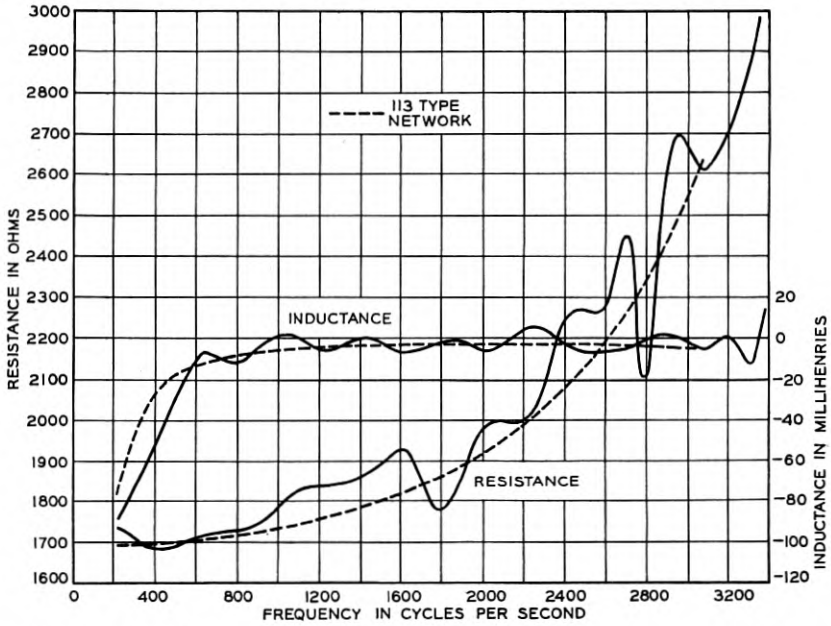

FIG. 9—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 10.8 miles 16-ga., 0.062 mf. per mile, 0.0483 mf., 32.8^{w} per load section; 48.7 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^{w} per load section.

not to lend themselves to such rigorous analysis as those previously made.

To show one of the effects of the Toronto toll entrance cable, Figs. 9, 10, and 11, showing the resistance and inductance components of the impedance measured at Toronto, are included. These indicate that the important departure from the network characteristic for these pairs occurs in the inductance component at the lower frequencies. This departure is probably due to the difference in the loading section

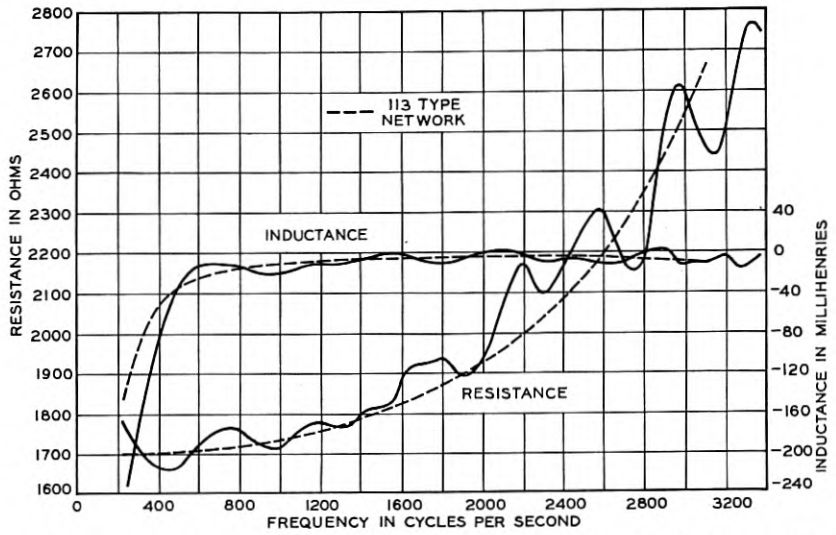

Fig. 10—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 9.5 miles 19-ga., 0.062 mf. per mile, 0.0483 mf., 66.8^{w} per load section; 1.3 miles 16-ga., 0.062 mf. per mile, 0.0483 mf., 32.8^{w} per load section; 48.7 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^{w} per load section.

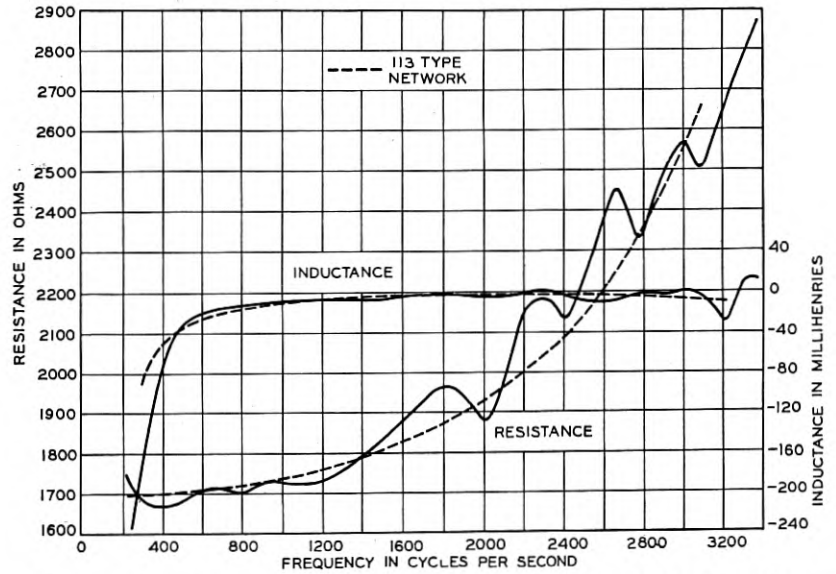

Fig. 11—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 9.5 miles 19-ga., 0.062 mf. per mile, 0.0483 mf., 66.8^{w} per load section; 50.2 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^{w} per load section.

resistance from that of the main cable. (The geographical spacing on the quadded 0.062 mf. cable was 4100 feet as compared to 3000 feet on the non-quadded cable.)

Since representative return loss data had already been obtained for circuits under working conditions (Fig. 8), the completion return loss measurements were made for the network building-out capacity conditions assumed for the theoretical return loss characteristic (curve '*A*,' Fig. 8). The results thus obtained are shown in Fig. 12 for Barrie and Fig. 13 for Toronto. In Fig. 12, the theoretical curve is shown for

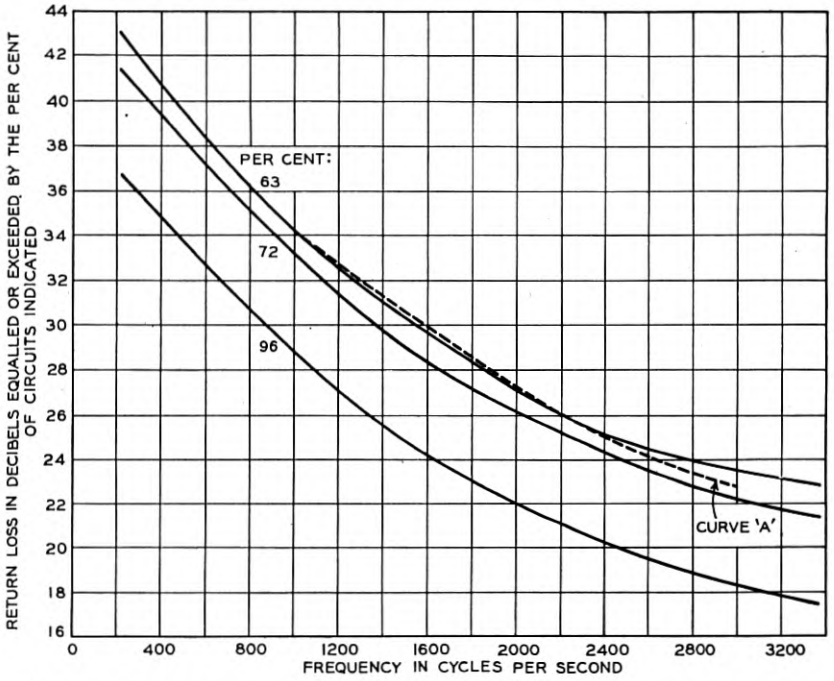

Fig. 12—Return loss—frequency characteristics. Measured from Barrie to Toronto on 53 pairs; building-out condenser adjusted to theoretical value; curve '*A*' is the 63 per cent return loss computed from attenuation and irregularity function measured on cable.

comparison, and it is to be noted that the agreement with actual results is remarkably good. The results obtained at Toronto are better than those at Barrie, except below 600 cycles, which is the frequency range of the impedance departures discussed in connection with Figs. 9, 10, and 11.

Analysis of the near-end crosstalk measurements indicated that at Barrie 98.8 per cent, and at Toronto 96.3 per cent of the combinations were equal to or better than the 74.5 db assumed for the preliminary

calculations. Investigation of the combinations poorer than 74.5 db indicated that most of the pairs involved could be assigned either to non-repeatered short circuits or to repeatered short circuits on which the repeater gains were considerably lower, with consequent lower crosstalk gains, than on the full length circuits assumed for the limit of 74.5 db crosstalk. This required the opening of one splice near Toronto for pair rearrangement.

It was decided to place the cable under permanent gas pressure in order to control service interruption as far as practicable. As there was no previous experience available for cables of this size, an investiga-

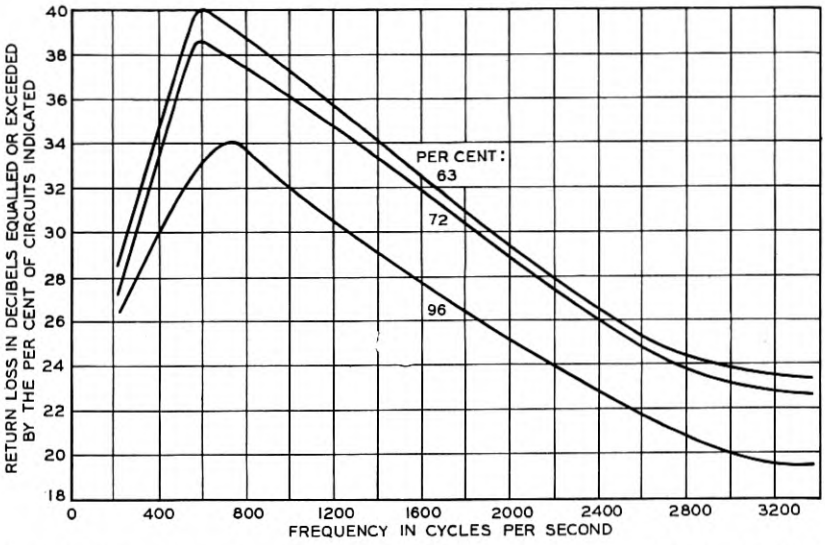

Fig. 13—Return loss—frequency characteristics; measured from Toronto to Barrie on 87 pairs; building-out condenser adjusted to theoretical value.

tion on the job was undertaken to obtain the information necessary for successful application of the gas pressure installation. Based on the results obtained, the installation of gas pressure was completed satisfactorily.

Acknowledgment

The design and installation of this cable represent the coordinated efforts of many people—members of the organizations of the Bell Telephone Laboratories Inc., the American Telephone & Telegraph Company, Northern Electric Company, Western Electric Company and the Bell Telephone Company of Canada—too many for anyone to be specifically mentioned. To all of these credit is due and is here given.

The Computation of the Composite Noise Resulting from Random Variable Sources

By E. DIETZE and W. D. GOODALE, Jr.

A statistical method is described for computing the meter reading which would be obtained on a sound level meter when used to measure room noise resulting from the random concurrent operation of a number of intermittent or continuous noise sources. The application of the method in the solution of practical problems is illustrated.

IT is generally recognized that the effects of noise upon the individual exposed to it are, to a large extent, dependent on the loudness of the noise. Various tests have been made of the relation between loudness and the different effects of noise, such as interference with hearing, reaction on the nervous system, disturbance of rest, reduction of working efficiency,[1] etc.

It has also been recognized that the ear itself, in general, is not a convenient means for the accurate measurement of loudness, especially in absolute terms. To overcome this difficulty sound level meters [2] have been made available for the measurement of acoustic noises or sound in general.

This paper is concerned with the application of such sound level meters in the study of noise problems and, in particular, with the question of determining the contribution of individual noise sources to the general "composite noise" including noise sources whose outputs are random, discontinuous variables. The paper does not concern itself with the attributes of loudness or the effects of noise, but merely with the computation of a meter reading of the total noise from available measurements of the noise components. It is recognized, of course, that not only are sound level meter readings an incomplete description of the effect of a change in noise but that considerable experience is required to appreciate properly the significance of the decibel unit employed.

Types of Problem

The method described in this paper has been developed to meet a very practical need experienced in the solution of a large variety of noise problems. To illustrate, consideration may be given to reducing

[1] For reference see Bibliography.

the noise in a room by excluding street noise, using quieter office equipment or sound absorbing material. Each of these measures will involve a certain expense and will reduce room noise a certain amount. Which of these measures will be of greatest benefit and most economical, i.e., give the greatest noise reduction per dollar expenditure?

In another assumed case, the noise from a certain noise producer, a piece of machinery, a ventilating system, etc., is known. Will this apparatus be objectionable in the particular location for which it is considered?

These and similar questions can be answered by computation while the project is still in the planning stage, whereas measurements can be made only after the change has been made, i.e., after the money for the project has been spent.

The computation method, as these illustrations show, is useful in specifying apparatus and in planning working or living quarters from the noise standpoint, in studying the comparative effectiveness of various noise reducing means, etc. The method has been used in many practical problems in this way, with satisfactory results. In a number of applications covering noise from 55 to 75 db sound level the computed and measured absolute values agreed, on the average within 1.0 db, and in the worst case within 2.0 db. Computations of the effect resulting from modifications of the noise sources were checked within closer limits. A few illustrations of applications are given at the end of this paper.

The Impulsive Character of Noises

Acoustical noise frequently is composed of sounds from a large number of sources each of which produces a relatively small proportion of the total noise. Usually these individual noise sources are discontinuous, consisting of a series of individual impulses. Consider, for instance, noise from a busy street. The hearers' first impression is that of a general roar. After a period of listening, however, a variety of individual sources may be distinguished, such as: The movement of automobiles, squeaking of brakes, whistles, street car wheels and bells, hammering and riveting from building operations, footsteps and conversations of people, etc. Each of these sources has a distinct time pattern and even those that appear most steady can frequently be broken up into impulses. For instance, the noise from an automobile passing down the street is composed of a series of impact noises which depend on unevenness of the pavement, the driving gears, number of cylinders in the engine, etc.; the hum of conversation of people in the street is composed of individual syllabic speech sounds from the different talkers.

These impulses occur at a rate which usually is not uniform. Provided, however, that the general conditions do not change, the rate approaches uniformity if the time interval considered is sufficiently large. Some of the impulses from the different sources are superimposed upon each other while others fall in the intervals between the impulses from other sources. As the amount of noise increases, two general phenomena are observed: First, the loudness of the noise increases due to the superposition of impulses; secondly, the noise becomes steadier due to the more complete filling in of relatively silent intervals (20 db or more below the average).

Definition of Terms

The solution of the problem of computing the total noise from its component parts requires the definition of a number of terms and a study of the characteristics of the implied measuring instrument.

Definition 1

Each individual producer of noise is referred to as a "noise source."

Illustrations of noise sources are: For the case of room noise—the conversation of one person, the noise from a typewriter or from a fan in the room. A number of sources of street noise have been mentioned above in illustrating the impulsive character of common noises.

Definition 2

The deflections on the measuring device (sound level meter) produced by the impulses of a single source are called "source peaks."

A peak is obtained by passing a noise impulse into a sound level meter. Depending on this measuring device, the characteristics of a peak differ from those of an impulse. The characteristics of the sound level meter, therefore, are important in connection with this computation method. Three of these, the frequency response, the rule of combination of the frequency components of a complex wave, and the dynamic characteristic of the indicating meter, are here considered in detail. These are defined in the "American Tentative Standards for Sound Level Meters" approved by the American Standards Association [2] from which the following abstracts are made:

1. The free field frequency response of a sound level meter, provided only one response is available, shall be the 40 decibel equal loudness contour modified by differences between random and normal free field thresholds. Methods are given in the ASA specification for correcting the reading when the microphone of the sound level meter responds differently to sound waves arriving with different angles of incidence.

2. The rule of combination is specified so that the power indicated for a complex wave shall be the sum of the powers which would be indicated for each of the single frequency components of the complex wave acting alone.
3. The dynamic characteristic of the indicating instrument is to be such that the deflection of the indicating instrument for a constant 1000-cycle sinusoidal input shall be equalled by the maximum deflection of the indicating instrument for a pulse of 1000-cycle power which has the same magnitude as the constant input and a time of duration lying between 0.2 and 0.25 second.

In addition, the method of reading the sound level meter is important. Where the noise is steady, it is fairly obvious how the meter should be read. When, however, the noise fluctuates, a certain amount of judgment is involved in obtaining an average. A satisfactory procedure in this event is to take a series of instantaneous readings of the noise peaks at approximately 5-second intervals for a period of time sufficient to include all noise sources. One or more of these series of measurements may be made depending on the regularity of occurrence of the noises of interest. The average and standard deviation of the fluctuating noise may then be determined from these measurements.

Using simplifying approximations based on these specified characteristics a peak may be defined as follows:

A peak is an impulse integrated by the measuring device. Its frequency components are weighted in accordance with the loudness weighting incorporated in the meter and combined by direct power addition.

It will be seen from the foregoing that the duration of the source peaks depends on the period of the indicating meter. It has been found that 0.2 second gives satisfactory correlation between computed values and actual sound level meter readings, and is in reasonable agreement with the above specified characteristics. Due to the meter characteristics, full magnitude is not indicated for impulses shorter than 0.2 second. Several impulses in the same integration period appear as a single peak on the meter. Impulses lasting longer may be regarded as producing a number of consecutive peaks. A steady noise, for instance, would be considered as consisting of a series of consecutive peaks of equal magnitude.

On the assumption of discrete integration intervals the average reading on a single source is the arithmetic mean of the intensities of the source peaks. Hence for a source, j, producing on the average m_j peaks per minute of intensities, I_{1_j}, I_{2_j}, I_{m_j}, the average reading on the meter is given by

$$I_j = \left(\frac{1}{m} \sum_{i=1}^{m} I_i \right)_j . \tag{1}$$

On the assumption of discrete integration intervals, furthermore, a source can produce no more than one peak every 0.2 second. The maximum number of peaks per minute that can be obtained from a single source, consequently, is 300.

Definition 3

The noise from all sources as measured by the indicating meter is called composite noise.

Room noise measured at a given observing position in the room is an illustration of composite noise. The peaks of a composite noise are called "composite peaks." Composite peaks have similar characteristics to source peaks as regards duration, frequency weighting, etc.

Statistical Method of Combining Noise Sources

In developing this computation method the principal aim of the authors has been to provide a practical, working method which is easy to handle yet is sufficiently reliable for engineering purposes. In accordance with this objective, a number of simplifying assumptions have been made. Some of these have been indicated in connection with the discussion of the assumed characteristics of noise peaks. The division of the time into discrete 0.2 second intervals is another approximation which has been made. The statistical treatment, in addition, includes approximations which are usual in probability mathematics of this type. Practical experience has shown that these approximations do not lead to errors which affect the usefulness of the method.

In the following an expression is derived for computing the average intensity of the composite noise from the average intensities of the source peaks and their number. Consideration is first given to the case when only a single source peak may occur in each 0.2 second interval. The consideration is then extended to cover the general case when more than one source peak may occur in a 0.2 second interval.

When only one source peak may occur in a 0.2 second interval, the average intensity $\bar{I}$ of the composite noise for these intervals is the arithmetic mean of the intensities of the source peaks, weighted by their frequency of occurrence.

$$\bar{I} = \frac{m_1}{N} I_1 + \frac{m_2}{N} I_2 + \cdots \frac{m_n}{N} I_n, \tag{2}$$

where

$I_1, I_2, \cdots I_n$ = average intensities of the sources 1, 2, $\cdots n$,
$m_1, m_2, \cdots m_n$ = number of peaks of each source per minute,

$N = \sum_{j=1}^{n} m_j$ = total number of source peaks per minute.

If several source peaks occur in the same 0.2 second interval, they will appear as a single composite peak on the meter. On the assumption of discrete 0.2 second intervals, these source peaks coincide. Their intensities, consequently, add up directly. For instance, if two source peaks occur during each integration period, the average intensity of the composite noise will be twice the arithmetic mean of the intensities. Similarly, the average intensity of the composite noise, when the number of source peaks per 0.2 second interval averages α, will be

$$I = \bar{I} = \alpha \left(\frac{m_1}{N} I_1 + \frac{m_2}{N} I_2 + \cdots \frac{m_n}{N} I_n \right). \tag{3}$$

Let M = the total number of composite noise peaks per minute. The maximum value that M can have is 300, the number of integration periods per minute. Unless the composite noise is continuous, however, there will be a certain proportion of time, t_0, in which no composite peaks occur. M then can be determined from the relation:

$$M = (1 - t_0)\, 300. \tag{4}$$

If, on the average, α source peaks per 0.2 second occur, the following relation holds between the total number of source peaks, N, and the number of composite peaks:

$$M = \frac{N}{\alpha}. \tag{4a}$$

Introducing this expression in equation (3) gives

$$I = \frac{m_1}{M} I_1 + \frac{m_2}{M} I_2 + \cdots \frac{m_n}{M} I_n. \tag{5}$$

As shown by equation (4), M is a function of t_0, the proportion of time in which no composite noise peaks occur. The value for t_0 can be found, as follows: The proportion of time when source j has a peak is equal to the probability $p_j = m_j/300$, and the proportion of time when source j has no peak is $q_j = 1 - p_j = 1 - (m_j/300)$. The proportion of time, t_0, when there are no peaks from any source then is equal to

the product

$$t_0 = q_1 q_2 \cdots q_i \cdots q_n.$$

This expression can be simplified, when the number of sources is large and none is particularly outstanding, by considering, instead of the individual sources, an average source having $m = N/n$ peaks.

The average probability then is

$$p = \frac{m}{300} = \frac{N}{300n},$$

and

$$q = 1 - \frac{N}{300n},$$

which leads to the approximation:

$$t_0 = q^n = \left(1 - \frac{N}{300n}\right)^n.$$

This expression can be further simplified when the number of sources is large and $p = N/300n$ is small by using the Poisson exponential limit:

$$t_0 = \left(1 - \frac{N}{300n}\right)^n \leqq e^{-N/300},$$

where

$$e = 2.718 \cdots$$

so that for this case

$$M = (1 - e^{-N/300})300. \tag{4b}$$

Measurement of Source Distributions

The method outlined in this paper for computing the composite noise assumes that information on the noise sources is available. Such data, therefore, must be obtained before the method can be applied. Representative measurements for a particular type of noise source, however, when once obtained, can be used in any future noise computation involving such a source.

It is necessary to consider carefully the acoustic conditions under which the sources are measured. For greatest accuracy the ambient noise level at the point of measurement should be 20 db or more below the average level of the source. Errors due to reflections can be minimized by making the measurements at a relatively short distance from the source out of doors or in a room that contains a large amount of absorbing material. A distance of 2 feet is a convenient value for most cases, and will be used as a reference value throughout the rest of this paper.

Readings should be obtained on all the noise peaks while the source is being operated in a normal manner. If the scale of the particular sound level meter used is limited, it may not be possible to read the highest as well as the lowest peaks with a single potentiometer setting. In such cases, the distribution of peaks may be measured in two or more groups.

Figure 1, Curve *A*, illustrates the measurement of source peaks in the laboratory and represents a cumulative distribution of the peaks from a typewriter as measured at a horizontal distance of about 2 feet from the type bar guide. The machine was operated by an experienced typist at an average rate.

When it is not possible to simulate actual conditions of use of a device sufficiently well in the laboratory, measurements on the source

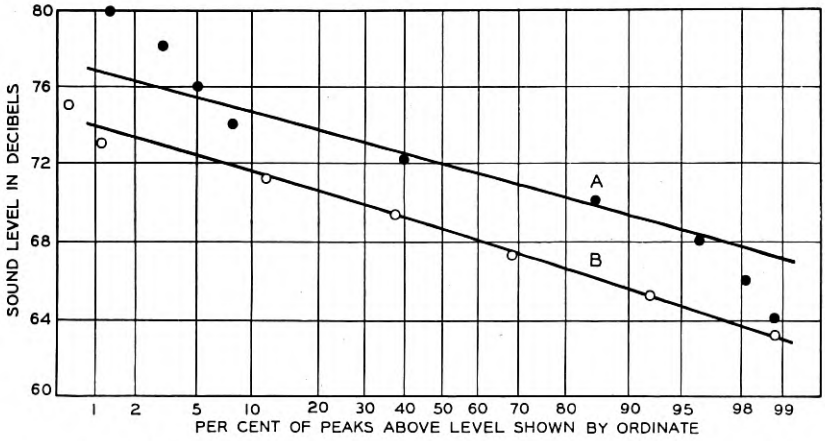

Fig. 1—Distribution of noise peaks in a typing room.

will have to be taken in the field. This may involve measuring in the presence of considerable noise from other sources. In general, it is feasible only to measure source peaks which are above the ambient noise level. If, however, an appreciable number of peaks is above this noise level, the rest of the distribution can be estimated and the average value determined. Statistical methods for doing this have been worked out for the case of normal distribution curves.[3] Experience has indicated that the distributions of noise in db frequently are approximately normal, so that these methods are applicable.

Figure 2 is an illustration of a distribution of a group of sources measured under adverse noise conditions. This curve shows the noise which came from the metal trays in a cafeteria. The distribution had to be obtained in the field because it was not feasible to estimate in

the laboratory how the customers would handle the trays. The points on the curve indicate the peaks that could be measured in the cafeteria which had an average composite noise of 66.5 db. It will be seen that the lowest peaks that could be measured satisfactorily were at 74 db sound level. The rest of the curve was estimated using the statistical methods referred to above.

The curves in Figs. 1 and 2 are plotted on "arithmetic probability paper." On this paper, cumulative normal distributions appear as straight lines.

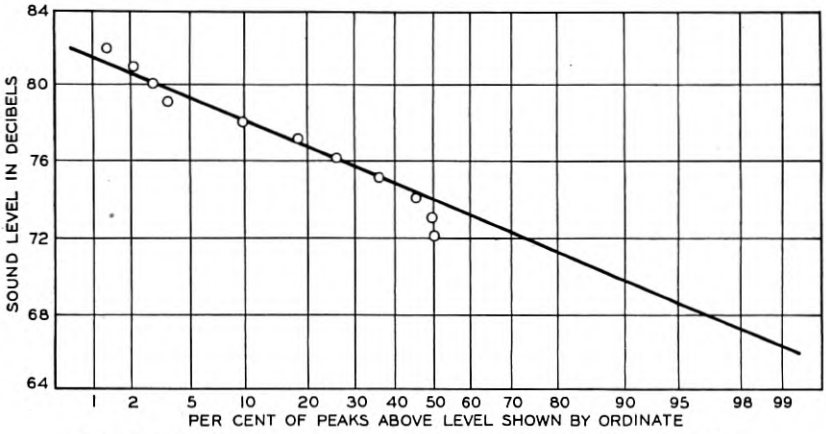

Fig. 2—Distribution of noise peaks from metal trays in a cafeteria.

Effect of Room Characteristics

Generally the noise sources are at various locations so that it is necessary to determine how much the noise from each is reduced by its distance from the observing point assumed for the computation.

Since it is not the primary concern of this paper to discuss the distribution and decay of sounds in rooms, only a very simple approximate method of computing distance losses, based on the classical theory of the steady-state distribution of sound in a room, is given here. This method has been found adequate for practical purposes in rooms having relatively simple geometric shape and large enough dimensions so that the sound is diffused. For a more complete treatment of room acoustics the reader should refer to the literature on this subject.[4]

The total steady-state intensity, I_T, at an assumed observing position in a room consists of two parts: I_R, the reflected sound intensity and I_D, the direct sound intensity, so that:

$$I_T = I_R + I_D.$$

Assuming the reflected sound to be uniformly distributed in the room, it can be shown that:[4]

$$I_R = \frac{.0038E}{\frac{aS}{S-a}},$$

where E = power emitted by source, in ergs per second,
S = total surface area of the room in square feet,
a = absorption in square feet of equivalent open window.

Introducing $F = aS/(S - a)$, the above becomes:

$$I_R = \frac{.0038E}{F}.$$

Assuming the sound source to radiate hemispherically, as is frequently the case because it is associated with a large surface acting as a baffle, the direct sound intensity is:

$$I_D = \frac{E}{2\pi r^2 v},$$

where r = distance from source, in feet,
v = velocity of sound, in feet per second.

In the above expression the direct sound intensity decreases inversely as the square of the distance from the source. This shows that room absorption is effective mainly in reducing the noise from sources at a considerable distance from the observing point, but has relatively little effect on nearby sources.

The curves in Fig. 3 give the variation in the total sound intensity, I_T, with distance from the source for different values of $F = aS/(S - a)$, as computed by means of the above expressions.

Computation of Composite Noise

In the following, the application of the statistical method outlined above is discussed. Since noise measurements are usually expressed in db sound level, it is necessary to change the form of the equations given in the preceding sections. For this purpose equation (5) is rewritten as follows:

$$\frac{I}{I_0} = \frac{m_1}{M}\frac{I_1}{I_0} + \frac{m_2}{M}\frac{I_2}{I_0} + \cdots \frac{m_n}{M}\frac{I_n}{I_0}, \tag{5a}$$

where I_0 = reference sound intensity.

This equation can also be written

$$\frac{I}{I_0}\frac{M}{300} = \frac{m_1}{300}\frac{I_1}{I_0} + \frac{m_2}{300}\frac{I_2}{I_0} + \cdots \frac{m_n}{300}\frac{I_n}{I_0}. \tag{5b}$$

Equation (5b) is somewhat more convenient in computing than (5a). In this equation a weight factor is associated with each intensity ratio, which is in each case the actual number of peaks divided by the maximum possible number of peaks.

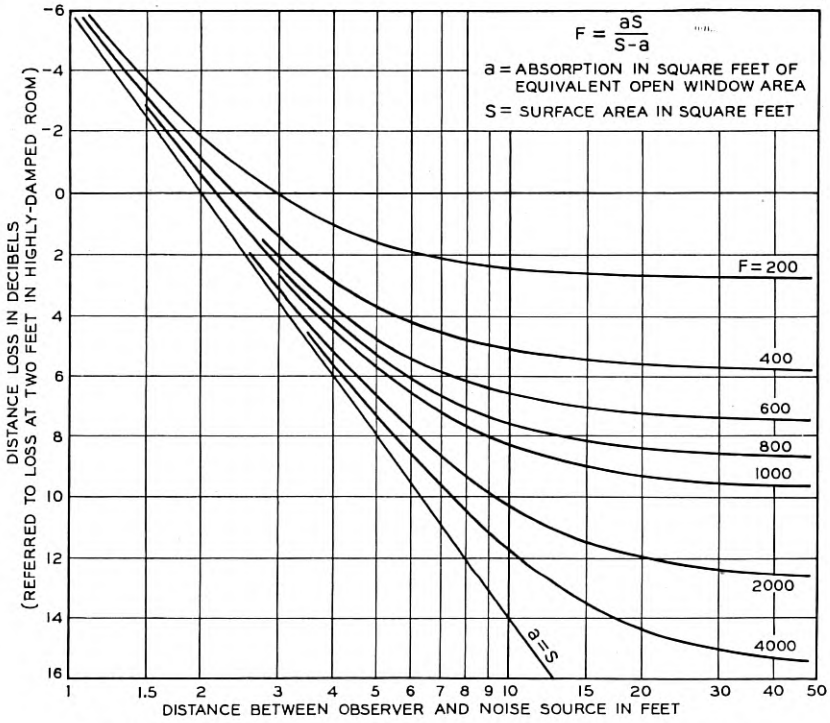

Fig. 3—Loss of intensity with distance from noise source for various amounts of room absorption.

Assuming that the intensity corresponding to the average of the db distribution of each source may be used, the following relations exist for the source noises in db sound level:

$$\left.\begin{aligned} A_1 &= 10 \log_{10} \frac{I_1}{I_0} \\ A_2 &= 10 \log_{10} \frac{I_2}{I_0} \\ &\cdots \\ A_n &= 10 \log_{10} \frac{I_n}{I_0} \end{aligned}\right\} \tag{6}$$

and for the average composite noise in db sound level:

$$A = 10 \log_{10} \frac{I}{I_0} . \tag{7}$$

It is usually convenient to use logarithmic weight factors

$$\left.\begin{aligned} w_1 &= 10 \log_{10} \frac{m_1}{300} \\ w_2 &= 10 \log_{10} \frac{m_2}{300} \\ &\cdots\cdots \\ w_n &= 10 \log_{10} \frac{m_n}{300} \\ w &= 10 \log_{10} \frac{M}{300} \end{aligned}\right\} \tag{8}$$

Figures 4 and 5 permit ready computation of these logarithmic weight factors. The chart in Fig. 5 is based on the relation between M and N given in equation (4b). It should be recalled that the derivation of this equation involved a number of approximations. This expression especially does not apply when one or more of the noise sources are continuous, in which case the exact expression (eq. 4) gives $M = 300$ (for $t_0 = 0$). Hence $w = 0$ in this case.

The terms of equation (5b) then can be rewritten in logarithmic form by using equations (6), (7) and (8), as follows:

$$\left.\begin{aligned} A_1 + w_1 &= 10 \log_{10} \frac{m_1}{300} \frac{I_1}{I_0} \\ A_2 + w_2 &= 10 \log_{10} \frac{m_1}{300} \frac{I_2}{I_0} \\ &\cdots\cdots \\ A_n + w_n &= 10 \log_{10} \frac{m_n}{300} \frac{I_n}{I_0} \\ A + w &= 10 \log_{10} \frac{M}{300} \frac{I}{I_0} \end{aligned}\right\} \tag{9}$$

This gives the following formula:

$$10^{\frac{(A+w)}{10}} = 10^{\frac{(A_1+w_1)}{10}} + 10^{\frac{(A_2+w_2)}{10}} + \cdots 10^{\frac{(A_n+w_n)}{10}}, \tag{10}$$

from which the average composite noise A can be found.

The application of this expression is materially simplified by the use of the chart shown in Fig. 6. Power addition of a number of

components may be carried out with this chart by first adding two components, then adding the resultant to a third component and continuing until all components have been summed up. Incidentally,

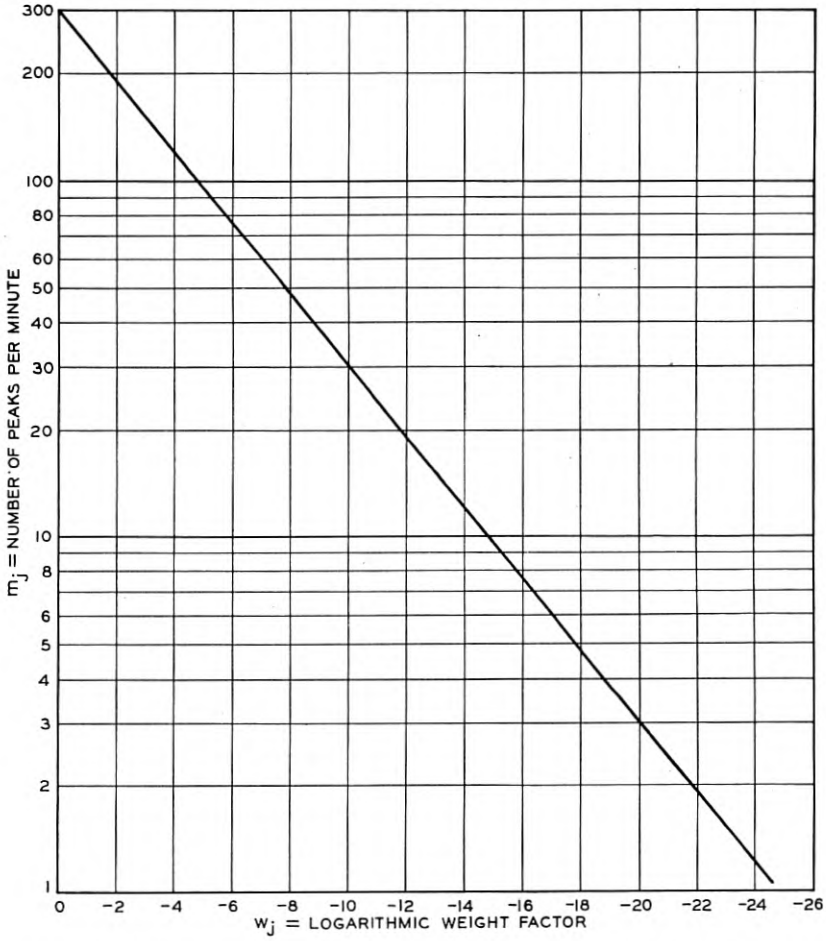

Fig. 4—Relation between peaks per minute (m_j) produced by a noise source and its logarithmic weight factor (w_j).

the chart shows that the contribution to the composite noise from a source whose weighted intensity ($A_1 + w_1$) is 20 db or further below that of another noise source ($A_2 + w_2$) is negligible.

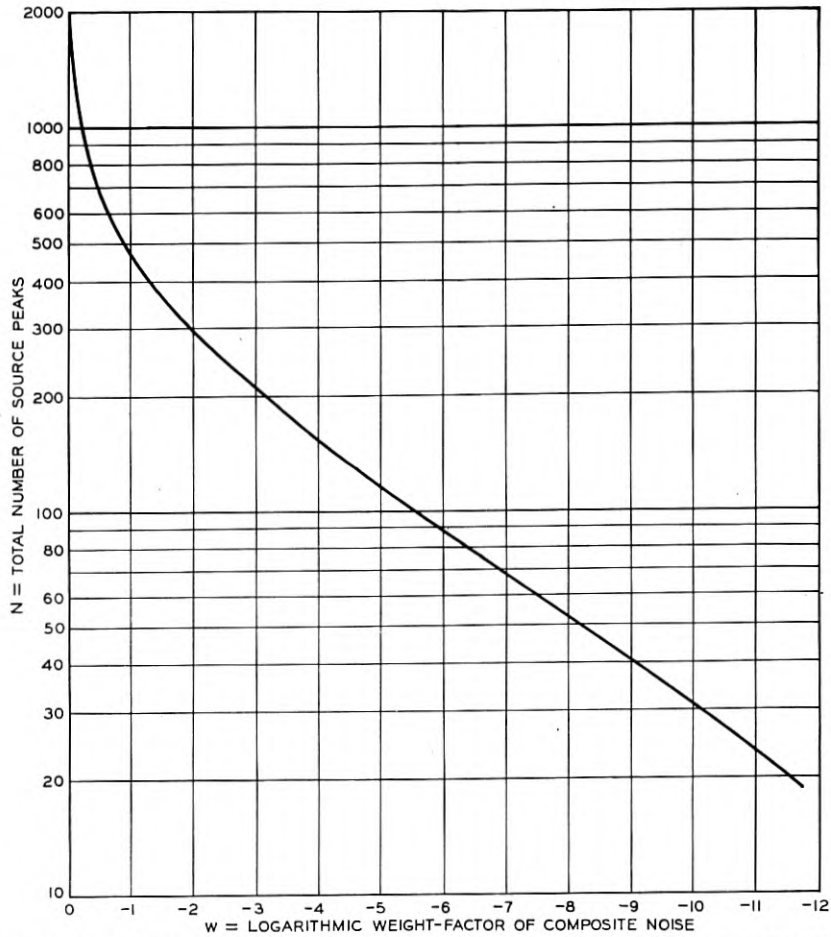

Fig. 5—Relation between total number of peaks per minute (N) and the logarithmic weight factor of composite noise (w).

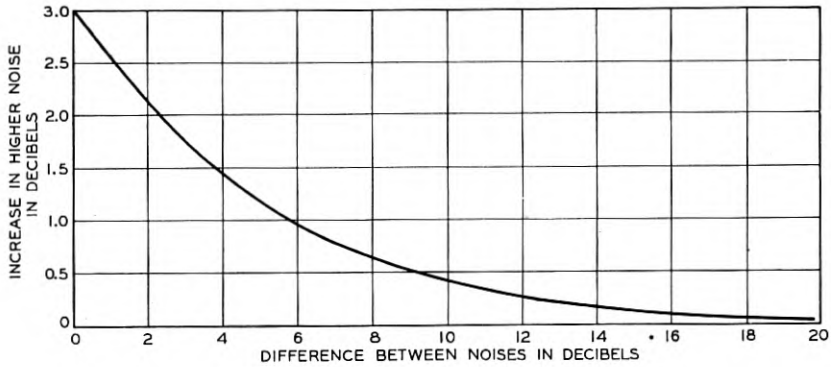

Fig. 6—Power addition of noises.

Applications

In the following, several applications of the theory are made to illustrate its practical usefulness.

I. The composite noise in a typing room is computed. This computed noise is compared with actual measurements. The effect of increased room absorption is discussed.

II. The effect on the composite noise of installing additional office equipment is computed for the two cases where this equipment produces a continuous noise and where its noise is intermittent.

III. The maximum permissible noise from added equipment is determined on the basis that the composite noise level shall not be increased by more than 0.5 db.

Problem I

For the purpose of computing the noise at a given location in the typing room, the following information was obtained:

A distribution of the noise from a typical typewriter was measured at a distance of 2 feet from the type bar guide while the machine was being operated at a normal rate. This is shown by Curve A of Fig. 1.

A location was chosen as a point of observation, and the distances between it and the typing desks were measured.

Estimates of the time spent in typing at each desk were obtained which, taken together with data on average typing speeds, gave information on the number of typing peaks produced per minute at each desk.

Computation of the absorption of the room using the usual values of absorbing coefficients [5] gave a value of 650 units for F.

Noise due to other sources, such as conversation and street noise, was negligible in this room.

The table shown below was then prepared.

In this table Column 2 gives the average noise A_i' produced by each of the sources at 2 feet distance. This value is the median point of Curve A in Fig. 1. Column 3 is the average number of source peaks per minute m_i produced at each desk. Column 4 is obtained from Column 3 by using Fig. 4. The total number of source peaks is $N = 750$, and from Fig. 5 the composite noise weight factor $w = -0.4$ db. The distances between the observing position and each source are given in Column 5 and the losses in db due to these distances are given in Column 6. These values were obtained from Fig. 3 for a value of $F = 650$. Column 7 is obtained by subtracting the losses of Columns 4 and 6 from the values of Column 2.

Adding the values of Column 7 successively on a power basis by means of the curve in Fig. 6 gives $A + w$ from which is obtained the total composite noise $A = 69.7$ db sound level. This differs by approximately 1 db from the average of the measured composite noise distribution shown by Curve B of Fig. 1.

The effect of sound treatment on the walls and ceiling in reducing the typing room noise may readily be calculated by means of the curves in Fig. 3. Supposing that the added absorption raises the value of F from 650 to 2000 units, this figure shows that noise produced by sources 20 feet or more away from the observing point would be reduced by approximately 5 db. At shorter distances the reduction would be less. For the observing position here considered, a computation similar to that carried out above indicates that the composite noise level would be reduced about 3 db by the added absorption in the room.

TABLE

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Source No.	A_i' = Average Source Noise at 2 ft.	m_i = Source Peaks per Minute	w_i = Freq. Weight Factor	Distance from Source	Intensity Loss vs. 2 Feet	$A_i + w_i$ = Weighted Source Noise at Observing Position
	db Sound Level		db	Feet	db	db Sound Level
1	71.9	105	−4.5	18	−7.4	60.0
2	71.9	90	−5.2	13	−7.2	59.5
3	71.9	90	−5.2	9	−6.7	60.0
4	71.9	120	−4.0	9	−6.7	61.2
5	71.9	120	−4.0	7	−6.0	61.9
6	71.9	40	−8.8	7	−6.0	57.1
7	71.9	65	−6.6	7	−6.0	59.3
8	71.9	120	−4.0	7	−6.0	61.9

Total Source Peaks: $N = 750$

$w = -0.4$ db (Fig. 5)

Power Addition $A + w = 69.3$ db (Fig. 6)

$A = 69.7$ db sound level.

Problem II

A piece of office machinery, such as an addressing or copying machine, which produces an average sound level of 75 db at 2 feet distance, is to be installed in the typing room considered in Problem I, 20 feet away from the observing position. How much will the composite noise level be raised?

(*a*) *The machine produces a steady noise.* The new value for the number of composite noise peaks is then 300 and the weight factor of this noise is zero. The distance loss of the machine noise (for $F = 650$) is -7.5 db from Fig. 3. Hence, the weighted value of

the noise from the new machine at the observing position is:

$$75 - 0 - 7.5 = 67.5 \text{ db sound level.}$$

Adding this figure to the weighted value of the existing composite noise ($A + w = 69.3$ db) on a power basis gives the new composite noise value of 71.5 db sound level (the new weight factor being zero). Hence, the composite noise at the listening position is increased 1.8 db by the machine.

(*b*) *The machine produces noise intermittently.* The increase in the composite noise level will not be as great in this case as in the preceding case. For example, assuming the rate to be 100 peaks per minute, the new value of N will be 850 peaks per minute and the corresponding weight factor from Fig. 5 will be $-$ 0.3 db. For the noise from the new machine, the weight factor (by Fig. 4) is $-$ 5.0 db. The distance loss as before will be $-$ 7.5 db. The weighted value of the machine noise at the observing position is then:

$$75 - 5.0 - 7.5 = 62.5 \text{ db sound level.}$$

Adding this figure on a power basis to the weighted value of the existing composite noise, 69.3 db, results in a new weighted composite sound level, $A + w = 70.1$ db. Since $w = -0.3$ db, this gives $A = 70.4$ db. Hence, the composite noise is increased 0.7 db by the intermittent machine noise.

Problem III

What is the maximum permissible noise, measured at 2 feet, which the machine considered in Problem II, may produce without raising the composite noise in the typing room by more than 0.5 db?

(*a*) *The machine produces a steady noise.* In this case, the composite noise has 300 peaks and its weight factor is zero. The existing composite noise was 69.7 db sound level (see Problem I). The maximum permissible value of the new composite noise level is consequently

$$A = 69.7 + 0.5 = 70.2 \text{ db sound level.}$$

Let the unknown machine noise be A_9 (its weight factor is zero), and since for the existing composite noise $A + w = 69.3$, equation (10) gives:

$$10^{\frac{70.2}{10}} = 10^{\frac{69.3}{10}} + 10^{\frac{A_9}{10}}.$$

Entering the ordinate of Fig. 6 at the value of $70.2 - 69.3 = 0.9$ db, the chart indicates that A_9 must be 6.3 db below 69.3. Hence $A_9 = 63$ db sound level at the observing position.

The distance loss for 20 feet is − 7.5 db. The machine then could produce a noise of:

$$63.0 + 7.5 = 70.5 \text{ db sound level}$$

at 2 feet distance without raising the composite noise level by more than 0.5 db at the observing position.

(*b*) *The machine produces noise intermittently.* The solution of the problem in this case follows the same lines as in Part (*a*) except that the weight factors are changed. The rate for the new machine noise, A_9, is assumed to be 100 peaks per minute, so that $w_9 = -5.0$ db, as in part (*b*) of Problem II. The maximum permissible value of the new composite noise is 70.2 db sound level (as before) but its weight factor now is − 0.3 db as in part (*b*) of Problem II. The weighted value of the existing composite noise as before is 69.3 db sound level. Equation (10) then gives:

$$10^{\frac{(69.9)}{10}} = 10^{\frac{69\ 3}{10}} + 10^{\frac{(A_9-5.0)}{10}}.$$

From Fig. 6 it is found that for a value of 69.9 − 69.3 = 0.6 on the ordinate, the abscissa is 8.3 db. Hence, $A_9 - 5.0$ must be 8.3 db below 69.3 or $A_9 = 66.0$ db sound level at the observing position. Applying the same distance loss as before, the machine could produce a noise of 73.5 db sound level at 2 feet without increasing the composite noise level by more than 0.5 db at the observing position.

From the computations, then, it may be expected that adding a steady noise will increase the general noise level more than adding an intermittent noise having the same average value, when there are a number of sources operating. That this is actually so can readily be verified by sound level measurements. As has been stated, sound level measurements under most conditions are directly related to the effects of noise upon the individual exposed to it, and the method described provides a convenient and reasonably reliable way of computing such readings and thereby makes possible the engineering analysis of noise problems.

Bibliography

1. "Environment and Employee Efficiency" by Harold Berlin and others—Office Management Series No. 81—Copyright 1937, American Management Association.
"City Noise"—Report of Noise Abatement Commission 1930, Dept. of Health, New York City.

2. "American Tentative Standards for Sound Level Meters for Measurement of Noise and Other Sounds"—Z24.3—Approved American Standards Association, Feb. 17, 1936.
"Indicating Meter for Measurement and Analysis of Noise" by T. G. Castner, E. Dietze, G. T. Stanton and R. S. Tucker—District Meeting A. I. E. E., April 1931, Rochester, N. Y.
3. "Determination of Normal Curve from Tail"—Table XI of "Tables for Statisticians and Biometricians" by Karl Pearson—Part I—Second Edition, 1924.
"Graduation by a Truncated Normal" by Nathan Keyfitz—*Annals of Mathematical Statistics*—Vol. 9, March 1938.
4. "Collected Papers on Acoustics" by W. C. Sabine,
"Reverberation Time in 'Dead' Rooms" by Carl F. Eyring—*The Journal of the Acoustical Society of America*, Vol. I, No. 2, January 1930,
"Architectural Acoustics" by V. O. Knudsen.
5. *Official Bulletin of the Acoustical Materials Association*, No. 6, March 1938.

Load Rating Theory for Multi-Channel Amplifiers *

By B. D. HOLBROOK and J. T. DIXON

The amplifiers of multi-channel telephone systems must be so designed with regard to output capacity that interchannel interference caused by amplifier overloading will not be serious. Probability theory is applied to this problem to determine the maximum single frequency output power which a multi-channel amplifier should be designed to transmit as a function of N, the number of channels in the system. The theory is developed to include the effects of statistical variations in the number of simultaneous talkers, in the talking volumes, and in the instantaneous voltages from speech at constant volume.

Introduction

IN A perfect multi-channel carrier telephone system, each channel would be entirely free from interference produced by the energy present in the other channels. Since all the channels are amplified by the same repeaters, which as a practical matter cannot have perfectly linear characteristics, this is an ideal that may be approached but not completely realized. The interchannel interference must be kept down to a value which will be satisfactory for the grade of transmission concerned, further reduction being uneconomic. To do this the repeaters must meet definite load capacity requirements and modulation (non-linearity) requirements. The load capacity requirement is most conveniently specified in terms of the maximum single frequency sine wave power which a multi-channel amplifier must transmit without appreciable overloading. The modulation requirement pertains to the performance of the amplifier for impressed loads equal to or smaller than the load capacity, and specifies the allowable power in the modulation products resulting from such loads. Because of the numerous factors which affect these requirements, their determination is a rather complicated matter and the present discussion will be restricted solely to a determination of the load capacity requirement. The object is to determine this quantity as a function of N, the number of channels in the system.

The criteria ordinarily used for determining the load capacity of single-channel amplifiers are of little use here because of two funda-

* Presented at Great Lakes District Meeting of A.I.E.E., Minneapolis, Minn., September 27–29, 1939.

mental differences between single-channel and multi-channel systems. In the first place, the modulation produced in a single-channel amplifier depends only upon the input to that channel and occurs only when the channel is energized. In addition, the most important frequencies resulting from modulation fall directly back upon frequencies already impressed and the net effect appears as a distortion of the original input, rather than as noise. The situation is entirely different in a multi-channel system. In this case, the modulation products falling into one particular channel are in the main unrelated either to the impressed frequencies or to the volume of impressed speech in that channel. Thus it is no longer possible to think of the interference as distortion; the effect must rather be considered as that of a particular kind of noise whose level depends upon the load on the other channels of the system. For a given grade of service, the ratio of signal to noise must be much larger than the ratio of signal to modulation products resulting in distortion; thus it is to be expected that the non-linearity requirements will be more stringent for multi-channel operation than for single-channel operation.

The second fundamental difference between single-channel and multi-channel systems arises from the character of the load which each system must be designed to handle. A single-channel amplifier must be capable of handling one channel at the maximum volume normally expected. Inasmuch as the amplifier will be loaded only about one-fourth of the time, even in the busiest hour, and as the average impressed volume will be some 15 db below the maximum that must be provided for, the ratio of maximum to average load of such an amplifier is inherently very high. In a multi-channel system, however, the several channels will very rarely be heavily loaded simultaneously. There is thus a favorable diversity factor, increasing with the number of channels, and multi-channel amplifiers may accordingly be worked successfully at lower ratios of maximum to average load.

Occasionally, of course, there will be short periods of excessive loading during which the interchannel interference in multi-channel systems will rise above the value normally permitted. This sort of thing often occurs when it is desired to make economical use of facilities of any kind in common. In machine switching systems, for example, it is common practice to associate a large number of lines with a smaller number of switches and trunks. The number of switches and trunks provided is sufficient to ensure a satisfactory service, with a very small probability of requiring more facilities than are available. The multi-channel amplifier problem presents a situation identical in principle, though the methods of solution are necessarily very different.

The application of probability theory is evidently indicated as the method of attack.

Those characteristics of multi-channel amplifiers which are important to the problem will be described first. Then a description will be given of the variables which must be taken into account in computing load capacity. Finally, the combined effects of these variables will be determined on a statistical basis to establish the required load capacity as a function of the number of channels in the system.

Characteristics of the Multi-Channel Amplifier

At the present time, multi-channel systems of primary interest employ single sideband transmission; the carrier frequencies are largely suppressed and different amplifiers are used for the two directions of transmission. For such systems negative feedback amplifiers have outstanding advantages, particularly with respect to stability of gain and reduction of modulation effects, and are thus being used almost exclusively in present day multi-channel systems. The following discussion is related particularly to such systems, although many of the calculations are also applicable to less common types.

At light loads the principal modulation products in a negative feedback amplifier increase approximately as the square or the cube of the fundamental output power. Beyond a certain critical point, however, the modulation increases very rapidly and the total output of the amplifier soon becomes practically worthless for communication purposes. This critical point will be called the "overload" point. For most tube circuits it is either the point at which grid current begins to flow, or that at which plate current cutoff occurs. This point obviously defines the instantaneous load capacity.

Below the overload point the higher order modulation products are negligible in comparison with second and third order products, and the interference may be regarded as due to the latter sources alone. Beyond the overload point, however, the higher order products become important very rapidly and the resultant disturbances appear in most, if not all, of the channels. With given tubes, the interference below the overload point may be altered by changing the amount of feedback. The interference above the overload point, however, may be little changed in this way because of the rapid loss of feedback as the amplifier overloads. Accordingly, in designing an amplifier, the necessary load capacity may be determined solely by insuring that the output will rarely rise above the overload point, afterwards adjusting the amount of feedback so that the inteference below the overload point will be tolerable. There are thus two problems which may be

handled separately, at least for negative feedback amplifiers, it being understood that the results are combined in the final design. As previously stated, only the load capacity problem will be considered in detail here but many of the methods used have been applied successfully to the interchannel modulation problem.

The Load on a Single Channel

The total load applied to a multi-channel amplifier varies rapidly between widely separated limits. A complete knowledge of the variations in the load applied to a single channel is necessary first; these variations arise from several recognizable causes which may be discussed separately.

Number of Active Channels

First of all, a single channel at a given instant may or may not be carrying speech; if not, it contributes nothing to the multi-channel load. A channel will be called "active" whenever continuous speech is being introduced into it; i.e., a channel is active during the time it is actually carrying speech power, and also during the short pauses that occur between words and syllables of ordinary connected speech. A channel is said to be "busy" when it is not available to the operator for completing a new call. Busy time is by no means all active time, for a busy channel is inactive during much of the time the connection is being completed, during pauses in the conversation, and finally during the time the other party is talking. The fraction of time during the busiest hour that a channel may be busy depends on the size of the group of circuits of which it is a member and on the methods of traffic operation. Measurements on circuits in large groups, made by Mr. M. S. Burgess, indicate that the largest fraction of the busiest hour that a channel may be active is about $\frac{1}{4}$. For channels in small circuit groups, this figure may become considerably smaller but it is unlikely that any probable increase in group size or improvement in operating practices will increase it appreciably. This figure, which will be represented by τ, may accordingly be taken as a conservative estimate of the limiting probability that a channel will be active in the busiest hour.

The number of channels that are active at a given instant in an N-channel system may be anything from zero to N. Inasmuch as the channels are independent, it is possible to write down at once the probability that exactly n of them are simultaneously active. This probability is

$$p(n) = \frac{N!}{n!(N-n)!}\tau^n(1-\tau)^{N-n}. \tag{1}$$

Talking Volumes

A second source of variation in the load on a given channel is that the impressed volume may have any value within rather wide limits when a channel is active. By "volume" is meant the reading of a volume indicator of a standard type. Its importance in the present problem arises from the fact that the volume is an approximate measure of the average speech power being introduced into the channel. Although some other instrument might give a better measurement of the latter quantity, only the volume indicator has been used sufficiently widely in the plant to give data on the distribution of average speech power per call under commercial conditions. The average speech power is dependent on the type of instruments, the character of the speech, and the time interval over which the average is determined. From an analysis of phonograph records of continuous speech it is found that the average speech power of a reference volume talker may be taken as 1.66 milliwatts, and the relationship between volume[1] and average power may be expressed by the following equation:

$$\text{Volume } (db) = \frac{10 \log_{10} \text{Average Speech Power in Milliwatts}}{1.66}. \quad (2)$$

This equation is based on the long average speech power. It will be understood that for purposes other than load rating computations, a different relation might be found more suitable.

The use of equation (2) to relate volume to average speech power is applicable to speech in a single channel. It is convenient to refer to a quantity related in the same way to the total average power contributed by a number of channels as the "equivalent volume."

The single-channel volumes on commercial circuits are conveniently measured at the transmitting toll test board, which will be taken as a point of "zero transmission level." Henceforth it is assumed that there is no gain or loss between this point and the output of the amplifier, so that the latter is also a point of zero transmission level. While this will seldom be the case in an actual system, the necessary change in the load capacity is easily computed. The volumes at this point are found to be distributed approximately according to the "normal" law; that is, the probability that the volume will be between V and $V + dV$ is given by

$$p(V)dV = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(V-V_0)^2}{2\sigma^2}} dV. \quad (3)$$

[1] Subsequent to the preparation of this paper, a new volume indicator was standardized for use in the Bell System. With the new volume indicator, volume is expressed in *vu*, + 8*vu* being approximately equal to reference volume (0 *db*) as used herein.

For calls on typical toll circuits, the best present values for the parameters are $V_0 = -16.0$ db and $\sigma = 5.8$ db. These parameters depend, of course, upon the character of the local plant and upon the habits of telephone users, and changes in either will affect their values. Curve A of Fig. 1 shows this distribution of talker volumes at a point of zero transmission level. Curve B of Fig. 1 is the talker volume distribution used for load rating computations when a particular amount of peak amplitude limiting occurs in the terminal equipment. This will be discussed later. Although the mean volume is V_0, the

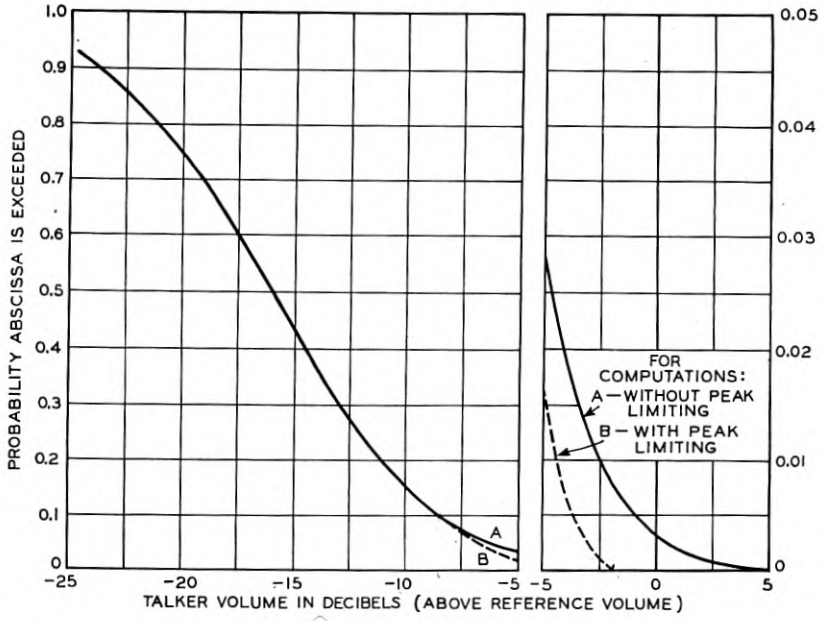

Fig. 1—Talker volume distribution.

volume corresponding to the mean power of the distribution (3) is equal to $V_0 + .115\sigma^2$, as may be seen by converting the volume scale of the distribution to power ratios, averaging, and reconverting the average to volume in db. For the values of the parameters given above, $V_0 + .115\sigma^2 = -12.1$ db.

Instantaneous Voltage Distribution

Finally, the voltage in an active channel fluctuates widely even at constant volume. Not only the differences between successive syllables and the differences between vowel sounds and consonants, but also the fine structure of single sounds, are important in this connec-

tion. The total voltage impressed on the amplifier is the quantity which determines whether or not it will overload, and the phases as well as the amplitudes of the frequency components in the several channels must be considered in determining this. It is most convenient for analysis to work directly with instantaneous voltages of speech, the frequency of occurrence of the magnitudes being expressed in the form of a distribution function.

This distribution function has been measured by Dr. H. K. Dunn, using apparatus which measures 4 samples per second of the instantaneous voltage out of a commercial subset and typical loop. By operating the apparatus until about 1000 successive samples have been measured, usable distribution curves of instantaneous voltage are obtained; this is readily checked by making repeated runs comprising the same number of samples on speech recorded on high quality phonograph records. It is, of course, known that commercial transmitters have considerable asymmetry as regards positive and negative voltages but the poling referred to the toll board is expected to be random. As the measurements were considerably simplified by doing so, it appeared desirable to average out this asymmetry by arranging a linear rectifier ahead of the sampling apparatus to obtain equal samples of positive and negative voltages.

Such measurements have been made for a number of different talkers, different commercial subsets, and different volumes, with the speech input held at substantially constant volume in each test. The various subsets now in commerical use all give essentially the same distribution curve. The resulting distributions, if they are considered as functions of the ratio of instantaneous to rms voltage, are also nearly independent of the speech volume at the subset. Specifically, the only important effect of volume is that which may be ascribed to amplitude limiting in the transmitter; i.e., to the fact that the transmitter itself has a limited load capacity. However, this effect does not appear until the volume is 10 db or more above the mean of the volume distribution curve, and is only of importance for talkers at still higher volumes. For all lower volume talkers, the instantaneous voltage distribution may be considered as the same for all volumes when expressed as a ratio of instantaneous to rms voltage. The cumulative distribution curve of the quantity E/U, where E is the rectified instantaneous voltage and U the rms voltage, is shown by the curve $n = 1$ of Fig. 2.

Voltage Limiting

While this curve of Fig. 2 is accurate for the bulk of the talkers, it changes for the high volume talkers who overload the subset trans-

mitters. It is also the custom to provide a certain amount of amplitude limiting in each channel by suitable circuit design of the channel terminal equipment. This limiting alters the shape of the instantaneous voltage distribution curve for a range of voltages below the maximum, the extent of the modification depending on the talker volume and the characteristics of the limiting device. Its effect on the load capacity will be considered later.

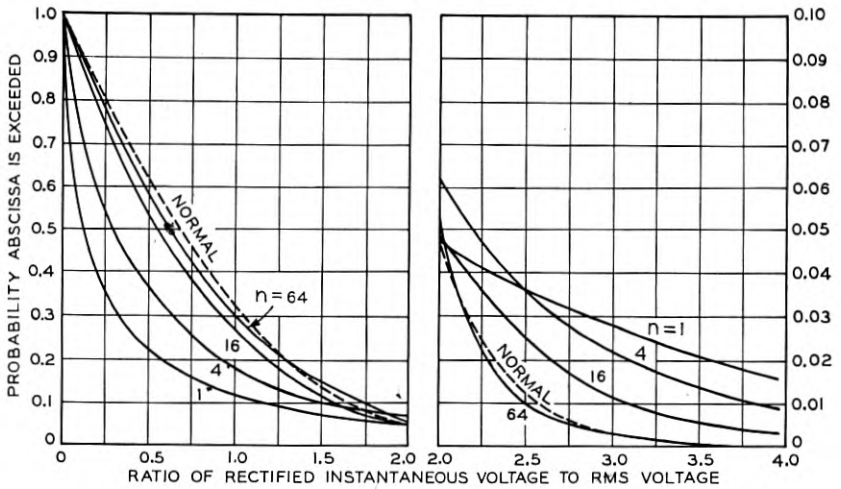

Fig. 2—Instantaneous voltage distributions for n talkers.

Multi-Channel Instantaneous Voltage Distributions

The number of variables with which it is necessary to deal makes the general load capacity problem rather a complicated one. The analysis will be easier to follow if the effects of the different variables are taken up one at a time, thus building up a complete theory in successive steps. To do this, it is advantageous to start with a case so simplified that it rarely, if ever, occurs in ordinary practice; i.e., that in which the volumes in all the channels are regulated to a common constant value, and in which the number of *active* channels is also kept constant. For this condition, it is necessary to consider only the effects of the distribution of instantaneous voltages in each channel. This distribution curve is the same for all of the channels, since all are at the same volume, but the voltage in any channel at a particular instant is entirely independent of the condition of the other channels.

Overload Expectation

The total voltage impressed on the amplifier by a number of channels at a given instant is the sum of the instantaneous voltages in the

separate channels. Since disturbances will be produced in many of the channels when the applied voltage goes beyond the overload point, it will be useful to know the fraction of the time that this may be expected to occur; this fraction will be called the overload expectation and denoted by ϵ. It is important to notice that this quantity ϵ is not necessarily the fraction of the time during which the performance of the amplifier will be unsatisfactory. This might perhaps be the case for a device having an instantaneous cutoff characteristic, but for an ordinary amplifier the time constants (among other things) affect the results of overloading. The interpretation of the overload expectation will be discussed further later; consideration must be given first to how it is obtained.

The n-Channel Voltage Distribution

The load in each channel is applied at voice frequency to the input side of a modulator, the voice frequency instantaneous voltage distribution being as shown by the curve $n = 1$ of Fig. 2. The overloading of the amplifier is determined, however, by the distribution of the sum of n such voltages after each has been shifted by the modulator to the appropriate carrier frequency, one side-band being suppressed. It may be shown that if the phases of the various components of the voice frequency input were random, the distribution of instantaneous voltage at side-band frequency would be identical with that measured at voice frequency. It is known, however, that the phases at voice frequency are not entirely random, and there may thus be differences between the two distributions. The results of a number of tests bearing upon this point indicate that any error resulting from the use of the distribution measured at voice frequency will be small for systems of few channels, and will rapidly disappear as the number of channels is increased.

Theoretically, the resultant n-channel voltage distribution can be derived from the single-channel distribution by straightforward analytical methods; in the present case, however, expression of the result in useful form is very difficult because of the form of the single-channel curve. This difficulty might be resolved by using graphical or numerical methods, as applied later to the volume distribution curves; fortunately, the fact that the voice frequency voltage distributions may be used throughout permitted the resultant n-channel distributions to be obtained much more easily. Since the addition of voltages from the several carrier channels does not depend materially upon the frequencies at which the channels appear in the system, the addition of n channels at voice frequency will give the desired n-

channel distribution directly. Mr. M. E. Campbell effected this addition by the use of phonograph records, the n-channel distributions being determined by means of the instantaneous voltage sampling apparatus previously mentioned.

As material for this process, 16 high-quality phonograph records were made of the outputs of commercial subsets through representative subscriber loops. Both male and female voices were used. The speech was furnished by reading magazine stories containing considerable conversational material, due precautions being taken that the volume on each record was substantially constant throughout. A calibrating tone was cut on each record to enable it to be played at any desired volume and most of the volumes recorded were well below the point at which the transmitter began to act as a voltage limiter.

These individual records were then combined in groups of four, with all records adjusted to the same volume by means of the calibrating tones, and re-recorded. Several such 4-voice records were made; by combining them again in the same way, 16-voice records and finally 64-voice records were obtained. The instantaneous voltage distributions were measured before and after each re-recording to insure that the recording process introduced no errors. A few minor discrepancies were found, but all were small enough to be disregarded. Each single-voice record appeared several times in a 64-voice record, but since the phases of its different appearances were random, this had no appreciable effect on the resultant voltage distribution. This was verified by comparing the voltage distributions of the various possible 16-voice combinations. By this process n-channel voltage distributions were obtained for $n = 1, 4, 16$ and 64.

These distributions, together with a normal curve, are shown in Fig. 2 in cumulative form. To show the curves conveniently to the same scale, it has been necessary to plot for each case not the distribution of E, the rectified instantaneous voltage, but that of E/U, where U is the rms voltage. The rms voltage, it will be remembered, is directly related to the equivalent volume by equation (2). The figure shows clearly the gradual transition from the single-channel distribution to the normal one for large n, and also indicates that for 64 active channels the curve is normal within the precision of the measuring apparatus. Hence, the normal distribution may justifiably be used for any value of $n > 64$.

Further significance is accorded the above data by plotting the ratio of the voltage exceeded a fraction ϵ of the time to the single-channel rms voltage, as a function of the number n of active channels, for several fixed values of ϵ. From the data given, points on such

curves can be obtained for $n = 1, 4, 16, 64$; furthermore, the fact that the distribution for $n > 64$ is normal permits drawing the asymptote for large values of n. The points read from Fig. 2 and replotted in this way give the full lines shown in Fig. 3.

In order to make practical use of these curves, it is necessary to know what value of ϵ corresponds to satisfactory performance of the amplifier. Experiments have been conducted on a number of different multi-channel amplifiers, each loaded by various numbers of active channels all at the same volume. It has been found that for low enough values of ϵ, no audible disturbance is produced but that as ϵ is increased by increasing the load on the amplifier, the disturbance falling into a channel not energized increases rapidly to a large value. Two different amplifiers having the same computed load capacity may show noticeable differences in performance in this respect when subject to identical fixed loads of the type being considered, thus indicating the influence of circuit design on the value of ϵ. In general, however, the allowable values of ϵ measured for all of the amplifiers that have been tested lie in a relatively narrow band on either side of the curve for $\epsilon = 0.001$. The broken curve of Fig. 3 represents the

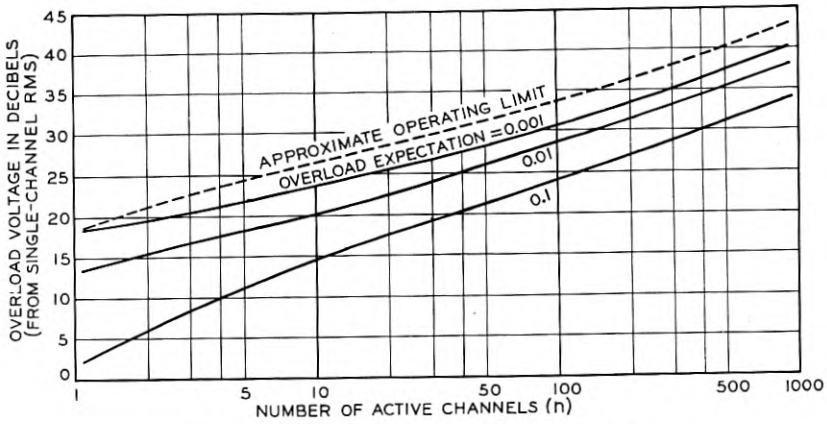

Fig. 3—Overload voltage for n active channels.

approximate upper limit of the observations, extrapolated parallel to the $\epsilon = 0.001$ curve above $n = 14$. It is possible that some amplifiers would overload even if operated in accordance with this curve, but for the great majority of amplifiers of types thus far tested the operation would be satisfactory, with perhaps a small margin.

Multi-Channel Peak Factor

It is useful at this point to introduce the concept of "multi-channel peak factor," which is defined as the limiting ratio of the overload

voltage to the rms voltage for a given number of active channels at constant volume. The ratio of the overload voltage for n active channels to the rms voltage of one active channel is given directly by the broken curve of Fig. 3, and the rms voltage for n channels is simply $\sqrt{n}$ times that for one channel. A simple computation then gives the multi-channel peak factor. This is plotted in Fig. 4 as a function of the number of active channels n. The reduction in multi-channel peak factor as the number of active channels increases reflects the transition from the single-channel distribution curve to the normal curve, as depicted in Fig. 2.

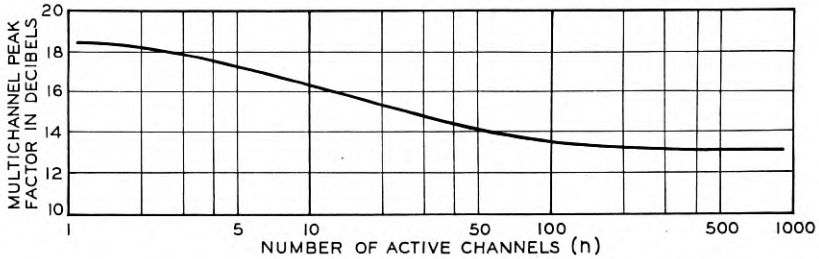

Fig. 4—Multi-channel peak factor for n active channels.

The Distribution of Equivalent Volume

The multi-channel peak factor deals only with the effects of changes in the instantaneous voltages of the channels, all other variables being fixed. It is next necessary to extend the treatment to include the effects of the other load variations that occur in practice—those in number of active channels and in channel volumes. It is important, first of all, to notice that the instantaneous-voltage variations occur very rapidly, while changes in the other two quantities are, in comparison, very slow. In the experiments described above, the loads were so fixed that the equivalent volumes could be changed only by changing the operating transmission level of the amplifier; in practical cases the amplifier transmission level is kept fixed, but the equivalent volume is constantly changing because of changes in number of active channels and in channel volumes.

The amplifier is thus loaded with a constantly changing equivalent volume but because of the great difference in the time-scales of the two classes of variations the load may be regarded as a succession of equivalent volumes, each constant for a small interval of time that nevertheless is long enough to include a representative sample of the resultant instantaneous voltage distribution. If the distribution function for equivalent volume is computed, and then corrected by the

multi-channel peak factor, the fraction of such intervals during which the amplifier will be unsatisfactory from the standpoint of overloading may be determined. For a particular amplifier, the operating transmission level must be so chosen that this fraction will be small enough to make any adverse effects on transmission unimportant. For systems of very many channels the proper value of this fraction is probably about 1 per cent. During the busiest hour, this corresponds to 36 seconds during which audible interference *may* occur and as this will be broken up into many very short intervals, the total effect should be slight. For systems of very few channels, the equivalent volume may reach objectionably high values during these intervals and it might be necessary to make this fraction smaller than 1 per cent to secure good performance. For illustrative purposes, the 1 per cent figure will be used in what follows without implying that it may not need alteration in some cases. The methods used are applicable no matter what value is chosen for the fraction of time overloading is permitted.

Controlled Volumes

As the simplest case to which the above procedure may be applied, and one that may occasionally be of practical interest, consider a commercial system with all the channels controlled to the same volume. If there are N channels in the system, the probability that exactly n channels will be active at any given time is given by equation (1), with $\tau = 0.25$. By computing the value of $p(n)$ for all values of n, and taking the cumulative sum, the value of n which makes the sum 0.99 (or the next greater n) is readily determined. This determines the number n of active channels that is exceeded 1 per cent of the time. A plot of these values of n is given by the curve of Fig. 5, as a function of N, the number of channels in the system. For small values of N this curve has been drawn in a manner to smooth out the steps introduced because n must of necessity be an integer and when the value of n read from the curve is not an integer, the next higher integral value is to be used. It is of interest to compare this curve with the two straight lines of the figure. The lower straight line represents the asymptote for sufficiently large N and the upper straight line is for the condition where all channels are active simultaneously ($n = N$).

The average power for n channels is n times that of one channel, and the equivalent volume expressed in db is $10 \log_{10} n$ above that in one channel. The equivalent volume may thus be computed as a function of n, and by means of Fig. 5, as a function of N. Curve A of Fig. 6 shows the values of equivalent volume so determined as a function of N, the number of channels in the system; it applies specifically to the

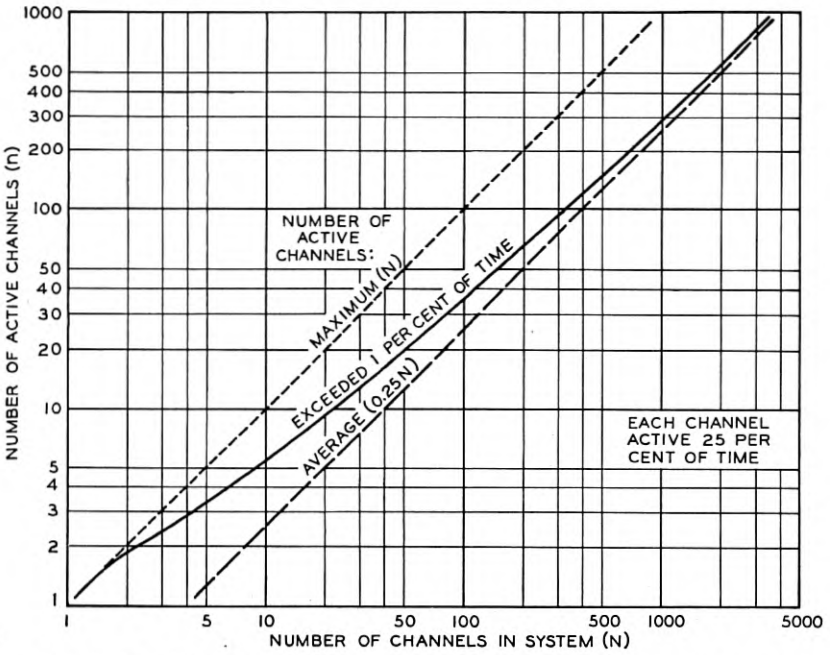

Fig. 5—Number of active channels as a function of the number of channels in the system.

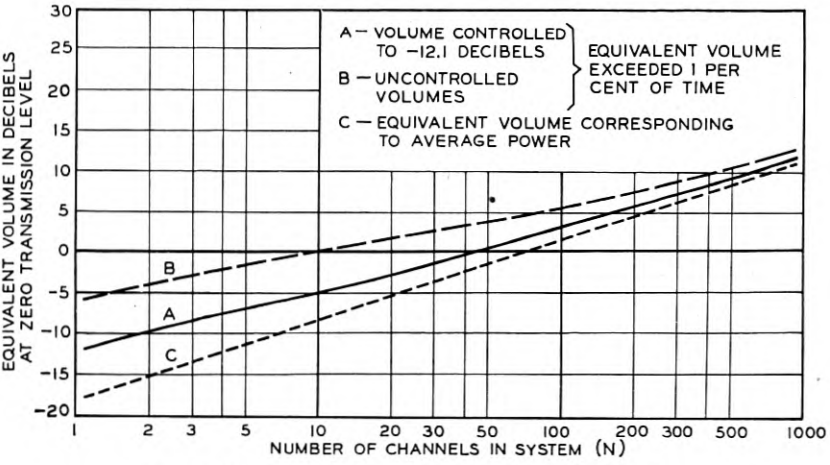

Fig. 6—Equivalent volume for systems of N channels.

case where the volume of each of the active channels is controlled so as to be 12.1 db below reference volúme. The choice of this particular volume is purely arbitrary, but it corresponds to the average power of the single talker volume distribution.

The equivalent volumes given by curve *A* of Fig. 6 are a measure of the average power of the *N* channels, as computed by means of equation (2). To determine the required instantaneous load capacity of the system, the average power must be corrected by the multichannel peak factor which is read directly from Fig. 4, using for the number of active channels the values read from Fig. 5.

For design purposes, it is more convenient to use the rms power of the single frequency test tone whose peak value represents the

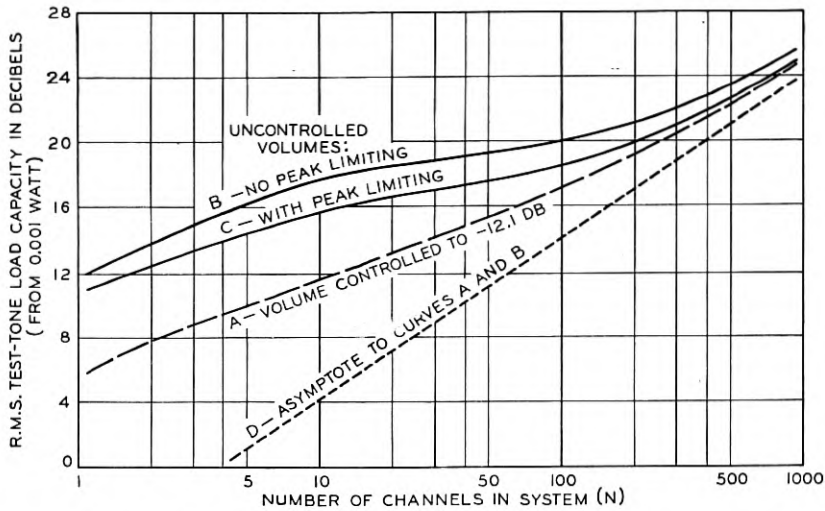

Fig. 7—Load capacity for systems of *N* channels.

instantaneous load capacity. As the ratio of the peak to rms power of a single frequency tone is 3 db, this test power is obtained by subtracting 3 db from the instantaneous load capacity. This required test-tone capacity is plotted as a function of *N* in curve *A* of Fig. 7, which gives the output capacity required for an *N*-channel system with volume control as specified above.

Uncontrolled Volumes

For systems in which volume control is not used, the application of this procedure becomes more involved. To study this more general case, it is convenient first to interchange the conditions of the preceding section, letting the number of active channels be fixed at any value *n*

and examining how the distribution curve of equivalent volume may be obtained for this fixed number of channels. The relation between volume and average speech power given in equation (2) may be rewritten for this case in the form

$$V_i = 10 \log_{10} \frac{W_i}{W_0} \text{db},$$

where $W_0 = 1.66$ milliwatts, W_i is the average speech power in milliwatts, and V_i is the volume in db for any one of the active channels, all at a point of zero transmission level.

Likewise, the relation between equivalent volume and average speech power for n active channels is given by the expression

$$V = n\text{-channel equivalent volume} = 10 \log_{10} \frac{\sum_1 W_i}{W_0} \text{db}.$$

Since the distribution of the channel volumes V_i is known and the volumes of the various channels are independent, the straightforward procedure to obtain the distribution of the n-channel equivalent volume V would involve the following steps: (1) the obtaining of the distribution function of W_i by a transformation of that of V_i; (2) the calculation of the distribution function for the quantity $Y(n) = \sum_1^n W_i$; (3) the transformation of the $Y(n)$ distribution to that of V by inverting the process used in step (1).

The difficulties in this process are all in the second step, where, having given $p_1(W)$, the distribution of average powers for a single channel, it is required to obtain $p_n(Y)$, the distribution for n active channels, with Y defined in terms of W by the relation given immediately above. The formal solution requires the evaluation of integrals of the following type:

$$p_n(Y) = \int_0^Y p_{n-k}(W)\, p_k(Y - W)\, dW.$$

By successive calculation of such integrals for $n = 2, 4, 8 \ldots$, taking k each time equal to $n/2$, the required distributions may be obtained for the necessary range of values of n.

As in the case of the instantaneous voltage distributions, it has not proved feasible to perform the integrations analytically. It was necessary to resort to numerical evaluation of these integrals; by combining the transformations in steps (1) and (3) with the process

of evaluating the integral, the process was somewhat shortened. In this way equivalent volume distributions have been obtained for $n = 2, 4, 8, 16 \ldots$; needed points on the distribution curves for intermediate values of n are obtainable by interpolation.

The accuracy of such a process depends upon the number of division points used in the numerical integration and this as a practical matter must be kept fairly small. When the process must be repeated many times, the errors introduced at each step may accumulate and lead to inaccuracies for large n. It is thus desirable to have some control over the accuracy other than by repeating the calculation with a larger number of division points. This is provided by calculating the moments of $p_n(Y)$ from those of $p_1(W)$ without the use of numerical integration.

The moments S_k of $p_1(W)$ are defined by

$$S_k = \int_0^\infty W^k p_1(W)\, dW,$$

and the moments $T_k^{(n)}$ of $p_n(Y)$ similarly. By the use of the semi-invariants of Thiele,[2] it may be shown that

$$T_1^{(n)} = nS_1,$$
$$T_2^{(n)} = nS_2 + n(n-1)S_1^2, \text{ etc.}$$

By comparing the moments of the distributions obtained by numerical integration with those calculated in this way, and making occasional minor alterations in the curves to bring the first and second moments into agreement, assurance was obtained that all the distributions used are reasonably accurate, with no accumulation of error as n becomes large.

Examples of the cumulative distribution curves of equivalent volume for 1, 4, 16 and 64 active channels are given in Fig. 8. The decrease in standard deviation which occurs as n increases is of interest for it indicates how the fluctuations in load due to talker volume variations are reduced by combining a large number of channels in one system.

Having now n-channel equivalent volume curves for a range of values of n, the resultant equivalent volume curves may be calculated when the restriction to a fixed number n of active channels is removed. Let $p_n(V)$ denote the probability that, with n channels active, the equivalent volume lies between V and $V + dV$ and let $p(n)$ denote the probability that just n channels *will* be active. Then the total proba-

[2] T. N. Thiele, "The Theory of Observations," 1903; reprinted in *Annals of Mathematical Statistics*, Vol. 2, 1931. See especially Sections 22, 29.

bility that the equivalent volume will be between V and $V + dV$ is given, for an N-channel system, by

$$p(V) = p(1)p_1(V) + p(2)p_2(V) + \cdots + p(N)p_N(V).$$

The $p_n(V)$ are given by equivalent volume curves such as those in Fig. 8 and the $p(n)$ by equation (1). Examples of curves thus computed are given in Fig. 9, which shows the equivalent volume distributions at a point of zero transmission level for 3, 12 and 240 channel systems. The equivalent volume that is exceeded 1 per cent of the time, read from such curves, is plotted as curve B of Fig. 6.

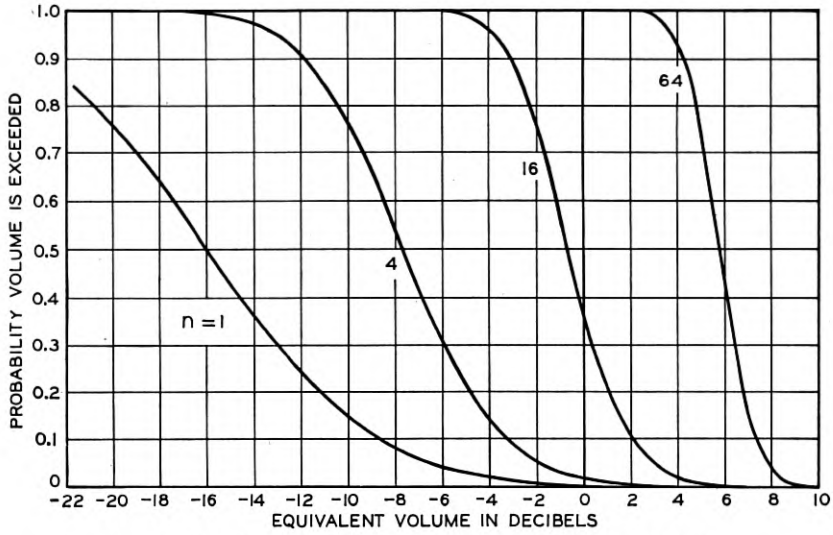

Fig. 8—Equivalent volume distributions for n active channels.

This curve gives, for any number of channels having uncontrolled volumes, the equivalent volume which will be exceeded just 1 per cent of the busy hour. To obtain the necessary load capacity, this must be corrected for the multi-channel peak factor. In the controlled volume case, for a given number N of channels in the system, there was no difficulty in deciding the value of n, the number of simultaneously active channels, for which the multi-channel peak factor should be taken. Now, however, there is no unique relation between equivalent volume and the number n; in addition, the multi-channel peak factors were measured with all n channels at the same volume, which represents a condition rarely holding on a system without volume control. It is apparent, however, that in the majority of cases in which the equivalent volume approaches values on curve B of Fig. 6, the number

of simultaneously active channels will be greater than the average number $N\tau$ of active channels. Since the multi-channel peak factor decreases as n increases, the peak factors for $n = N\tau$ active channels may be safely used. A more detailed analysis, feasible only for very small systems but avoiding the use of this approximation, shows that its effect is small and tends to give load capacities slightly higher than actually required, but the difference diminishes rapidly as the size of the system is increased.

For the uncontrolled volume condition, therefore, the multi-channel peak factors are read from Fig. 5 for values of $n = N\tau$. They are added to the equivalent volumes obtained from curve B of Fig. 6, and

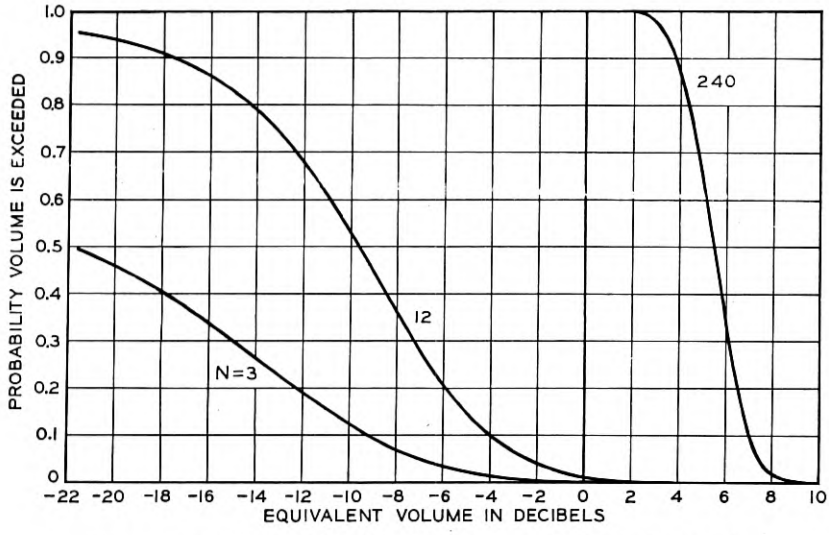

Fig. 9—Equivalent volume distributions for systems of N channels.

reduced to single frequency power as previously described for the volume controlled case. Curve B of Fig. 7 is obtained in this manner and shows the load capacity required in an amplifier for an N-channel system in which the volumes of each channel are distributed in accordance with curve A of Fig. 1. The load capacity which is approached asymptotically as N increases indefinitely is represented by curve D of Fig. 7.

The load capacities given by Fig. 7 are valid only for systems for which the basic single-channel data apply. As these may not hold in specific cases, and may be subject to modification in the future, estimates of the effects of small changes in these data are useful. These effects cannot be described simply for moderate numbers of

channels but for large numbers of channels the effects are readily estimated from the change in the location of the asymptote shown on Fig. 7. The equation of this asymptote is as follows:

$$L = 10 \log_{10} N\tau + (V_0 + .115\sigma^2) + MPF + P_0 - 3 \text{ db},$$

where L = test tone load capacity,
MPF = asymptotic multi-channel peak factor,
P_0 = long average power of a reference volume talker in db above .001 watt.

The other quantities are as defined before.

Peak Voltage Limiting

The curves referred to in the preceding discussion have so far neglected the effects of peak voltage limiting in the transmitters and in the channel terminal equipment. Fundamentally, the effect of such limiting is to modify the distribution of instantaneous voltages in the individual channels. The extent of the modification, however, depends on the volume. For single-channel systems it is obvious that the improvement in load capacity due to limiting will be substantially equal to the reduction in the maximum peak voltage. For a large number of channels the improvement will approach the reduction in the rms voltage per channel. An approximate method of accounting for these complicated reactions is to consider that peak voltage limiting modifies the upper end of the single-channel volume distribution. Strictly the amount of such modification is a function of the number of channels as well as of the characteristics of the limiters. Curve *B* of Fig. 1 represents a compromise between the different effects which is believed to give reasonably accurate results for both small and large numbers of channels for the limiting characteristic of present terminals.

With the talker volume distribution modified in accordance with curve *B* of Fig. 1, computations of the load capacity with voltage limiting present may be made in a manner identical with that previously described. Curve *C* of Fig. 7 shows the results obtained for this amount of limiting.

All of the load capacity curves of Fig. 7 are based on the equivalent volume which would be exceeded 1 per cent of the time, irrespective of the number of channels in the system. Where voltage limiting is used, it appears reasonable to consider this percentage as fixed because the action of the limiters serves to restrict the range of voltages above the overload point, thus reducing the severity of any overloading effects. When there is no limiting, and particularly

for a small number of channels, the range of overloading voltage is not so restricted and overloading effects may become undesirably severe during the 1 per cent of time when the overload voltage is exceeded. If voltage limiting is not provided in some form, it may be important to reduce the percentage of time during which overloading may occur for small numbers of channels. This is a matter to be determined by experience and, if necessary, would require modification of curve *B* of Fig. 7 in the direction of requiring more load capacity for a small number of channels, thus increasing the spread between curves *B* and *C*.

Operating Margins, etc.

The curves which have been given for output capacity versus number of channels apply to a single amplifier, or to a system in which all amplifiers are identical and work at the same output level without appreciable impairment of overall performance. In practice, the number of amplifiers in tandem in a long system may be very large and problems of equalization and regulation may make it difficult to maintain exactly the same level conditions at all amplifiers. In addition, aging of tubes, and other effects will introduce some impairment. It is important, therefore, to allow a margin for these effects in the design of an amplifier for a multi-channel system. The proper margin is essentially a matter of system design and it is often economical to build a liberal margin into the amplifiers in order to allow greater latitude and economy in the design of equalizing and regulating arrangements.

In addition to the speech loads, there are also impressed on the amplifiers various signaling and pilot frequencies, carrier leaks, etc. It is not always possible in practice to make these negligibly small and the load capacity requirements must be corrected to allow for their presence. Multi-channel telephone systems are also required to transmit other types of communication circuits, such as program channels and voice-frequency telegraph systems, superposed on one or more telephone channels. Modifications of the methods applied to speech loads may readily be made to determine the effect of these on the amplifier load capacity.

Acknowledgment

Many members of the Bell Telephone Laboratories, in addition to those mentioned in the text, have contributed to various phases of this work. The authors take this opportunity to acknowledge their indebtedness to these colleagues, and in particular to Dr. G. R. Stibitz, who first developed the theoretical approach here used.

The Quantum Physics of Solids, I
The Energies of Electrons in Crystals

By W. SHOCKLEY

It is proposed to make this paper the first of a series of three dealing with the quantum physics of solids. This one will be concerned with the quantum states of electrons in crystals. The discussion will commence with an introductory section devoted to the failure of classical physics to account for phenomena of an atomic scale. Next, the quantum theory of electrons in atoms will be discussed, together with the resultant explanation of the structure of the periodic table; this is designed to illustrate the meaning of various quantum mechanical ideas which are important in understanding solids. Furthermore, much of the detailed information about atomic quantum states of particular atoms will be needed in the later discussion of the properties of certain solids. As an introduction to the modification of the quantum states occurring when atoms are put together to form a crystal, a short section will be devoted to structure of diatomic molecules. The next section will be concerned with quantum states for electrons in crystals. Whereas in an atom there are a series of isolated energies possible for an electron (corresponding to the various quantum states), in a crystal there are bands of allowed energies. This concept of energy bands is essential to the theory of crystals in much the same way that the concept of energy levels is essential to that of atoms. In terms of energy bands, the energy holding crystals together can be interpreted on a common basis for a wide variety of crystal types. This will be followed by a brief description of various crystal types and by a discussion of thermal properties in which the smallness of the electronic specific heat will be shown. The last section will be devoted to a discussion of para and ferromagnetism on the basis of the energy band picture.

In the second paper, problems connected with electric currents and the motion of electrons through crystals will be discussed. This leads to the concept of the Brillouin zone which is complementary to that of the energy band, the two together forming the basis for discussing the quantum states of electrons in crystals.

The third paper of the series will contain a comparison between theory and experiment for the alkali metals, the principal emphasis being placed upon the physical picture of the state of affairs in these simple metals.

Introduction

"THE parts of all homogeneal hard Bodies which fully touch one another, stick together very strongly . . . I . . . infer from their Cohesion, that their Particles attract one another by some Force, which in immediate Contact is exceeding strong, at small distances

performs the chymical Operations above mention'd, and reaches not far from the Particles with any sensible Effect. . . . There are therefore Agents in Nature able to make the Particles of Bodies stick together by very strong Attractions. And it is the Business of experimental Philosophy to find them out." But it was not destined for experimental philosophy to finish the business which Sir Isaac Newton set for it in the above words [1] until two centuries had elapsed. Only since the advent of quantum mechanics have scientists had laws capable of explaining the cohesive forces of solid bodies and predicting their numerical magnitudes. The new laws were developed first in order to explain the behaviors of independently acting atoms but, as we shall see, they are laws capable of extension to systems containing large numbers of atoms and thus to solid bodies. The fact that a solid body remains a solid body, resists being pulled apart, and exerts the cohesive forces of which Newton wrote, is explained by showing from theory that atoms packed together in a solid are in a state of low energy, and to change the state requires the expenditure of work. In this paper we shall describe how the quantum mechanical concepts developed for isolated atoms are applied to interacting atoms and lead to methods of calculating the energies and forces binding atoms together in crystals.

A crystal is a regular array of atoms. The regularity of this atomic array is frequently exhibited in the macroscopic appearance of the crystal. A crystal of potassium chloride—sylvine—is a good example (Fig. 1A). The natural growth faces of the crystal are parallel to planes passing through the atoms, which are arranged in the microscopic array pictured in Fig. 1B. It is evident that the microscopic arrangement of the atoms in the crystal is one of its most basic features. In sylvine the atoms are arranged on the corners of cubes in an alternating fashion. The arrangement of the atoms in the crystal is called a "lattice." Sodium chloride—rock salt—has the same arrangement as sylvine and the type lattice pictured in Fig. 1B is known as a "sodium chloride lattice." The distance between atoms in a given lattice is specified by giving the value of the "lattice constant," which for a cubic crystal is defined as the distance between like atoms along a line parallel to a cube edge. Lattice constants are usually expressed in angstroms; 1 angstrom $\equiv$ 1A $= 10^{-8}$ cm. The lattice constant of sylvine, designated by "a" in Fig. 1B, is 6.28A. Figure 1C suggests how a large number of atoms, arranged as in Fig. 1B, produce the shape of the crystal photographed for Fig. 1A. Studies of the directions of the natural growth faces and cleavage faces of crystals are

[1] "Opticks" 3rd ed., 1721, p. 363.

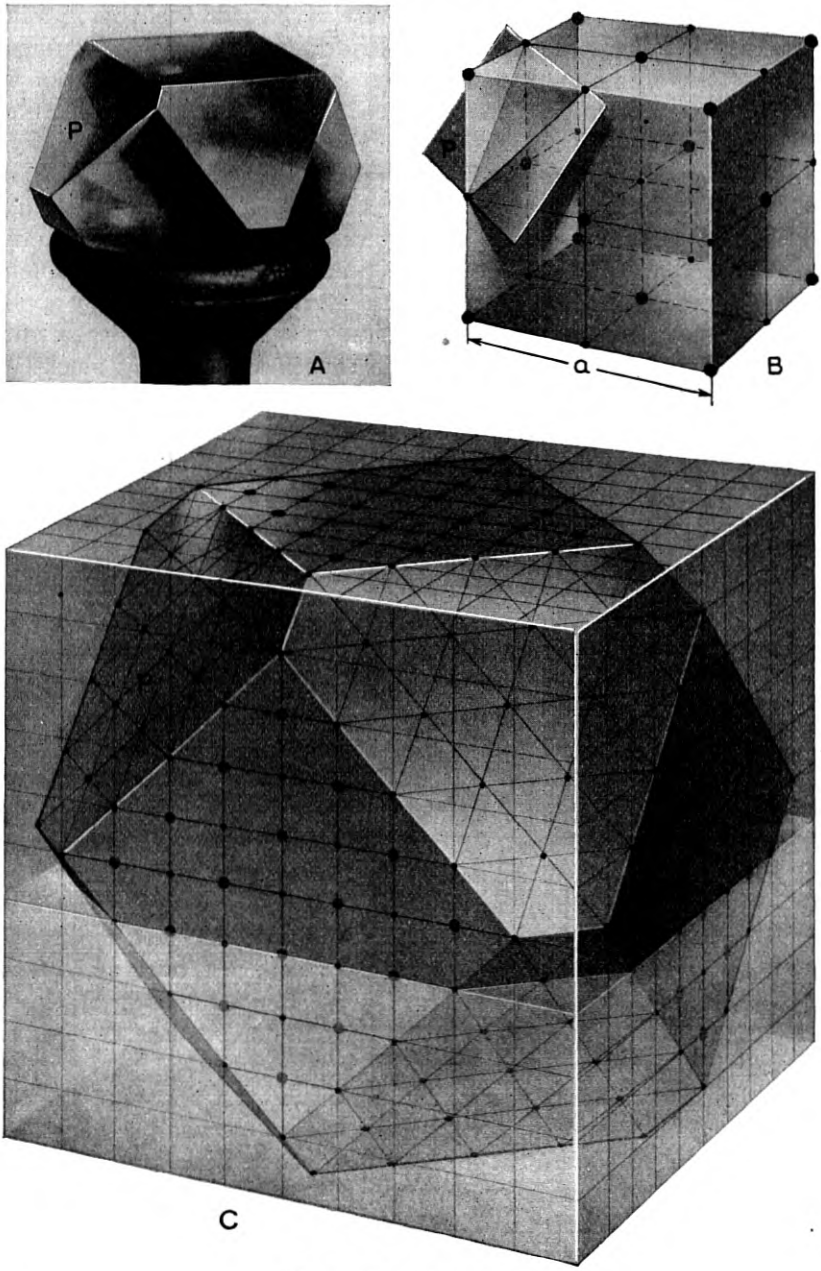

Fig. 1—Crystal structure.

A. Macroscopic appearance of a crystal; retouched photograph of a sylvine crystal.
B. Microscopic arrangement of atoms in crystal showing natural planes.
C. Large number of atoms arranged as in B to show formation of A.

primarily of importance as an aid in classifying and identifying minerals; and although they do give some information about the arrangement of the atoms in planes within the crystal, the information is too meager to permit a determination of the microscopic structure. The latter can be deduced by the methods of x-ray diffraction. X-rays are light waves of very short wave-length and they are diffracted from crystals in much the same way as light is diffracted from a ruled grating. From studies of x-ray diffraction patterns, the arrangements of atoms in a large number of crystals have been determined. Exceedingly strong forces act to hold the atoms in these arrangements and, by application of the laws of quantum mechanics, we shall try to find them out.

There is now no question that the elementary building blocks of the material world are primarily electrical and of two sorts.[2] The negative particles, electrons, are all alike and have the same charge $-e$ and the same mass m; the positive particles, atomic nuclei, are not all alike and may differ in charge and mass from one another. The positive charge is always some integral multiple, Z, of the fundamental charge e and we shall not be concerned with the mass except to say that it varies upwards from about 2,000 times the electron mass. An atom of a chemical element consists of one nucleus surrounded by enough electrons to neutralize its charge; all atoms of a given chemical element have the same nuclear charge, Z, which is appropriately known as the "atomic number"; atoms having the same nuclear charge but different masses are called "isotopes"; their chemical behaviors are slightly different, so that it is possible by chemical processes to separate one isotope of a chemical element from the others, but this difference is so slight that we can neglect it here. An atom, then, consists of a number of electrons circulating about and attracted by the nucleus, which, by virtue of its relatively great mass, is effectively an immobile center for their motions. A simple molecule consists of an assemblage of a few such atomic systems and a crystal of an immense number. The fundamental problem of atomic mechanics—which is now solved quite satisfactorily but not yet perfectly—is to find the laws governing the motion of these particles.

The necessity of finding such laws is made most apparent by considering the failure of the older laws of "classical mechanics," Newton's laws. These laws were satisfactory for dealing with large bodies—but not perfect; for, as is well known, they are approximations to the more adequate laws of relativity—and they were successfully applied

[2] Since we are here concerned with problems of a chemical nature, we may disregard those particles such as positrons, mesotrons, neutrons, etc., which are concerned with cosmic rays and nuclear processes but not with ordinary atomic behavior.

even to single atoms so long as no attempt was made to investigate the internal structure of the atom. Considering the atom to be a perfectly elastic miniature billiard ball having size, mass, and velocity but no internal properties, classical mechanics was able to handle in a statistical fashion the dynamics of large systems of atoms in a gaseous form and to deduce a number of valid conclusions concerning the specific heat, gas laws, viscosity, and diffusion constants of gases. On the other hand, failure attended all endeavors to apply these laws to the swarm of electrons surrounding a nucleus. A system of this sort is unstable classically and can never come to thermal equilibrium. Applying the classical laws of statistical mechanics, one finds that some of the electrons will move very close to the nucleus, the energy lost in this process being acquired by other electrons which move farther out. According to classical mechanics this process will continue without ever reaching equilibrium and during it the atom will be thoroughly torn apart.

Another difficulty in the classical theory arises from the electrodynamics of an accelerated electron. An electron moving in the field of a nucleus is accelerated, and classical electromagnetic theory predicts that under these circumstances electromagnetic energy will be radiated—the atom being in effect a microscopic radio transmitting station in which the charging currents in the antenna are represented by the motions of the electrons. According to this theory an atomic system would continually radiate energy, and it could be proved that no equilibrium like that actually observed between matter and radiation would ever be achieved.

Thus, classical mechanics and electromagnetics were incapable of taking the electrons and nuclei as building blocks and constructing solids or even atoms from them. To put it bluntly, the classical laws were wrong; although adequate for large-scale phenomena, they were inapplicable to phenomena of an atomic scale.

Nevertheless, modified applications of the classical theory had a great number of successes in the atomic theory of solids. Dealing with the atoms as elastic idealized billiard balls led to the correct value for the specific heat of solids, at least at normal temperatures, and the electron theory of conduction in metals was in many respects quite successful. None of the successes of the conduction theory were completely satisfying, however, because the assumptions needed to explain one set of facts were incompatible with other sets of facts and the whole field was greatly lacking in unity. According to this classical theory a metal contained free electrons which could move under the influence of an electric field and thus conduct a current. Their motion was

impeded by collision with the atoms (ions, really, since they are atoms which have given up free electrons) according to some theories, and with the spaces between atoms according to other theories, and this impeding process gave rise to electrical resistance. The free electrons were capable of conducting thermal as well as electrical currents. Although the theory gave reasonable values for the electrical and thermal conductivities of metals at room temperature, the predicted dependence upon temperature was wrong: the resistance of a pure metal is known from experiment to be very nearly proportional to the absolute temperature; the classical theory, unless aided by very unnatural assumptions, predicted proportionality to the square root of the absolute temperature. Another difficulty, the greatest in fact which beset the old theory of free electrons in metals, was concerned with the specific heats of metals. According to the billiard ball theory of gases, the specific heat arose from the kinetic energy of motion of the gas atoms; thus the specific heat at constant volume of one gram atom of a monatomic gas was $(3/2)R$, where R is the gas constant. This was in good agreement with experiment. For solids this specific heat was just doubled, giving $(6/2)R$ because of the addition of potential energy to the kinetic. For a metal the free electrons were regarded as having kinetic energy. In order to explain the observed electrical properties of a metal, the number of electrons was taken as approximately equal to the number of atoms. Hence, as for a monatomic gas, a specific heat of $(3/2)R$ was expected for the electron gas and, therefore, a specific heat of $(9/2)R$ was predicted for a metal. Measurement shows that most crystals, metals included, fit quite well the value of $(6/2)R$ and that $(9/2)R$ is incorrect. Thus classical theory was left with the dilemma that to explain electrical properties one free electron per atom was needed while to explain specific heat one free electron per atom was far too many. This dilemma is very neatly resolved in the new theory; in this paper we shall show why the free electrons are not free for specific heat and in a later paper why they are free for conduction. We shall also show that the new theory leads to quite proper values for the conductivity and also explains facts concerning the resistance of alloys, which the classical theory could not do.

According to the classical theory there was one quantity that should be the same for all metals and this was the ratio of the thermal to the electrical conductivities. This ratio, known as the Wiedemann-Franz ratio, was predicted to be equal to the absolute temperature times a universal constant L called the Lorentz number. This prediction was in reasonable agreement with experiment. The new wave me-

chanical theory predicts the same result, but with a slightly different value for L. According to the old theory $L = 2k^2/e^2 = 1.44 \times 10^{-8}$ volts^2/degree^2 where k is Boltzmann's constant, while the new gives $L = \pi^2k^2/3e^2 = 2.45 \times 10^{-8}$ volts^2/degree^2, and the experimental values for several elements are Cu 2.23, Ag 2.31, Au 2.35, Mo 2.61, W 3.04, Fe 2.47—all times 10^{-8} volts^2/degree^2. We see that the constancy of the Lorentz number predicted by both theories is in reasonable agreement with experiment, but that in predicting the numerical value of the constant the new theory is better than the old.

The fundamental problem of how the electrons and nuclei form stable atoms and crystals was, as we have said above, inexplicable on the older theory. The newer quantum mechanics of Bohr and later that of Schroedinger, Heisenberg, and Dirac were needed. Bohr postulated that out of the infinity of possible motions for the electrons of an atom, only a certain restricted set was permitted. Each permitted motion corresponded to a definite energy for the atomic system as a whole. This concept of energy levels for the atom gave a natural interpretation to nature of atomic spectra and explained the meaning of the combination principle. In order to restrict the atomic motions to certain energy levels, Bohr supposed that the laws of atomic dynamics were such that only those modes of motion were permitted for which certain dynamical quantities, called phase integrals, had values equal to multiples of Planck's constant h. For the case of the hydrogen atom these laws led to the now well-known Bohr orbits for the electron and to energy levels which were in good agreement with experiment. For atoms with more electrons it was very difficult to apply Bohr's laws except in a very approximate and unsatisfactory way. However, two very valuable concepts came from his theory which are preserved in the newer wave mechanical theory. These were that the individual electrons could be thought of as restricted to certain orbits and that these orbits were specified by giving them certain quantum numbers. It was found that three quantum numbers were needed to specify the orbit. All atoms were found to have the same general scheme of orbits. The number of electrons moving in these orbits varies from atom to atom and for any given atom is equal to the atomic number Z. In order to explain the facts of spectroscopy and the periodic table of the elements, it was necessary to introduce a rule known as Pauli's principle. This principle states simply that no more than two electrons may occupy the same orbit in an atom; that is, no more than two electrons of an atom may have the same three quantum numbers. As we shall discuss in the next section, a complete specification of the state of an electron in an atom requires four quantum numbers; two

electrons in the same orbit have different values for their fourth quantum number. We shall use the term "quantum state" to signify the permitted behavior corresponding to specified values for the four quantum numbers. In this language, Pauli's principle asserts simply that no two electrons in a given atom can be simultaneously in the same quantum state; that is, Pauli's principle is a quantum mechanical analogue of the classical principle that two bodies cannot occupy the same place at the same time. The two ideas—first that the motions of the electrons are quantized so that only certain quantum states are allowed, and second that in an atom only one electron can occupy a given quantum state—form the basis of all quantum mechanical thinking. We shall make use of them continually in the following discussion. We shall use them, however, not in connection with the orbits of Bohr but instead with the wave functions of Schroedinger.

The Bohr theory can be applied only with difficulty to any atom but hydrogen. The difficulty lies in determining the motions of the electrons in the complex interacting fields of the electrons and the nucleus. This problem is even more difficult in the case of a solid where there are many atoms, and it would seem hopeless to try to find out why the electronic orbits in insulating crystals such as rock salt or diamond do not permit electrons to move through the crystal and carry a current, while the orbits in metals do. Indeed not only does the Bohr theory have the foregoing disadvantage but it is probably wrong. Fortunately there is a theory both sounder and easier to apply embodied in the "wave equation of Schroedinger."

One feature, probably not sufficiently stressed, about Schroedinger's equation is its relative convenience. The word "relative" must be used here because it is usually very laborious to obtain solutions for the equation and only in the simplest cases can we obtain exact solutions. Compared to the classical equations and the equations of Bohr, however, it *is* convenient. Quite satisfactory approximate solutions can be obtained for Schroedinger's equation even for the complex case of solids, where it would be prohibitively difficult to obtain as good solutions for the classical and Bohr equations.

Electrons in Atoms

According to the Schroedinger theory, a differential equation can be written down for any system consisting of electrons and atomic nuclei. This equation contains an unknown wave function and an unknown energy and the instructions of the theory are to solve the equation for the unknown quantities. Furthermore, the wave function must satisfy a certain mathematical requirement which embodies

in a generalized form the restrictions imposed by Pauli's principle. As is too frequently the case in mathematical physics, it is much easier to state the problem than to solve it; the solutions of Schroedinger's equation are, in fact, so difficult to obtain that exact solutions have been found for atomic systems only of the simplest type, namely those consisting each of a single nucleus and a single electron. For this case, the quantum states and their energies are all exactly known. For other cases approximations of varying degrees of exactness must be used. The difficulty arises from the interactions between the electrons. If it were not for these interactions, one could obtain exact solutions for atoms having many electrons. The difficulty is that the interactions—they are merely electrostatic repulsions—prevent each electron from being independently in a definite quantum state. The interaction of each electron with another is in general small compared to its interaction with the nucleus. To a first approximation, then, the electrons are treated as not interacting and then corrections are applied to this over-simplified picture. (In this first approximation, the generalized mathematical statement of Pauli's principle reduces to the one we gave in the last section—only one electron may occupy a given quantum state.) As a result of this procedure of over-simplification followed by corrections, our exposition will commence with a discussion of the quantum states of an electron in an atom as if these quantum states were private possessions of the electron and not influenced or disturbed in any way by the other electrons. We shall then correct this picture to some extent by considering how the energy of a given electron depends upon the behavior of the other electrons. One correction term which we shall introduce in this way is the important "exchange energy" discussed below. Thus atomic theory represents a field of endeavor in which further progress is made largely by improvements and refinements. It should be emphasized, however, that the corrections and refinements are not additional assumptions, which are added to the theory, but that they represent instead only steps forward in improving the wave mechanical solutions.

The last paragraph mentions that an approximate treatment of Schroedinger's equation leads to a set of possible quantum states for an electron in the atom. We shall discuss Schroedinger's equation and the wave functions corresponding to the quantum states in more detail later and at present be concerned only with a description of the results. In a neutral atom the electrons arrange themselves in the quantum states in such a way as to make the energy of the atomic system a minimum. Consistent always with Pauli's

principle, only one electron can occupy a particular quantum state. When the atom is in the arrangement of lowest energy, we can say that each electron has a definite energy corresponding to whichever quantum state it occupies. This energy is most conveniently defined in terms of the amount of work required to take an electron from its state in the atom and put it in a standard state defined as zero energy. An electron in the zero energy state is to be thought of as at rest and so far removed from the atom that there is no energy of interaction between them. In this way we can define the energy of every occupied quantum state in the atom. Each of these energies must be taken as negative—since potential energy is yielded up when the electron returns to the atom—and by definition represents how tightly the electron is bound to the atom. One of the electrons will be the most loosely bound (it may be that there are several with the same energy) and the energy required to remove it is called the "ionization energy." From our definition this is obviously the minimum energy required to convert the atom to a positive ion. The definition of the energy of a quantum state given above can be used only when an electron is in the quantum state; we can, however, define the energy of an unoccupied state conveniently in terms of the energy the atom would have if the state were occupied by "exciting" one of the electrons to this state by giving it the proper amount of energy.[2a]

The Quantum States of the Atom

Using this definition of the energy of a quantum state, we find that for all atoms the arrangement of quantum states in energy is as shown in Fig. 2, where the ordinates represent energies and states of equal energy appear as divisions of the horizontal lines. Figure 2 does not indicate which states are normally occupied, nor could it unless we knew how many electrons there were in the atom. The general scheme of Fig. 2 is applicable, with certain changes discussed below in the energy scales, to any neutral atom in its normal state, and the energy

[2a] This definition is subject to restrictions because the energy of an electron in the state in question depends upon the arrangement of the other electrons in the atom and this arrangement depends in turn upon which electron was excited to the initially unoccupied state. In constructing the figures we have supposed that the electron (or one of the electrons in case there are several) that is most easily removed from the atom is caused to shift from its normal state to the unoccupied state in question; this shift will change the state of the atom and since the atom was initially supposed to be in the state of lowest energy, the change in energy cannot be negative and will in general be positive but may in certain special cases be zero. The energy of the unoccupied state is defined as the energy of the occupied state from which the electron is taken plus the change of energy caused by shifting the electron. This is equivalent to saying that the energy of the unoccupied state is the ionization potential of the atom after an electron has been shifted from the highest state normally occupied to the normally unoccupied state in question.

levels represented on Fig. 2 apply to unoccupied and occupied states as well.

I have already mentioned that four quantum numbers are required to specify each quantum state. These are indicated by the letters

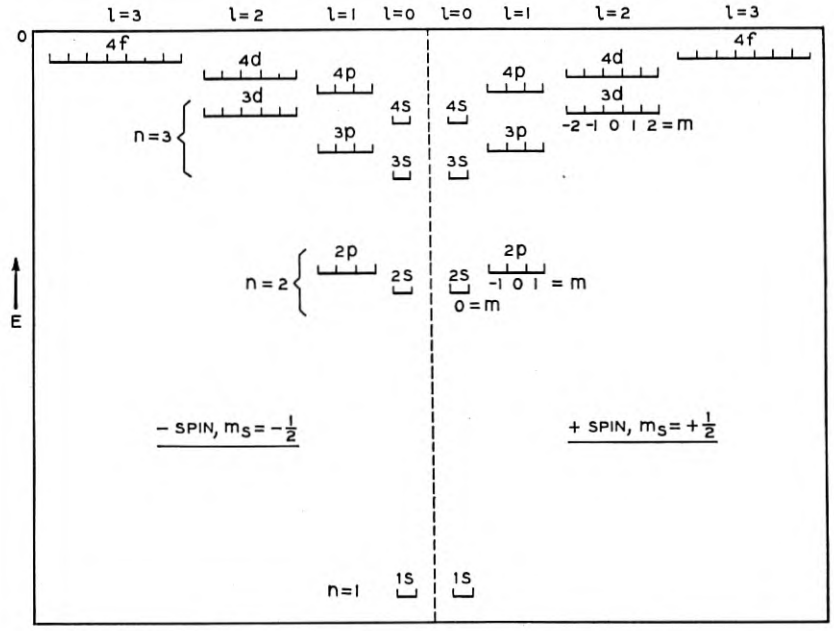

Fig. 2—Quantum states for electrons in atoms.

n, l, m, and m_s. Roughly speaking, the "principal quantum number n" fixes the "energy level" of the state; however, there is some dependence upon the "angular momentum quantum number l." The dependence of energy upon the third quantum number, "the magnetic quantum number m," is slight and will be neglected in this paper. We shall consider the energy to be specified by giving n and l. A notation borrowed from spectroscopy is applied to this pair of quantum numbers and one uses the apparently quite fortuitous choice of letters s, p, d, f, g, h, etc., to stand for $l = 0, 1, 2, 3, 4, 5$, etc., and a state with $n = 3$ and $l = 2$ is known as a "$3d$ state" and an electron occupying such a state is called a "$3d$ electron." The quantum laws permit the following values for n, l, and m:

n takes on all positive integral values. (All states with n greater than four have been omitted from the figure; they lie between the highest states shown and zero energy.)

For a given n, l takes on all positive integer values from 0 to $n - 1$ inclusive.

For a given n and l, m takes on all integer values including zero from $-l$ to $+l$ inclusive.

The difference between right and left sides of the figure corresponds to the fourth quantum number: an electron, in addition to its electric charge, possesses angular momentum or "spin" about its axis. The rotating charge resulting from this angular momentum produces a magnetic moment. The angular momentum is quantized and there are two possible values $+1/2$ and $-1/2$ for the "spin quantum number m_s," corresponding to the right and left halves of Fig. 2. Electrons occupying states on the right half of Fig. 2 have their spins parallel to each other and directly opposite to electrons occupying states on the left half. As already implied, the quantum numbers l and m also correspond to angular momentum and magnetic moments for the electron "orbits" (really wave functions) in the atom.[3]

For our purpose we need two results of the theory of the spinning electron, first that its *spin introduces a duplicity of quantum states* as indicated by the two halves of Fig. 2, and second that *all the electrons of one spin have their magnetic moments parallel and opposite to those of the other spin.* Later when we consider the question of magnetism, we shall be concerned with the direction in space of the spin vector and the magnetic moment, but not now.

Several units of energy are employed in describing atomic processes. The simplest of these is the electron volt; it is the energy acquired or lost by an electron in traversing a potential difference of one volt. For example, in a vacuum tube operating with one hundred volts between cathode and plate, the electrons strike the plate with a kinetic energy of one hundred electron volts, 100 ev. Another unit is the ionization potential of hydrogen, and as hydrogen has only one electron, which normally occupies the 1s state, this is also the energy of the 1s state. This energy is called the "atomic unit" of energy or the "Rydberg." Another unit of energy useful in chemical processes is the kilogram calorie per gram atom; this is related to the others as follows: if the energy of each atom in one gram atom is increased by one electron volt then the energy of the whole system is increased by 23.05 kilogram calories. The conversion factors are: 1 Ry = 13.5 ev, 1 ev per atom = 23.05 Kg.-cal./gm. atom.

[3] For a discussion of the quantum states of the electrons from the point of view of angular momentum see "Spinning Atoms and Spinning Electrons" by K. K. Darrow, *Bell System Technical Journal*, XVI, p. 319, or standard texts on spectroscopy.

Variation of the Energy Levels with Atomic Number

All atoms have the same general scheme of quantum states indicated in Fig. 2. Quantitatively the energy scale varies from atom to atom. Thus the 1*s* state lies at $-$ 13.5 ev for hydrogen and at $-$ 24 ev for helium. This decrease (i.e., becoming more negative or moving lower down on Fig. 2) is due to the increase in nuclear charge, $Z = 1$ for hydrogen and 2 for helium, which results in greater attraction and tighter binding for electrons in helium. This steady downward motion of the levels continues as one goes from element to element in the periodic table. However, the ionization potential, the energy required to remove the most easily removed electron, does not steadily increase. In Fig. 3 we show the ionization potentials of the first twenty elements.

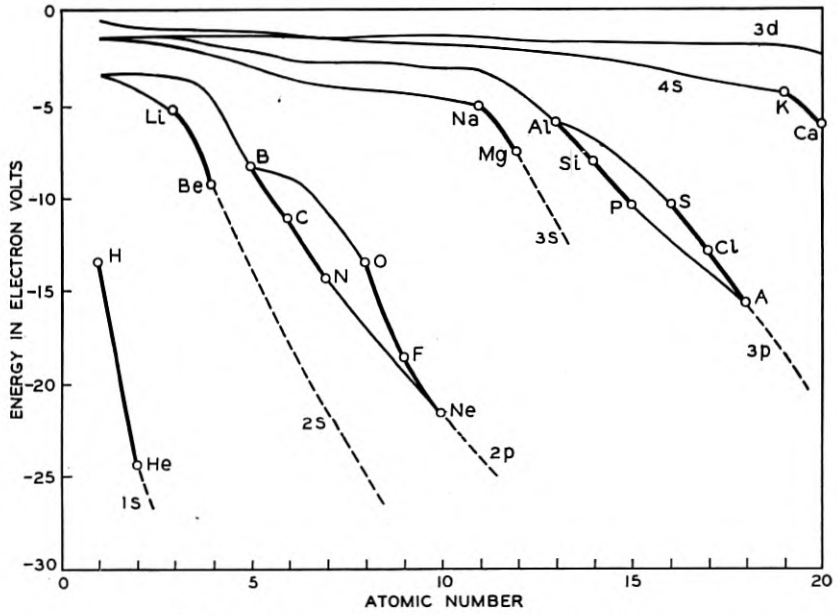

Fig. 3—Ionization potential versus atomic number.

Since we are interested in the energies of electrons rather than in ionization per se, the ionization potentials have been plotted as negative giving in this way the energies of the states in the atom. The main features of this figure can be explained by using Fig. 2 and the Pauli principle.

The Pauli principle, also known as the exclusion principle, permits only one electron to occupy each of the states of Fig. 2. The electrons in a many-electron atom will tend to go to the states of lowest energy.

Thus in helium, since its two electrons can have oppositely directed spins, each fills one of the $1s$ states; we say the "electron configuration" of helium is $1s^2$ (read as "one ess squared"). For lithium, $Z = 3$, the third electron, which cannot go to the completely filled $1s$ states, goes to the next highest, $2s$, giving $1s^2\,2s$. In going from helium to lithium, all the states move to lower energies but not so much lower as to make $2s$ for lithium as low as $1s$ for helium. For this reason lithium can be relatively easily ionized, as is seen in Fig. 3.

Before continuing the discussion of particular atoms, we must point out that two changes accompany each advance from one element to the next in the periodic table. In each step the nuclear charge increases by one plus unit and at the same time an electron is added to the atom and the combined effects produce the results of Fig. 2. Quite different results are obtained if one electron alone is added to the atom. Then instead of the general falling of the levels which accompanies the double change, there is a general rising of all the levels. This is due to the unbalanced negative charge on the added electron, whose presence on the atom raises the potential energy of all the electrons and therefore raises their energy levels. For some atoms, the raising of the energy levels produced by an unbalanced electron may be so great that the electron is not bound at all or at least only very slightly, and for these atoms negative ions do not form. On the other hand, when an electron is removed from an atom all the remaining electrons become more tightly bound and the energy levels are lowered.

Exchange Energy

In Fig. 4 we show the electron configurations for the elements from lithium to neon. The decrease in ionization potential in going from beryllium to boron is due to the completed filling of the $2s$ states and the consequent start of filling of the $2p$ states. The decrease in going from nitrogen to oxygen suggests that not only do the $2s$ and $2p$ states lie at different levels but that the $2p$ states themselves lie at two different levels. This is true but in a rather special sense: *the difference in energy between the two sets of 2p states depends upon how they are occupied.* This difference is an "exchange energy." We shall discuss the origin of the exchange effect in the next paragraph but one; however, the aspect of it needed for this paper is illustrated in Fig. 4. We there imagine that the quantum states are represented by little trays upon which are placed weights to represent occupancy by electrons. The exchange effect corresponds to hanging the trays on springs; in this way we see that as the electrons fill up the $2p$ states

with one spin (the same effect occurs for either spin; the figure shows + spin), these states are depressed in respect to the $2p$ states with the other spin. The springs must, however, be considered to pull the

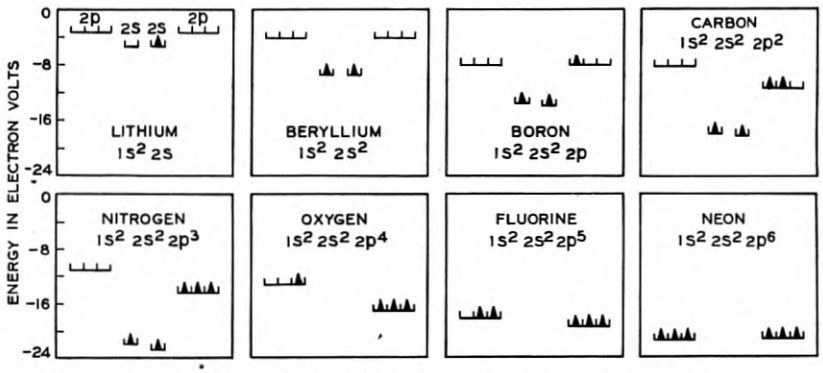

Fig. 4—Electron configurations illustrating the exchange effect.

trays up against stops with a force such that a single weight upon a tray will produce no lowering whereas two or three weights will. This effect seems contradictory to the simple idea that adding electrons raises the potential energy and the energy levels; however, it must be remembered that we are here discussing neutral atoms and that with each added electron there is also an added plus charge on the nucleus. These two charges produce the dominant variation in the energy levels and upon this variation the exchange effect is superimposed.

The reader may verify that so far as the distribution of electrons in the $2p$ states is concerned, the exchange effect will lead to the configurations shown in Fig. 4 for the states of lowest energy for the atoms. Let us consider carbon for example; if the electrons have opposite spins—that is, if there is one weight on each $2p$ tray—there will be no lowering due to the exchange effect; if the electrons have the same spin, however, then each loses energy because of exchange and the energy of the atom is less than for the case of parallel spins. The fact that one electron is not enough and that two or more electrons are required to produce the exchange effect is a natural consequence of the origin of the exchange energy.

The exchange energy is due to the electrostatic repulsion between the electrons and results directly from the application of Pauli's principle to Schroedinger's equation. The exchange effect emerges in a quite straightforward fashion from a consideration of wave functions, but usually no attempt is made to explain it in non-mathematical

terms. It seems to the writer, however, that the explanation given below does contain the mathematical essence in physical language.[3a] Pauli's principle, we have said, is the quantum mechanical analogue for electrons of the classical law that two bodies may not occupy the same place at the same time; it is, however, more general in the sense that it does not apply alone to location but rather to a combination of location and velocity and spin, and it requires that any two electrons differ essentially in one or more of these. Now a difference in the values for the spin quantum numbers of two electrons is a sufficiently great difference to permit them to have the same velocity and the same location (i.e., be very near together compared to atomic dimensions). If the spin quantum numbers are the same, however, there must be a difference in location or in velocity. Now two electrons having the same values of n and l, as for example two $2p$ electrons, move in similar orbits and have much the same velocities; hence, if their spins are the same they must differ in location—that is, they will satisfy Pauli's principle by keeping away from each other. If, however, their spins are different, then they need not keep away from each other, and in their motion about the nucleus they are, on the average, closer together than for the case of the same spin. Since the energy of repulsion between the two electrons decreases as they move farther apart, the average energy of the electrons is less for the case of parallel spins, for which Pauli's principle requires most difference in location; and this is just the effect shown in Fig. 4. Furthermore, if the electrons differ in their values of n and l, then their velocities are quite different and the restriction upon location is not so important and their electrostatic energy of repulsion for parallel spins is nearly the same as for opposite spins. There is, however, a small exchange effect between electrons of different n and l values as may be appreciated in Fig. 4 for boron, for example, by noting that one $2s$ level i depressed compared to the other owing to the presence of the $2p$ electron.

We see that helium and neon correspond to electron configurations which fill all the levels below $n = 2$ and $n = 3$ respectively. One sometimes refers to the states with $n = 1$ as the K shell, and to those with $n = 2$, 3, 4, etc. as L, M, N, etc., shells. The rare gases helium and neon then correspond to electron configurations consisting of "closed shells"—that is, to shells all of whose states are occupied.

[3a] As the aspects of exchange energy needed for the exposition are those discussed above in connection with Fig. 4, this explanation is not essential to the later argument of this article and is given in the hope that it may invest the concept of exchange energy with the appearance of a little more physical reality. If it fails in this, the reader is requested to disregard it.

The Periodic Table

The elements from lithium to neon constitute the first short period of the periodic table, Fig. 5. The second short period, running from sodium to argon, is built up in a similar way by filling the $3s$ and $3p$

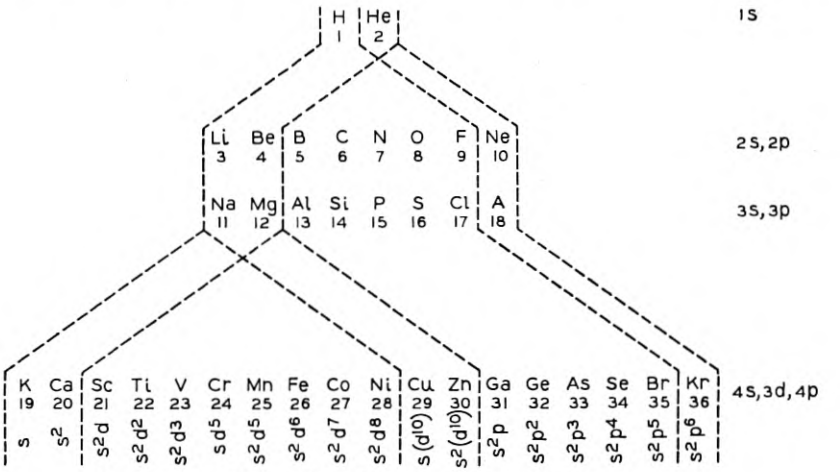

Fig. 5—First part of the periodic table.

levels. Some of the chemical properties of the elements of these periods are quite easily understood in terms of the electron configurations of the atoms. An atom of lithium or of sodium has one easily removed electron and thus can become a positively charged ion; only one electron, however, can be so easily removed, and for this reason sodium and lithium do not have doubly charged positive ions in chemistry and are, therefore, monovalent positive elements. Similarly beryllium and magnesium have two easily removable electrons, and are divalent; however, their electrons are harder to remove than those of the alkali metals; hence the alkaline earth metals, beryllium and magnesium, are not so electropositive as the alkalis.[4] The halogens, fluorine and chlorine, present a contrasting picture. Instead of having one loosely bound electron, they have one low-lying empty state. They can therefore hold tightly an extra electron, and thus be negative ions. The rare or noble gas elements helium, neon and argon consist only of closed shells. They can neither gain nor lose electrons in chemical compounds and are, therefore, generally aloof to chemical urges.

[4] For brevity we shall refer to the alkali metals and alkaline earth metals simply as alkalis and alkaline earths.

The Transition Elements

Actually argon does not correspond to a complete system of closed shells, since for it none of the $3d$ states in the M shell is occupied. For the elements below copper, $Z = 29$, these $3d$ levels lie above the $4s$ and below the $4p$. They are filled up progressively in the series of elements scandium, titanium, vanadium, chromium, manganese, iron, cobalt, and nickel, which are known as the transition elements of the first long period of the periodic table. The first two elements after argon are potassium (an alkali) and calcium (an alkaline earth); these are similar to sodium and magnesium in having respectively one and two s electrons. The first transition element, scandium, however, is not like beryllium or aluminum, for with it the filling of the $3d$ states begins. The electron configurations for several of the transition elements are shown in Fig. 6. An interesting case occurs at chromium;

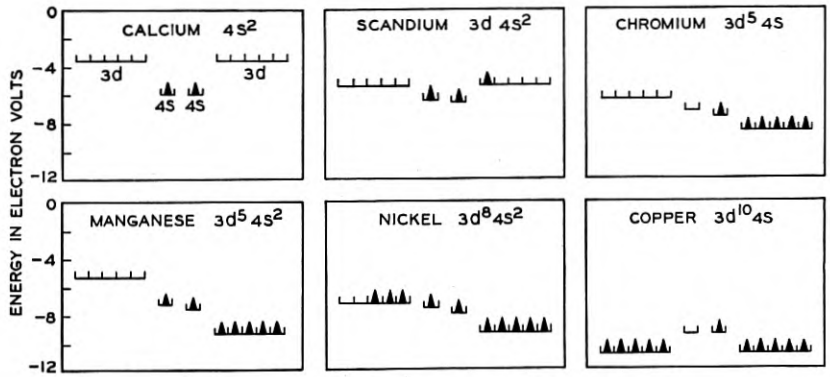

Fig. 6—Electron configurations for transition elements.

for it the exchange effect is so great that the $3d$ levels drop below the $4s$ and one $4s$ electron is transferred. Since there is an exchange effect between all electrons of the same spin, the remaining occupied $4s$ state in chromium has the same spin as the occupied $3d$ states. A similar transfer of a $4s$ electron occurs at copper, which is then left with one $4s$ electron and tends therefore to be monovalent (in the divalent copper ion the $4s$ electron and one $3d$ electron are removed). These transition elements are of particular interest because three of them, iron, nickel, and cobalt, are ferromagnetic in the solid. The atoms themselves are magnetic, as may readily be seen for chromium, for example; in it all the electrons have their spins parallel and hence their magnetic moments add to give a free chromium atom a magnetic

moment six times as large as the spin magnetic moment of the electron.[5] The same exchange effect which causes the $3d$ quantum states to fill unevenly in the isolated atom causes, in the case of the metal, an uneven filling of the "energy bands" which arise from these $3d$ states. We shall return to this topic in the section on ferromagnetism.

Solving Schroedinger's Equation

The possible quantum states of an atom are obtained by solving Schroedinger's equation for an electron moving in the potential field of the nucleus and the other electrons. In Fig. 7*a* we have represented the potential energy of an electron in an atom. If this potential energy (call it U) is known as a function of the position x, y, z, of the electron, then the Schroedinger equation is

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}+\frac{8\pi^2 m}{h^2}(E-U)\psi=0, \tag{1}$$

where m is the mass of the electron, h is Planck's constant and E is an unknown energy and ψ an unknown wave function, for which a physical interpretation will shortly be given. It is found that this equation possesses proper solutions only for certain values of E; once these values are known, the equation can be solved for the unknown wave functions. The fact that only certain values of E are possible will probably seem more natural after reading the discussion given below of a mechanical system. The permitted energies and wave functions give the system of quantum states of Fig. 2.

The wave equation of Schroedinger is similar in form to many of the other wave equations of mathematical physics. In Fig. 7*g* to 7*j* we represent a stretched membrane like a rectangular drumhead. If the mass per unit area of the membrane is σ and the surface tension is T, then the wave equation for it is

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial z^2}+\frac{4\pi^2 f^2\sigma}{T}\varphi=0, \tag{2}$$

where f is the unknown frequency of vibration and φ is the unknown vertical displacement. Applied to the membrane, this equation has solutions only for certain values of f; the standing wave patterns corresponding to the four lowest frequencies are shown in Figs. 7*g* to 7*j*.

[5] For transition elements other than chromium, the motions of the electrons in their wave functions produce magnetic moments that must be considered as well as the spin; for a discussion of this point the reader is again referred to "Spinning Atoms and Spinning Electrons" by K. K. Darrow, *Bell System Technical Journal*, XVI, p. 319 and to texts on atomic physics.

The type of vibration of the system in one of these patterns is called "a normal mode." The patterns are described by two "quantum numbers" p and q which are equal to one plus the number of nodal lines (indicated by arrows) running across the membrane from front to back and from right to left respectively. In Figs. 7*b* and 7*c* we show the wave patterns corresponding to the 1*s* and 2*s* states of the atom; the quantum numbers of the ψ waves are also correlated to nodes. In the case of the membrane the frequencies and standing

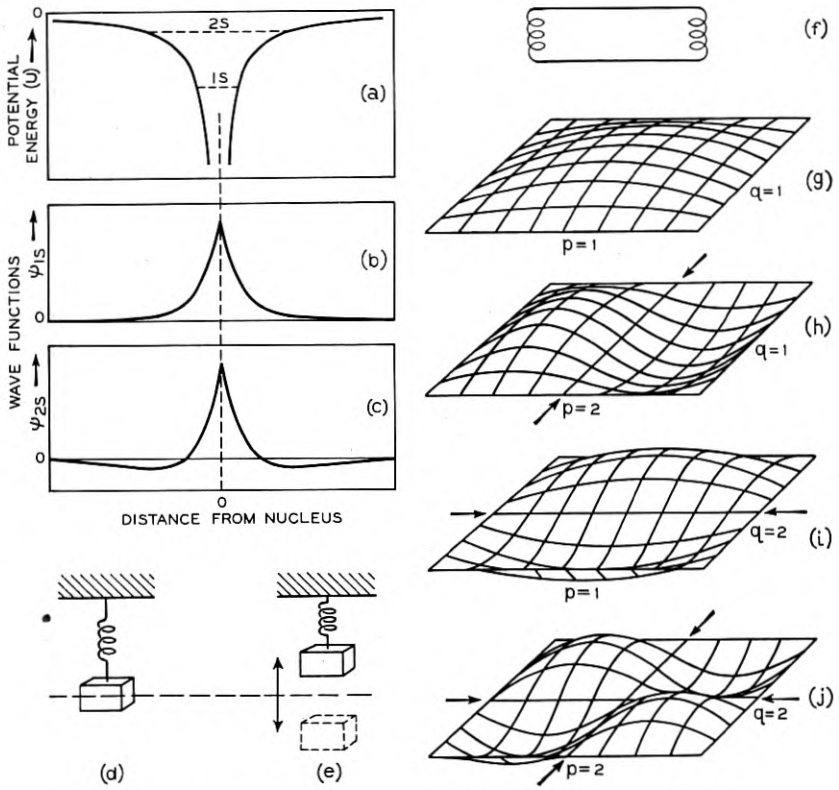

Fig. 7—The atom and some mechanical and electrical analogues.

(a) Potential energy of an electron in the atom.
(b) The 1*s* wave function.
(c) The 2*s* wave function.
(d) A mechanical analogue and
(e) its normal mode of vibration.
(f) An electrical analogue.
(g) to (j) The first four normal modes of vibration of a stretched drum head. (From "Vibration and Sound" by P. M. Morse, McGraw-Hill, New York, 1937. Courtesy of the McGraw-Hill Book Co.)

wave patterns are determined not only by the values of mass per unit area, σ, and tension, T, of the membrane but also by the "boundary condition" that it be clamped on its rectangular edge; at this edge the vertical displacement φ must vanish. The corresponding boundary condition for the atom is that the wave function ψ vanish at all points infinitely far from the nucleus.

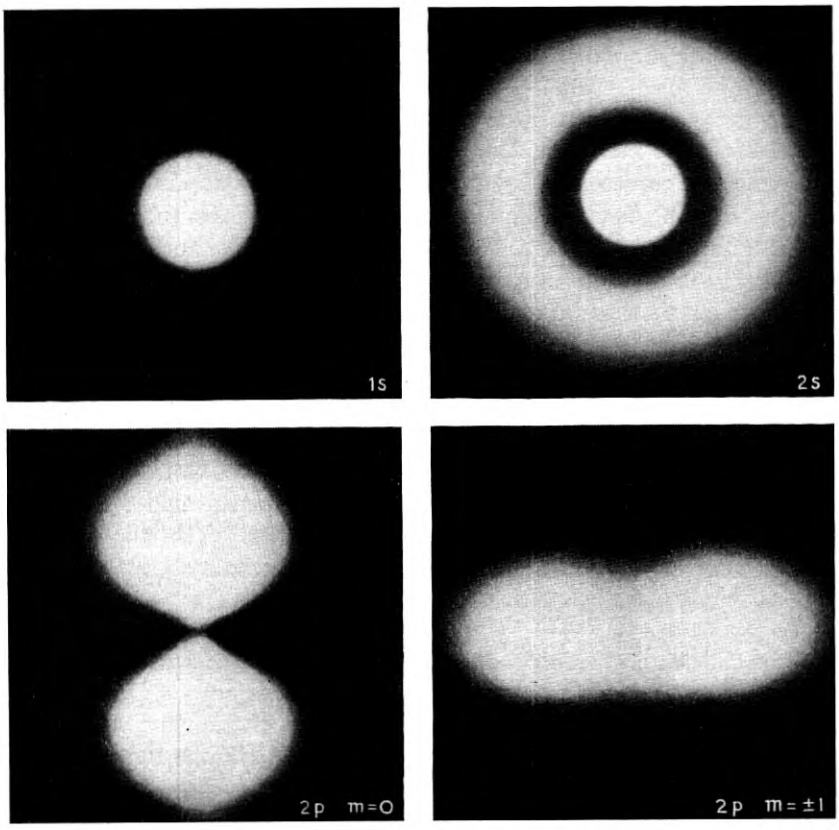

Fig. 8—The electron charge densities for four wave functions. Cross-sections are given for the $1s$ and $2s$ wave functions and perspective views for the $2p$. $1s$ represents a ball of charge; $2s$, a ball surrounded by a shell; $2p$ $m = 0$, a dumbbell-like distribution; $2p$ $m = \pm 1$, a doughnut-like distribution seen edgewise.

The quantity $|\psi|^2$ has a direct physical interpretation: its value at any point in space gives the probability of finding the electron at that point. If it were possible to take a photograph of the electron's motion with a time exposure so long that a true average of its positions would be obtained, this photograph would represent $|\psi|^2$. In Fig. 8

we show the predicted patterns as obtained by H. E. White,[6] who photographed a model representing the wave functions. We see that for the 2*s* wave function the electron is much farther from the nucleus on the average than for the 1*s*; this accounts for higher energy of the 2*s* state. For a hydrogen atom the 2*s* and 2*p* actually have the same energy. For other elements the 2*s* lies lower as shown in Fig. 2; this is because an electron in the 2*s* state penetrates the K shell and feels the full charge of the nucleus whereas an electron in the 2*p* state stays outside of the K shell and is thus shielded from the nucleus by the two electrons of the K shell.

For purposes of illustration we have considered the rectangular drumhead as a mechanical analogue for the wave equation. Other analogues are represented by sound waves in rooms and in organ pipes and by standing electromagnetic waves in wave guides, tuned cavities, and rhumbatron oscillators. We shall use two simple analogues in our later discussion. One is the mechanical vibrator represented in Fig. 7*d* which we consider to be restricted to vertical motion. It is a system with a single frequency—like an imaginary atom with only one possible state—and its one normal mode of vibration is a simple harmonic motion up and down equally far above and below its equilibrium position as indicated in Fig. 7*e*. The other is an electrical analogue, Fig. 7*f*, consisting of a section of transmission line terminated at each end by a high inductance. This system has a series of normal modes of vibration and a related series of allowed frequencies. The allowed frequencies correspond to the energy levels of the atom.

Electrons in Molecules

We shall next consider what happens when two atoms are brought so close together that their quantum states "interact." Two similar atoms widely separated have each a distinct set of quantum states and wave functions and the scheme of energy levels for the two atoms is obtained by duplicating the energy level scheme of Fig. 2. However, if the atoms move so near together that the wave functions for the corresponding quantum states of the two atoms overlap, there is an alteration in the energy levels. Figure 9 is intended to illustrate this process. Figure 9*a* shows the potential energy of an electron for points on a line passing through the centers of the two nuclei, and Figs. 9*b* and 9*c* show for points on the same line the values of the correct wave functions in this field. These wave functions are obtained, approximately, by using the 1*s* wave function for the two separate

[6] *Physical Review*, *37*, 1416 (1931). I am indebted to Professor White for the photographs used for these illustrations.

atoms; *b* represents the sum of the wave functions and *c* the difference. The process involved in getting these molecular wave functions is mathematically similar to that of finding the normal modes for a system of two similar coupled oscillators. In Fig. 9*d* we represent two

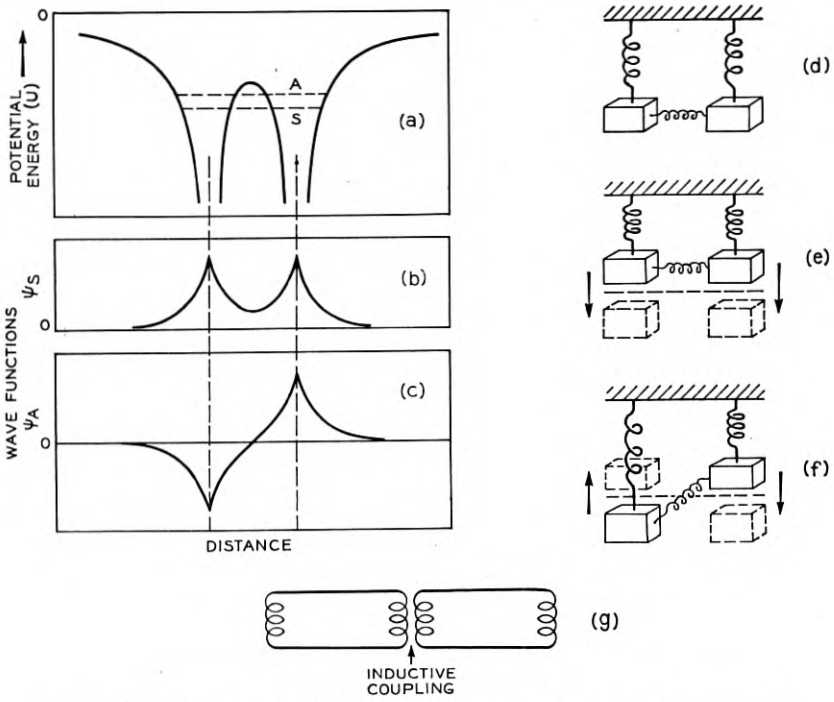

Fig. 9—A diatomic molecule and some mechanical and electrical analogues.

(a) The potential energy of an electron for points on a line through the two nuclei.
(b) and (c) Values of two wave functions for points on the same line.
(d) Two coupled oscillators.
(e) and (f) Their normal modes of vibration.
(g) Two coupled circuits.

weakly coupled oscillators. The normal modes of vibration for the coupled system are as indicated in Figs. 9*e* and 9*f*. These two modes have different frequencies. Similarly if two electrical circuits are placed so that there is some inductive coupling between them, we find that each frequency is split into a pair. This inductive coupling is similar to the overlapping of the wave functions; thus the coupling between the circuits is large when the electromagnetic field of one reaches over to the other. We may summarize the situation by saying that before coupling each frequency occurred twice, once for each sys-

tem; after coupling two frequencies are present and the corresponding modes of vibration belong not to the individual systems but instead to the pair of systems. For the case of atoms each quantum state occurs twice, once for each atom, before the atoms interact; after interaction there are still two quantum states but now they have different energies and are shared by both atoms.

As the atoms are brought closer together the energies separate more and more. The behavior is indicated qualitatively in Fig. 10. The

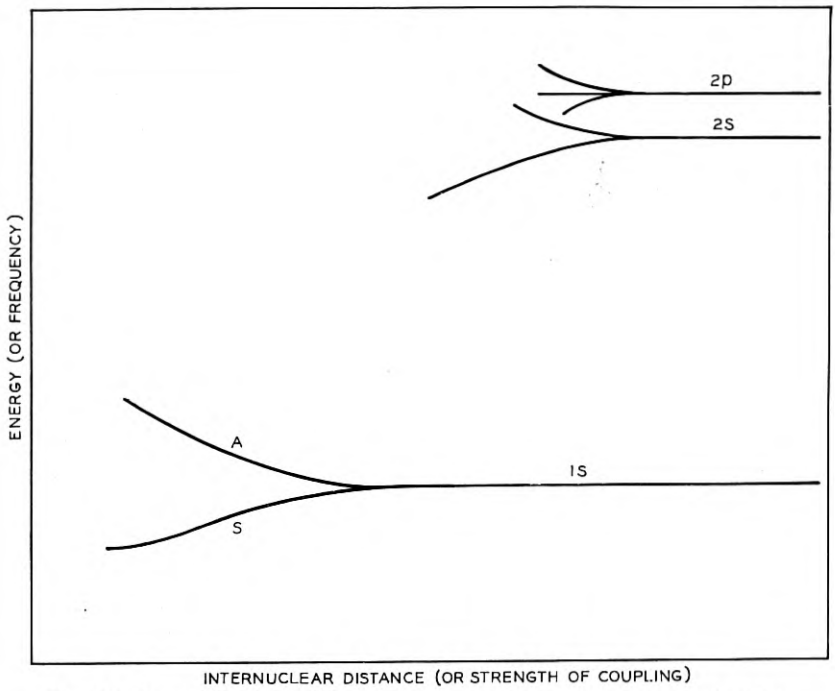

Fig. 10—Energy levels of a diatomic molecule versus internuclear distance.

L levels ($2s$ and $2p$) split at larger distances than the K levels because their wave functions extend farther from the nucleus (see Figs. 7 and 8) and overlap at greater distances. The details of the splitting are somewhat complicated and only the start is shown here. For the mechanical analogues shown in Fig. 8, the coupling raises one frequency and leaves the other unaltered. On the other hand, the quantum mechanical interaction results, at large distances, in equal displacement up and down for the energy levels.

We can use Fig. 10 to describe the formation of a molecule of hydrogen, H_2. We start initially with the single electron of each atom

in the 1*s* state. When the atoms have come together the 1*s* states have split into two energies with two states for each energy—one with each spin. In the H_2 molecule, the two electrons will both go into the lower 1*s* states, which both have the wave function of Fig. 8*b*. Bringing the atoms closer together decreases the energy of the electrons and results in the binding together of the atoms. This tendency of the electrons to reduce their energies by drawing the atoms together is opposed by the electrostatic repulsion between the nuclei. The repulsion between the nuclei is inoperative when the atoms are sensibly separated because then each nucleus is shielded from the nucleus of the other atom by the electron of that atom. When the atoms are closer together, however, the electrons no longer perform this shielding perfectly and the nuclear repulsions are important. Hence with decreasing interatomic spacing the electronic energy decreases and the energy of repulsion of the nuclei increases, and the equilibrium internuclear distance is the one which makes the total energy of the molecule a minimum.

The situation is quite different for two helium atoms. There being two electrons in each, for them all four of the "1*s* molecular orbitals," as the states of Fig. 9 are called, are occupied. When all the molecular orbitals are occupied, there is no decrease in energy when the two atoms are brought together: in this case the decrease of energy for the electrons in the two lowest states is compensated by the increase for the electrons in the upper states—more than compensated, as a matter of fact, because the upper states rise slightly more rapidly than the lower ones fall. This effect results in a repulsion between two helium atoms. This repulsion is a consequence of the closed shell nature of the helium atom and always occurs between such closed shells even if the atoms are different, as, for example, a neon and an argon atom. We shall refer to this closed shell repulsion, which occurs when the wave functions of the two closed shells encroach upon each other, as an "encroachment energy." The encroachment energy, as we have said, always corresponds to a repulsive force between the closed shells. We shall find that it plays a very important role both in ionic crystals and in metals.

The encroachment energy occurs not only between rare gas atoms but also between ions of elements which as neutral atoms have partly filled shells but in the ionic form have closed shells. Consider, for example, an alkali halide molecule such as LiF. For this case the 2*s* valence electron of lithium is transferred to the vacant 2*p* level of fluorine (see Fig. 4), thus leaving two ions with closed shell configurations, the Li^+ being He-like, the F^- being Ne-like. These two oppo-

sitely charged ions attract each other and draw together until the encroachment repulsion between their closed shells balances the attraction and holds them apart. Conversely if one of two atoms having closed shells normally is converted to an ion, the closed shell arrangement will be destroyed and an attraction will result. For example, He_2^+ ions, which may be thought of as formed from an atom and an ion, have been observed in the mass spectrograph.[7] The attraction is explained by noting that in this case there are three electrons and the effect of two of them in the lower 1*s* molecular orbital overbalances the one in the upper orbital and gives rise to a net attraction.

Electrons in Crystals

We must now investigate the quantum states and their energy levels for electrons in crystals. As in the case of the diatomic molecule we shall study the dependence of the energies upon the distance between atoms, which in the case of a crystal is called the lattice constant. We shall treat the lattice constant as a variable and shall refer to the values for it found experimentally as "observed" or "experimental lattice constants" and indicate them on the figures by the symbol a_0. We shall consider the allowed states to be occupied in accordance with Pauli's principle and on this basis find how the energy of the crystal as a whole depends upon the lattice constant. In this section we shall deal with crystals at the absolute zero of temperature and leave the complicating features of thermal effects to a later section. According to theory, the equilibrium state of a system at absolute zero is that one which makes the energy least. Hence, a knowledge of the dependence of energy upon lattice constant can be used to predict the equilibrium lattice constant—that is, the one which should be found experimentally—for according to the theory quoted above, the equilibrium lattice constant is the one which makes the energy of the crystal least.

In Fig. 11 we show the potential energy for an electron in a one-dimensional crystal, the distance being measured along a line passing through the atomic nuclei of the constituent atoms. In the interests of simplicity we imagine that high potential walls through which the electron cannot pass bound the crystal at both extremities. These boundary conditions lead to a simpler set of wave functions than would boundary conditions like those discussed for the free atom. The simplification of problems by arbitrarily choosing certain boundary conditions is a standard device in some branches of quantum mechanics; it introduces an error, but if the crystal is large, the error is negligible;

[7] F. L. Arnot and Marjorie B. M'Ewen, *Proc. Roy. Soc.*, *171*, 106, 1939.

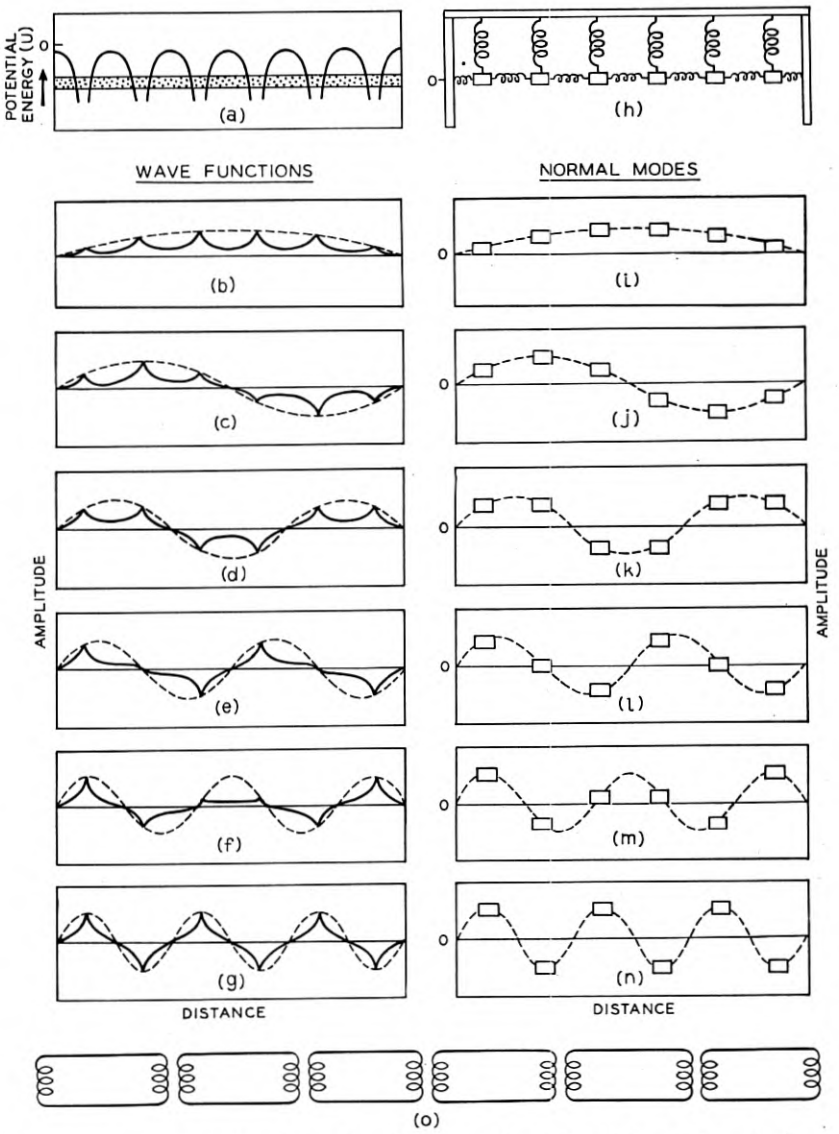

Fig. 11—A one-dimensional crystal and some mechanical and electrical analogues.

(a) The potential energy of an electron for points on a line through the nuclei.
(b) to (g) Wave functions for points on the same line.
(h) Coupled oscillators.
(i) to (n) Their normal modes of vibration.
(o) Coupled circuits.

the situation is similar to that which arises through neglecting "edge effects" in calculating the capacity of a parallel plate condenser.

In Fig. 11 we show also a series of coupled oscillators with boundary conditions corresponding to those prescribed for the atoms. For this case there are six coupled oscillators, which when uncoupled had six independent normal modes of vibration all with the same frequency, like that shown for the single oscillator of Fig. 7*d*. After coupling there are six normal modes all having different frequencies; the standing wave patterns corresponding to these are shown in Figs. 11*i* to 11*n*. A similar splitting of frequencies occurs when the members of a set of electrical circuits are placed in close proximity as indicated in Fig. 11*o*. For them the situation is more complicated than for the mechanical oscillators; each mechanical oscillator has but a single frequency, whereas each circuit has a fundamental and a sequence of overtones. Each possible frequency for the electrical circuits is split by coupling into a set of six.

In Figs. 11*b* to 11*g* are shown the proper electron wave functons which arise from the 1*s* atomic states. These wave functions have different energies. When the atoms were separated there were six 1*s* wave functions for the six atoms and each of these gave two states—one for each spin. After coupling we find six crystal wave functions and twelve crystal quantum states, the same number of states for each spin as before. This illustrates a fundamental theorem concerning wave functions in crystals which holds for two and three dimensions as well as for one and is true no matter how large the number of atoms in the crystal. This theorem, which we shall refer to as the "conservation of states," may be stated as follows: consider a set of N similar isolated quantum mechanical systems; they may be single atoms or molecules. Any particular quantum state is then repeated N times over, once for each system. Now bring the systems together so that the energy levels have split up. Then for each N-times-repeated quantum state of the isolated systems, we find a set of N crystal quantum states. In other words, putting the systems together may change the energies and wave functions of the quantum states but no states are gained or lost in the process.

In Fig. 12 we indicate how the energy levels of the states depend upon the lattice constant. Each energy level in the figure corresponds to two states, one for each spin. For simplicity only two atomic levels are shown here. Higher energy levels split appreciably at larger lattice constants because of the greater spatial extension of their wave functions. For any particular lattice constant the energy levels arising from a given atomic state lie in a certain band of energy. The

number of states in the band is, of course, proportional to the number of atoms; however, if the number of atoms is large, the *width of the band is independent of the number of atoms.* This concept of allowed bands of energies for the crystal states plays the same role in crystals as the concept of energy levels in the atom. We shall refer to 3*s* bands

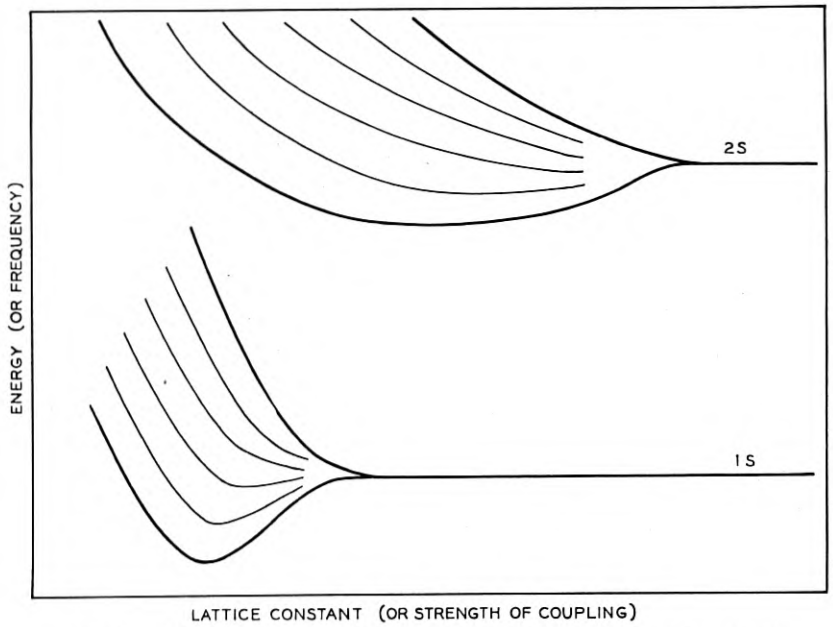

Fig. 12—Dependence of energy levels upon lattice constant or frequency of vibration upon strength of coupling.

and 3*d* bands of energy levels in crystals in much the same way as we refer to the 3*s* and 3*d* atomic energy levels from which these bands arise.

We must emphasize that like the molecular states, the crystal states do not belong to the atoms individually but instead belong to entire system of atoms.

Before proceeding with the application of these ideas to crystals with large numbers of atoms, we shall digress by anticipating several subjects to be taken up in the next paper. For the energy levels of isolated atoms the quantum numbers n, l, m, and m_s were satisfactory. For a crystal, however, there will be many crystal quantum states in an energy band all arising from atomic levels having the same values of n, l, m, and m_s. A new quantum number is therefore needed to distinguish the various crystal states in an energy band one from another. In Fig. 11 we see that the wave function of each crystal state is asso-

ciated with a wave form, shown dashed. This wave form is in every case of such a wave-length that it has an integral number of half wave-lengths along the edge of the crystal.[7a] The number of half wave-lengths is a suitable quantum number for the wave functions in the crystal and a more general consideration of it in the next paper will lead us into the theory of the "Brillouin zone" and the zone structure of energy bands. The second subject concerns the transmission properties of the crystal. The set of coupled circuits of Fig. 11 constitutes a length of transmission line. A line of this type is a simple filter network and as such it has bands of frequency in which power will be transmitted and bands in which it will not. The allowed frequencies lie in the transmitting bands. The system of coupled mechanical vibrators likewise constitutes a mechanical filter. Just as the mechanical and electrical systems can transmit power in their allowed bands, a crystal can transmit an electron whose energy is in an allowed band. The electrons in an allowed band, however, can produce a net current only if the band is partially filled. Electrons in wholly filled energy bands, although individually representing tiny currents to and fro in the crystal, can produce—we shall find—no net current as their individual currents cancel out in pairs. On this basis the theorist explains the difference between metals and insulators as follows: in a metal some of the energy bands are partly filled, but in an insulator each energy band is either completely filled or completely empty.

Distributions of Quantum States in Energy Bands

When there are a very large number of atoms in the crystal, it is impractical to represent the energy levels by distinct lines as was done for the case of six atoms in Fig. 12 and another scheme must be used. For a crystal of macroscopic dimensions the number of levels in the band is of the order of 10^{24}, that is a million million million million. When so many levels are placed so close together, a continuous band of allowed energies is suggested. Actually, of course, only a discrete set of allowed energies is possible, the total number in the band being that required by the conservation of states. We shall now consider the distribution in energy of these quantum states; that is, how many lie in a given range of energy between E and $E + dE$. Let us call

[7a] For a three-dimensional crystal having the external shape of a cube, the three-dimensional wave function has an integral number of half wave-lengths along lines parallel to each edge of the crystal. This condition is illustrated in a simplified form by the wave patterns for the two-dimensional drum head shown in Fig. 7; for each normal mode, there is an integral number of half wave-lengths parallel to each boundary of the membrane, and, in fact, the values of these numbers are given by p and q.

this number dN; it will depend upon E and be proportional to dE and we may write

$$dN = N(E)dE, \tag{3}$$

where the function $N(E)$ represents the "number of quantum states per unit energy" at E. This equation, like many in statistical mechanics requires special interpretation, because if dE is small enough—less than the spacing between levels in the band—it may include no levels. If, however, we always use small but not infinitesimal values for dE, so many levels will be included in it that equation (3) is quite satisfactory. In Fig. 13*a* we represent qualitatively the distribution in

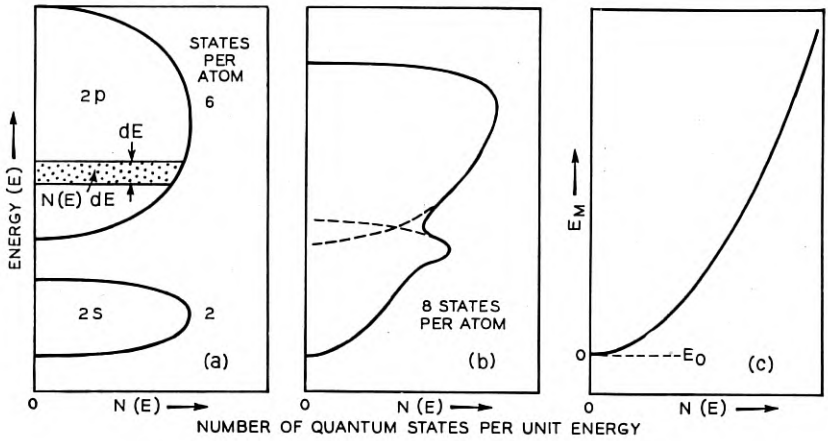

Fig. 13—Distribution of energy states in energy.

(a) For two separate energy bands.
(b) For overlapping energy bands.
(c) For free electrons.

energy for two energy bands. We plot $N(E)$ horizontally so as to retain the vertical scale for E. The area under the curve for the $2p$ levels is three times that for the $2s$. This is because the number of states in the $2p$ and $2s$ bands are respectively six times and two times the number of atoms in the crystal. The $1s$ band lies too low to be shown on this figure; its levels will be concentrated over a very narrow range in energy in keeping with the small splitting suggested in Fig. 12.

It is possible for the energy band arising from one atomic energy level to overlap the energy bands arising from other atomic levels. We shall be concerned below with several cases where this occurs for various crystals. When it does occur the states in the bands become mixed up and it is no longer possible to decide which atomic level was

the parent of each state in the band. This confusion is of no consequence, however, for it does not interfere with using the distribution in energy curves when they are obtained. Furthermore, the conservation of states holds when the bands overlap so that the total number of states per atom in the combined bands is the sum of the number of states per atom in the separated bands. In Fig. 13*b* we represent a distribution qualitatively similar to that occurring in various metals where *s* and *p* bands overlap. The number of states per atom in the combined bands is eight, four for each spin.

A very important distribution-in-energy curve is that of the case of "free electrons." This is the distribution one obtains by imagining that the electrons in a crystal are perfectly free—that is, subjected to no electrostatic forces whatever—but that they are required to remain within a certain prescribed volume. The distribution of quantum states in energy for this case is represented in Fig. 13*c*. At low temperatures the electrons tend to occupy the lowest states consistent with Pauli's principle and the system is referred to as a "degenerate electron gas." With the aid of the distribution curve, the energy and pressure of this gas can be calculated. We shall require its energy for a discussion of the binding energy of sodium, but we shall give here only the equation of the curve, leaving the calculation of the energy until later.[8] According to the theory, then, for the case of free electrons the number of states per unit energy is given by

$$N(E) = \frac{4\pi V}{h^3}(2m)^{3/2}E^{1/2}, \tag{4}$$

where V is the volume of container, h is Planck's constant, m the mass and E the energy of the electron; for free electrons E is all kinetic energy, there being no potential energy. For the case of the alkali metals, calculations show that the wave functions for the valence electrons are very similar to the wave functions for free electrons. For these metals we can use Eq. (4) to calculate energies.

Before utilizing the concepts of energy bands in a discussion of the binding energies of crystals, we must define two symbols to be used in describing the energy of a state in the band. For this purpose we arbitrarily separate the energy E of a crystal state into two parts: one of these is denoted by E_0 and stands for the energy of the lowest state in the band and the other is E_M which stands for the energy which the state possesses in excess of E_0—that is, its energy above the bottom

[8] The reader will find a derivation of this curve given in K. K. Darrow's article "Statistical Theories of Matter, Radiation and Electricity," *Bell System Technical Journal*, Vol. VIII, 672, 1929 or *Physical Review Supplement*, Vol. I, 90 (1929), and in various texts on quantum statistics and the theory of metals.

of the band. We shall find in the next paper that the quantum states in a band represent electrons traversing the crystal with various average speeds. The state E_0 has an average speed of zero. The subscript "M" has been assigned with these ideas in mind and stands for "motion," implying that an electron with energy greater than E_0 has an energy of motion E_M. In general both E_0 and E_M are actually composite energies containing both kinetic and potential energy; only in the case of free electrons is E_M purely an energy of motion. We shall not use in this paper the property of motion connected with the values of E_M; however, we shall use the division of the energy into two parts, E_0 and E_M, and we shall for convenience refer to the latter as an "energy of motion."

We shall next apply the concept of the energy band to a determination of the binding energies of several types of crystals. It is one of the principal merits of the theory of energy bands in crystals that we can treat many different crystal types on the basis of the same set of ideas. As we shall point out later, however, the band theory is most appropriate for metals: for ionic and valence crystals other theories are better suited.

Energy Bands and Binding Energies of Metals

For several metals the wave functions and distribution of states in energy have been found by solving Schroedinger's equation for the electrons in the metal. We shall discuss sodium since it constitutes one of the simplest cases and is the first metal for which good calculations were carried out.

A sodium atom, Na, contains ten electrons in filled K and L shells and one valence electron in the M shell; its electron configuration is $1s^2\ 2s^2\ 2p^6\ 3s$. When the atoms are assembled together as in the metal, the $3s$ atomic state gives a wide band which overlaps the $3p$ band while the lower levels widen only very slightly.

The formation of the energy bands[9] is shown in Fig. 14. Since the K and L bands are very narrow, it is possible to neglect the changes in the wave functions of the electrons occupying them and to concentrate upon the valence electrons. The valence electrons then move in a potential field produced by the Na^+ ions and the other electrons.

It can be shown by a lengthy argument that for the case of a monovalent metal, the energy of the metal as a whole is very nearly equal to the sum of the energies of the valence electrons.[10] It is rather

[9] J. C. Slater, *Phys. Rev.*, *45*, 794 (1934).

[10] An exact statement of the situation is too involved for this paper. The reader can find a more complete discussion in Mott and Jones "The Properties of Metals and Alloys," Chapter IV.

natural that the valence electrons should contribute so largely to the binding since the complete shells of electrons making up the Na^+ ion, that is the K and L electrons, are only slightly affected by bringing the atoms together to form a metal. The result that the energy of the metal is the sum of the energies of the valence electrons is

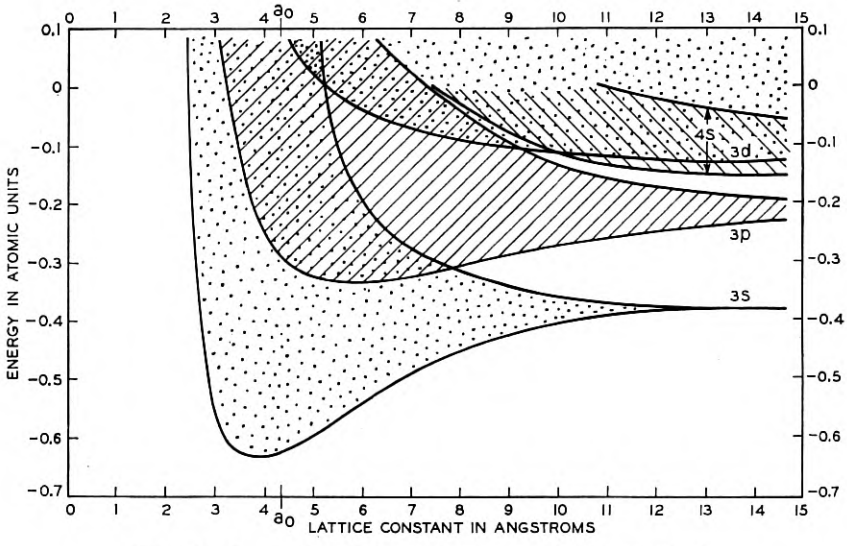

Fig. 14—Energy bands for sodium versus lattice constant.

of great importance in applying the theory. We shall discuss below how the energies of the various states in the band depend upon the arrangement of the atoms; some of these states are occupied and the energy of the crystal can be found by adding the energies of the occupied states. In this way we can find how the energy of the crystal depends upon the arrangement of the atoms and can find what arrangement makes the energy least. According to theory the arrangement of least energy is the stable one and the one which should be found in nature. The remainder of this section will be devoted to discussing the energies of the quantum states in metals and the energies of the electrons which occupy them.

The first satisfactory solutions of Schroedinger's equation for electrons moving in the field of a metal were obtained for sodium by Wigner and Seitz.[11] They assumed, in keeping with the findings of experiment, that the sodium atoms were arranged on a body-centered

[11] E. Wigner and F. Seitz, *Phys. Rev.*, *43*, 804 (1933) and *46*, 509 (1934) and E. Wigner, *Phys. Rev.*, *46*, 1002 (1934).

cubic lattice. They did not, however, assume that the lattice constant was that given by experiment but instead carried out calculations for each of several assumed values for the lattice constant lying on both sides of the experimental value. The results of their calculations are shown in Fig. 15. The curve marked E_0 is the energy of the lowest

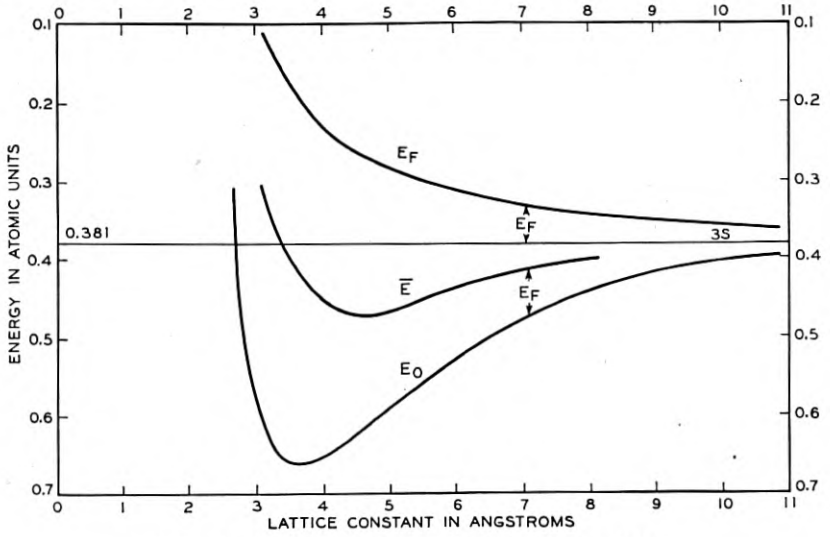

Fig. 15—Energy for sodium versus lattice constant.

level in the valence band. Only two electrons, one with each spin, can occupy this energy level and all others must occupy states of higher energy—that is, only two electrons can have zero value for the "energy of motion" E_M and all others must have larger values. By a method of calculation described below, it can be shown that the average energy of motion of a valence electron is given by the curve marked E_F in the figure. Hence the total energy per valence electron in the metal, which for a monovalent metal is equal to the energy per atom, is represented by the curve marked $\overline{E}$ in the figure; $\overline{E} = E_0 + E_F$. Figure 15 exhibits the dependence of this energy upon the lattice constant. The abscissa of the minimum in the $\overline{E}$ curve gives the theoretically predicted value for the equilibrium lattice constant. The binding energy or heat of sublimation is defined as the energy required to separate the metal into isolated atoms; it is the difference in energy between the minimum of the curve and the value of $\overline{E}$ for infinite lattice constant—that is, for free atoms. Finally, the curvature of the curve at its minimum is a measure of the energy required to compress or expand the crystal and from it a value for the compressibility can

be obtained. In Table I, we compare theoretical and experimental values of lattice constant, binding energy, and compressibility calculated by the method described above. The theoretical values were computed by Bardeen [12] who has added some refinements and corrections to the original calculations.

TABLE I

	Li		Na	
	Calc.	obs.	Calc.	obs.
Lattice Constant (angstroms) . . .	3.49	3.46	4.53	4.25
Heat of Sublimation (Kg. cal./gm. atom)	34	39	23	26
Compressibility (cm^2/dyne)	8.4×10^{-12}	7.4×10^{-12}	12.0×10^{-12}	12.3×10^{-12}

Although the theory can give quite satisfactory values for the various physical quantities shown in Table I, it cannot as yet predict precisely what crystalline form a metal like sodium will take. In carrying out the calculations discussed above, it was assumed that the atoms were arranged in a body-centered cubic lattice. Now the correct theoretical procedure would be to calculate the energy for all conceivable arrangements of the atoms and then to select that arrangement giving the least energy of all as the theoretically predicted equilibrium arrangement. This program is, of course, too laborious to be practical—furthermore experience shows that metals, with but few exceptions, crystallize in one of three forms: body-centered cubic, face-centered cubic, and hexagonal close-packed. For this reason it might be regarded as sufficient to calculate the energies for the face-centered cubic and hexagonal close-packed and to compare these with that for the body-centered cubic. When such calculations are carried out, however, it is found that the minimum energies calculated for the three forms differ among themselves by amounts which are negligible in view of approximations necessary in making the calculations. Hence the theory cannot predict with any certainty which form really has the lowest energy; it does predict, however, that all three forms do have nearly the same energy and gives a value for this energy. Actually the binding energy of sodium must be greatest for the body-centered cubic lattice because this form is the one that occurs in nature and so must be the form of lowest energy. However, it is probable that the difference in energy between the various possible allotropic

[12] J. Bardeen, *Jour. Chem. Phys.*, *6*, 367, 372 (1938).

forms for sodium is really very small—so small that we should not expect the present theory to evaluate it. Some indication that the energy of caesium is very nearly the same in the body-centered form and face-centered form (or possibly the hexagonal close-packed form) is furnished by a transformation at high pressures observed by Bridgman. In the next paper we shall meet a case where the theory does seem able to differentiate between the energy of face-centered and body-centered structures. In general, however, the procedure is to use the crystal structure found by x-rays and to calculate the energy for a series of values of the lattice constant as was done for sodium.

We must now return to a discussion of the curves E_0 and E_F of Fig. 15. About the curve E_0 we shall only say that it is obtained by solving Schroedinger's equation for and finding the energy and wave function of an electron in the lowest state in the energy band. The wave function for this state, however, possesses the interesting feature of being very nearly the same as the wave function for a free electron having zero energy of motion. From this fact it is possible to draw the conclusion that the distribution in energy of motion of the valence electrons in sodium is the same as the distribution in energy of free electrons in an electron gas. Accepting this conclusion, we can then use the formulas given for the distribution of states for free electrons in order to calculate the mean energy of motion of the valence electrons in sodium. The results of this calculation, which we give in a footnote,[13] lead to the energy curve E_F. This energy curve is, from its

[13] We shall first derive a general expression for E_F without specifying the particular form of $N(E)$. Since in this footnote all energies are measured from E_0, we shall omit the subscript M from E_M and use simply the symbol E in the equations. The total number, denoted by n, of atoms in the crystal is equal to the total number of valence electrons. Let the volume of the crystal be V. Because of the duplicity due to the spin there are $2n$ states in the band, and half of them will accommodate the n electrons so that the band will be filled only up to a certain energy $E_{\max.}$ We must therefore have

$$n = \int^{E_{\max.}} N(E)dE. \tag{i}$$

Once the distribution function $N(E)$ is known, this equation serves to determine $E_{\max.}$ The average energy of motion of an electron in these occupied states is, from the definition of an average, the total energy of motion divided by the total number of electrons:

$$E_F = \frac{1}{n}\int_0^{E_{\max.}} EN(E)dE. \tag{ii}$$

Substituting the value of $N(E)$ for free electrons into the first equation gives

$$n = \frac{8}{3} V(2mE_{\max.}/h^2)^{3/2} \tag{iii}$$

The quantity V/n is the volume per electron which in the case of a monovalent metal

definition, the average energy per electron of a degenerate electron gas. In a degenerate electron gas the electrons have the least possible energy consistent with Pauli's principle and with the distribution of quantum states in energy. For reasons associated with the origin of the statistical mechanics of electrons—that is, with the Fermi-Dirac statistics—the energy E_F is called the "Fermi energy" and given the subscript F. The energy E_F is far greater than the average energy per particle of an ordinary classical gas. We shall see below how this fact accounts for the very small specific heat of the electron gas. From the dependence of the energy upon volume, the pressure of the electron gas can be calculated. It is usually very large, for sodium it is about 50,000 atmospheres. The force that prevents this pressure from blowing the metal apart is represented by the E_0 curve, which gives decreasing energy with decreasing lattice constant and corresponds to a force pulling the atoms together. A more detailed discussion of these forces will be taken up in the third paper of this series.

Other Metals

Calculations similar to those for sodium can be carried out for other metals. The band structure as calculated for copper by Krutter [14] is shown in Fig. 16. Ten electrons per atom can be accommodated in the $3d$ band and two per atom in the $4s$. For copper the $3d$ band is filled—in keeping with the fact that the Cu^+ ion consists of filled K, L, and M shells. From the discussion of molecules given above

---

is the same as the volume per atom. Denoting by Ω the value of V/n, we find

$$E_{\max.} = \left(\frac{3}{\pi}\right)^{2/3} \frac{h^2}{8m} \Omega^{-2/3} = 36.1\Omega_0^{-2/3}\text{ev} \qquad \text{(iv)}$$

where Ω_0 is the volume per atom in cubic Angstroms. For a body-centered cubic lattice with lattice constant a Angstroms, $\Omega_0 = a^3/2$. Substituting the expression for $N(E)$ into the equation for E_F gives

$$E_F = \frac{3}{5} E_{\max.} = 21.6\Omega_0^{-2/3}\text{ev}. \qquad \text{(v)}$$

Expressing E_F in atomic units and Ω_0 in terms of the lattice constant, we find

$$E_F = 2.54a^{-2}. \qquad \text{(vi)}$$

This is the equation of the curve for Figure 15. The values of $E_{\max.}$ calculated from the above equations for a series of metals are

Metal	Li	Na	K	Rb	Cs	Cu	Ag	Au
$E_{\max.}$(ev)	4.74	3.16	2.06	1.79	1.53	7.10	5.52	5.56

[14] H. M. Krutter, *Phys. Rev.*, *48*, 664 (1935).

we should expect encroachment repulsions between these ions when their wave functions begin to overlap. In the band picture this repulsion results from the spreading of the 3*d* band; since the band spreads more to higher energies than to lower energies and since it is full, the average energy of an electron in it increases as the lattice constant decreases. Thus the same result, repulsion between closed shells, is

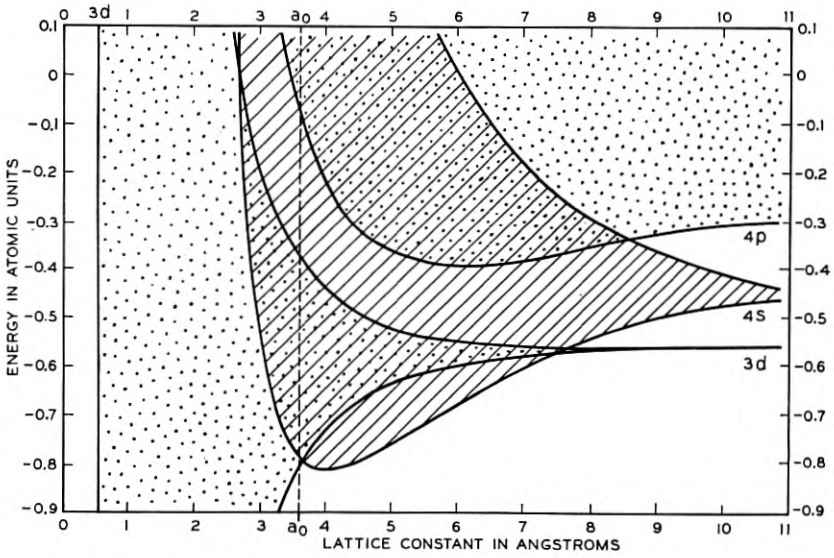

Fig. 16—Energy bands for copper versus lattice constant.

found for the ions in a metal as for the rare gas atoms. For elements whose atoms have partially filled 3*d* levels the situation is quite different. For them only part of the levels of the 3*d* band will be filled and there will be a decrease in the energy of the 3*d* electrons in the metal as compared to the atom. This has been proposed by Seitz and Johnson as an explanation of the fact that the highest binding energies for the metals of a transition series occur for those that have approximately half-filled 3*d* bands and for which consequently nearly all of the 3*d* electrons have lower energies than in the atomic state.[15] The very high melting point metals—columbium, molybdenum, tantalum, and tungsten—come approximately at the middle of their transition series. In Table II we give the binding energies for a number of the transition elements.

[15] F. Seitz and R. P. Johnson, *Jour. App. Phys.*, *8*, 84, 186, 246 (1937).

TABLE II

HEATS OF SUBLIMATION FOR SEVERAL METALS INCLUDING THE TRANSITION ELEMENTS IN KILOCALORIES PER GRAM ATOM *

K	Ca	Se	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
19.8	48	70	100	85	88	74	94	85	85	81	27.4
Rb	Sr	Y	Z	Cb	Mo	Ma	Ru	Rh	Pd	Ag	Cd
18.9	47	90	110	—	160	—	120	115	110	68	26.8
Cs	Ba	La	Hf	T	W	Re	Os	Ir	Pt	Au	Hg
18.8	49	90	—	185	210	—	125	120	127	92	14.6

* Taken from F. R. Bichowsky and F. D. Rossini "The Thermochemistry of the Chemical Substances," Reinhold (1936), except for T which was taken from D. B. Langmuir and L. Malter, *Phys. Rev. 55*, 1138 (1939).

Energy Bands of Diamond

In Fig. 17 we show the band structure for diamond as calculated by Kimball.[16] The configuration of the carbon atom is $1s^2 2s^2 2p^2$ and

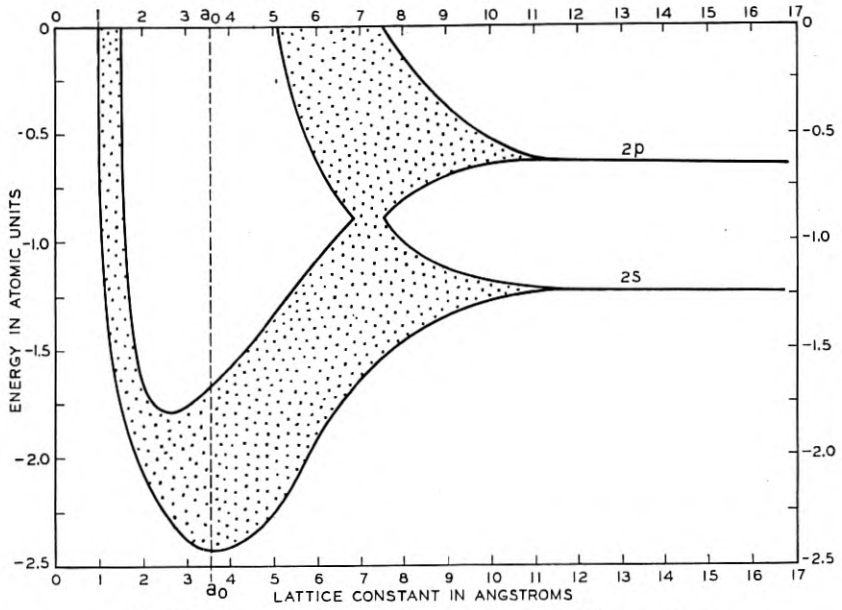

Fig. 17—Energy bands for diamond versus lattice constant.

all four of the L shell electrons are involved in the binding. At large lattice constants the lower band contains two states per atom and the upper six. The lower band is completely filled, the upper only one-

[16] G. E. Kimball, *Jour. Chem. Phys.*, *3*, 560 (1935). Some unimportant features resulting from approximations in Kimball's work have been modified in this figure.

third filled. To the left of the crossing of the bands, Kimball finds that both bands contain four states per atom so that the lower is filled and the upper is empty. The actual spacing in diamond occurs to the left of the crossover and, as we shall see in the next paper, the resultant filled band and empty band arrangement explains the absence of electrical conductivity for diamond. The diagram suggests an explanation for the conductivity in graphite; one of the lattice constants of graphite is known to be larger than the abscissa of the crossover of Fig. 17; hence in graphite there are partially filled bands and conduction.

The general downward trend of the bands in Fig. 17 indicates a strong binding energy for diamond; but quantitative calculations of the total energy have not been made.

The type of binding involved in diamond is quite like the binding of metals save that, owing to the absence of partially filled bands, there is no electrical conductivity. In both cases the energy arises from the lowering of energy levels as the atoms come together. In chemical terminology the binding of diamond is referred to as "homopolar" signifying that the atoms are all similarly charged, or rather uncharged. In crystals containing ions rather than neutral atoms, the cohesion is due largely to electrostatic forces and one refers to binding as "heteropolar" or "ionic."

Energy Bands and Binding Energies of Ionic Crystals

The energy band theory can be applied to the calculation of the binding energy of ionic crystals. Before discussing this application, however, it will be instructive to examine a somewhat simpler approach to the problem.

A sodium chloride molecule consists of a sodium ion and a chlorine ion. These ions have charges of $+e$ and $-e$ respectively and have a mutual electrostatic energy of

$$-\frac{e^2}{r}, \tag{5}$$

where r is their distance of separation. This electrostatic energy, which we shall refer to as the "coulomb energy," decreases with decreasing interatomic distance. If the ions are close together, as they are in a molecule, the energy of encroachment due to the overlapping of their closed shells must be considered; this energy increases with decreasing interatomic distance. The equilibrium distance is the one that makes the total energy, coulomb plus encroachment, a minimum.

Closely similar calculations can be carried out for a crystal. One finds the total coulomb energy of all the ions and the total encroachment energy; and then one finds the lattice constant that makes the total energy a minimum. The total encroachment energy is easily found; only atoms which are nearest neighbors in the lattice have appreciable overlapping with each other and it is therefore a straightforward and simple calculation to find the total number of encroachments in the crystal. The coulomb energy is not quite so simply found, however, because the electrostatic interaction of a given ion with its nearest neighbors is no more important than its interaction with its vastly larger number of more distant neighbors. The electrostatic problem is solved as follows: one considers a NaCl lattice which is perfect except for the absence at one lattice point of a Na^+ ion; one finds by known techniques of electrostatics the value at the vacant lattice point of the electrostatic potential due to the remaining ions; this potential is negative and has a value

$$-\phi = -\frac{Me}{4a} = -\frac{3.49e}{a}, \tag{6}$$

where a is the lattice constant and M is a numerical constant known as Madelung's constant, which has a particular value for any special lattice; for the NaCl lattice, $M = 13.94$. If now a Na^+ ion is placed in the vacant lattice point, its electrostatic energy will be $-e\phi$. Similarly the electrostatic potential at a vacant Cl^- lattice point is $+\phi$ and the electrostatic energy of a Cl^- placed there is $-e\phi$. The total electrostatic energy per NaCl molecule in the lattice, however, is not $-2e\phi$ but only $-e\phi$; the factor 2 does not occur since otherwise the electrostatic interaction between each pair of ions would be included twice.[18] The total energy per molecule for the crystal can be found by combining the coulomb and the encroachment energies, and the equilibrium lattice constant and binding energy per molecule thence can be derived.

Using wave functions for Na^+ and Cl^- ions obtained by D. R. Hartree, who has found solutions of Schroedinger's equation numerically, the encroachment energies in NaCl have been evaluated by R. Landshoff.[19] For the lattice constant and binding energy for NaCl he obtains 5.88Å and 165 Kg.-cal./gm. atom while experiment gives 5.63Å and 183 Kg.-cal./gm. atom.

Some very important theoretical work of a semi-empirical nature has

[18] To see that this is true in a simple case, use the procedure described above to calculate the electrostatic energy of an isolated NaCl molecule.

[19] *Zeits. f. Phys.*, *102*, 201 (1936).

been carried out for the alkali halides. In it an analytical expression suggested by theory and containing adjustable constants has been used for the closed shell repulsions. The adjustable constants have been determined from certain data and then used for predictions which can be compared with other data. Using a relatively small number of adjustable constants, Born and Mayer,[20] Mayer and Helmholz,[21] and Huggins and Mayer [22] have calculated a much larger number of values for lattice constant and binding energy for many alkali halides with an agreement with experiment of the order of one per cent.

Let us now consider NaCl using the band picture. We shall reach the rather surprising conclusion that there is no fundamental difference between the results obtained from it and those just deduced from the ionic picture described above.

In Fig. 18 we show qualitatively the behavior of the bands for NaCl.[23] In the ionic state, an electron is transferred from the Na $3s$ to the Cl $3p$. The general shifting of the bands is explained as follows. The wave functions corresponding to the Cl^- $3p$ band, like all energy band wave functions, are distributed over the whole crystal. They are not, however, equally intense at Na^+ and at Cl^- ions; instead they are definitely concentrated about the Cl^- ions. The electrostatic potential at a Cl^- ion, due to the remainder of the crystal, has the same value (6) as was found in discussing the ionic method. Since the charge on the electron is $-e$, the energy of each of the states in the Cl^- $3p$ band varies with a in the same manner as does $-e\phi$. A similar argument shows that the Na^+ energy bands vary as $+e\phi$. At a certain lattice constant, the Cl^- $3p$ and $3s$ bands and the Na^+ $2p$ and $2s$ bands begin to widen. Since these bands are full, this widening gives the customary encroachment energy just as it was obtained in the ionic picture. The shifting of the bands similarly gives the coulomb energy. To see this we note that per NaCl molecule there are 18 electrons in the Cl^- bands where energies vary as $-18e\phi$ and that there is also one chlorine nucleus with charge $+17e$ whose energy varies as $+17e\phi$. This leaves a net effect of $-e\phi$ for the Cl^- ions. Similarly a net effect of $-e\phi$ comes from the electrons and nuclei of the Na^+ ions. As in the case of the ionic method the sum, $-2e\phi$, of these energies really contains each ionic energy twice and the total electrostatic energy per NaCl molecule is $-e\phi$. So far as calculating energies is concerned, the two methods give equivalent results; the advantage, if any, lies

[20] *Zeits. f. Phys.*, *75*, 1 (1932).
[21] *Zeits. f. Phys.*, *75*, 19 (1932).
[22] *Jour. Chem. Phys.*, *1*, 643 (1933).
[23] J. C. Slater and W. Shockley, *Phys. Rev.*, *50*, 705 (1936).

with the ionic method rather than the band method because of the more immediate physical interpretation of the former.

We may remark that in the discussion of metallic sodium, it was not necessary to consider the potential energy of the nuclei and the closed

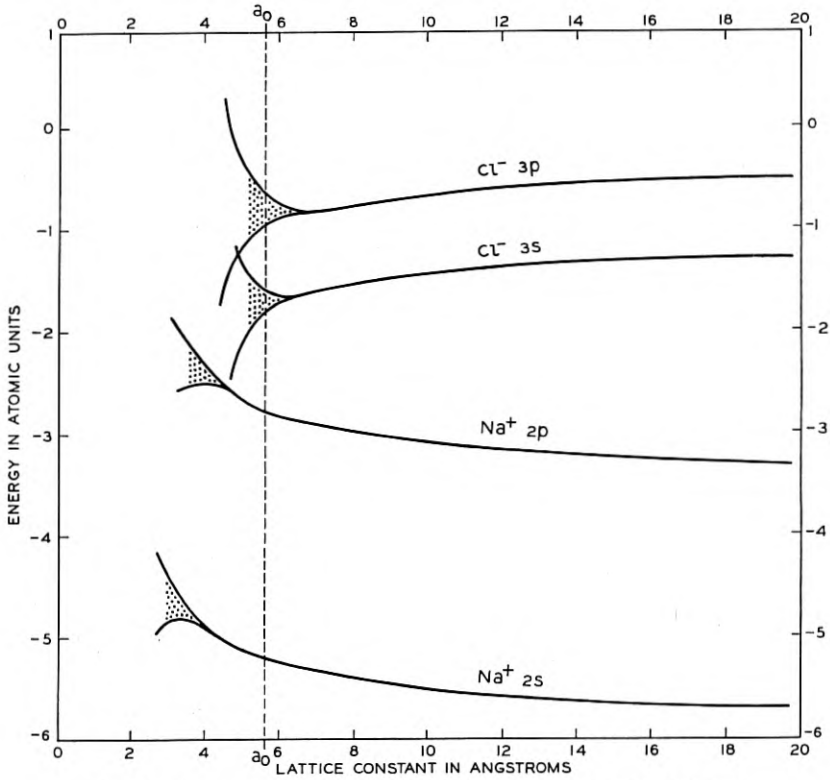

Fig. 18—Energy bands for sodium chloride versus lattice constant.

shell electrons as was done in NaCl. This is because sodium metal is not ionic—although we think of it as consisting in part of Na^+ ions, the electrostatic forces between them are suppressed by the shielding effect of the electron gas. In an ionic crystal, like NaCl, there is no electron gas and the coulomb energy must be considered in the manner described above.

Energy Bands for Other Crystals

There are chemical compounds which lie between the homopolar and ionic types. For example in the sequence of compounds NaF, MgO, AlN, SiC the compounds are progressively less and less definitely ionic—the least ionic, SiC or carborundum, being homopolar. Simi-

larly there are compounds, in particular intermetallic compounds, which are more like metals than like either ionic crystals or valence crystals. Thus there is an intermediate field which connects all three of the simple types of binding. Good computations are lacking for these intermediate cases; we shall return to a discussion of some aspects of them in connection with semiconductors in the next paper.

Concerning a Classification of Crystals

In the last section we saw how the concept of the energy band can explain the binding energies of a number of different types of crystals. Although the band theory has the merit of being very general it has the disadvantage of being at the same time rather abstract. Other theories have been developed to explain the cohesion of particular types of crystals; and, while lacking the generality of the band theory, they have the advantage of a more immediate physical interpretation in their own particular fields. In this section we shall digress from the exposition of the band theory in order to describe briefly some of the simpler viewpoints of the other theories.

We have discussed in the last section three types of binding. Sodium exemplified the metallic type; diamond, the homopolar or valence type; and sodium chloride, the ionic type. The distinction between the valence bond and the metallic bond is not very clearly indicated in the band theory; the only difference there had to do with the degree of filling of the bands. There is another difference, however, which has been long familiar to chemists. The homopolar compounds are usually characterized by "directed valence." Thus the "tetrahedral carbon atom" is a familiar concept of organic chemistry. In crystals in which homopolar binding is dominant the atoms are arranged so that each atom has the proper valence bonds with its neighbors. In diamond each carbon atom is tetrahedrally surrounded by four other carbon atoms. In silicon carbide, carborundum, a similar situation prevails: each carbon is tetrahedrally surrounded by four silicons and vice versa. These crystals are said to have a "coordination number" of four, or $z=4$, meaning that each atom has four nearest neighbors. In crystals of the divalent elements—sulphur, selenium and tellurium—each atom has two near neighbors and the valence condition is satisfied; these crystals have a coordination number of two. The monovalent halogens form crystals in which each atom has one near neighbor, coordination number one. In the metals, however, the neighbors of a given atom are as many as eight or twelve—do these large coordination numbers imply that the metals have eight or twelve electron pair bonds with their neighbors?

According to the quantum mechanical theory of valence, which in itself forms a theory with as many ramifications as the band theory, the electron configuration $1s^2 2s^2 2p^2$ of carbon is especially suited for forming "electron pair bonds" with other atoms. In forming these bonds the wave functions from one atom and another become distorted so as to overlap and form a high electron concentration along the line between the atoms; the energy levels being incompletely filled for the atoms, this overlapping does not produce a repulsion but instead a binding together like that produced by the overlapping wave function in Fig. 9*b* in the hydrogen molecule. The carbon atom is capable of forming four such bonds and forming them most effectively along four lines, making the tetrahedral angles with each other.

Recently Brill [24] and his collaborators using x-ray analysis have determined the electron concentration in diamond, in which the carbon atoms are arranged in a tetrahedral manner. The results of their investigations are shown in Fig. 19A.[25] It is easily seen that the electrons are concentrated in the bonding directions forming homopolar bonds between the atoms.

The energy band theory, we have said, does not give the clearest picture of the valence crystals; it is, however, especially suited to treatments of the metallic bond. According to the band theory the valence electrons constitute an electron gas—that is, instead of forming electron pair bonds with localized overlapping of the wave functions, they form instead a more or less uniform region of negative charge. In this negative charge the positive ions float. Since the ions repel each other they tend to arrange themselves so as to use their space to best advantage and this requires that they take up one of the "close-packed" arrangements. Let us see why this is true. The close-packed arrangements are those obtained by trying to pack rigid spheres as compactly together as possible. For these arrangements then, the volume per sphere is less than for other arrangements; that is, the close-packed arrangements are the ones which give a minimum volume per sphere for a prescribed value for the distance between sphere centers. Conversely, the close-packed arrangements must be the ones which give a maximum value for the distance between neighboring sphere centers for a given value of the volume per sphere. Since the energy of motion, E_F, of the electron gas and, although we have not shown why, the energy E_O, depend for a metal mainly upon the volume, in metals we are interested in cases where the volume per

[24] R. Brill, H. G. Grimm, C. Hermann and Cl. Peters, *Ann. d. Physik, 34*, 393 (1939).

[25] The writer is indebted to Professor Grimm for his permission to reproduce Fig. 19 from his article: *Naturwissenschaften 27*, 1 (1939).

atom is closely prescribed. For a given volume per atom, as we have seen above, the close-packed arrangements are the ones which give the largest separation between neighboring positive ions; and since the positive ions repel each other, the close-packed arrangements will give the lowest energies. This accounts for the fact that the custom-

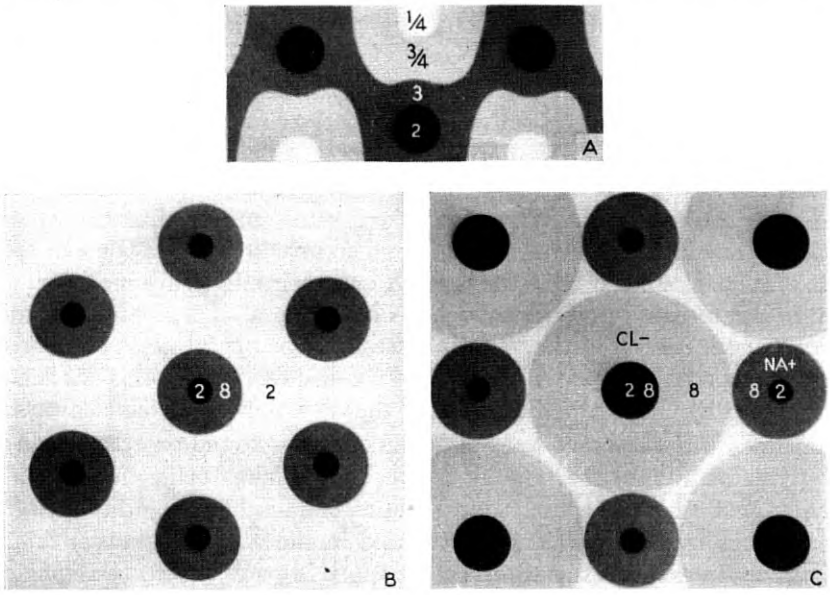

Fig. 19—Electron charge densities in crystals. The numerical values give the number of electrons per atom in the space corresponding to each intensity of shading.

(A) In diamond.
(B) In magnesium.
(C) In sodium chloride.

ary metallic lattices are the body-centered cubic lattice, the face-centered cubic, and the close-packed hexagonal. A further discussion of this physical picture of the nature of the metallic state will be given in the third paper of this series. In Fig. 19B we show qualitatively the electron density of metallic Mg according to Grimm.[25] It is seen that the valence electrons give a uniform negative charge in which the positive ions are embedded.

We have seen in the preceding section that the band theory of the alkali halides is essentially equivalent to the ionic theory. A large fund of evidence attests to the validity of the ionic theory, one item being the electron concentrations determined for sodium chloride by Brill [24] and his collaborators. These are represented in Fig. 19C; we

[25] Loc. cit.
[24] Loc. cit.

see that the electrons are closely held about the ions with very little overlapping of the closed shells.

In addition to the metallic, homopolar, and ionic bonds, there is still another interatomic force known as the Van der Waals force—a very weak force compared to the other three. We shall not discuss its origin here except to say that it arises from the spontaneous and mutual polarization of two atoms or molecules when in the neighborhood of each other. It is responsible for the "a" term in the Van der Waals equation for gases. When a crystal is formed from organic molecules, such as a crystal of benzene, the forces holding them together are the weak Van der Waals forces. This is the reason why "molecular crystals" have low melting points and binding energies. Although the Van der Waals forces are much smaller than the other three, they are not entirely negligible in comparison and in some of the calculations referred to in the last section, their effects are included.

It is interesting to note that in a single crystal of a given chemical compound, several of the various forces may be operative at once in a rather separable way. A classification of this sort for crystals has been discussed by Grimm.[25] For example the crystal mica, which cleaves so naturally into sheets, consists of planes of atoms bound together chiefly by valence forces, the binding between the planes being due to ions lying between and in the planes. Thus mica is held together in two directions by strong valence forces and in the other by weaker ionic forces. In asbestos the atoms are arranged in parallel rows, being held together in the rows by valence forces; the rows, on the other hand, are held to each other by ionic forces. The ionic bonds are more easily broken and asbestos crystals exhibit a typical fibrous structure. Mica and asbestos are intermediate members of a sequence of which diamond with all valence binding and sodium chloride with all ionic binding constitute the extremes. We shall give one more example: cellulose consists of long chains of carbon, oxygen and hydrogen, the chains held to each other by Van der Waals forces; it is an example of valence binding in one direction and Van der Waals binding in the other two.

This section has been a digression, as the main purpose of these papers is to illustrate the band theory of solids. It would hardly be fair to concentrate on this, however, without pointing out, as has been done in this section, that, although the band theory has great generality, it is best adapted for a certain class of solids and that other viewpoints are more natural for solids outside of this class.

[25] *Naturwissenschaften*, *27*, 1 (1939).

Thermal Properties of Crystals

In this section, as in the last, we shall digress from a straightforward exposition of the theory of energy bands and discuss the theories of specific heat and thermal expansion. These theories are well worth discussing on their own merits and furthermore their results and methods can be applied later to other topics. Thus the thermal vibrations that account for the specific heat will be shown in the second paper of this series to account for the resistance of metals. The discussion of thermal expansion given here will in the next section on magnetism be extended to an explanation of the unusual expansion properties of magnetic materials, in particular to an explanation of the very small expansion of invar. We shall, however, make use of the band theory once in this section by showing why the free electrons in a metal do not normally make an appreciable contribution to the specific heat.

In the introduction to this paper we pointed out that the specific heat per gram atom of a solid should be by classical theory $3R$—coming half from the kinetic energy and half from the potential energy of the atoms. This prediction is in reasonable agreement with experiment for many crystals at high temperatures. As the temperature is lowered, however, the observed specific heat decreases in such a way as to approach zero when the absolute zero of temperature is approached. This decrease in the specific heat at low temperatures, as well as the value $3R$ at high temperatures, is readily explained by quantum mechanics. In order to understand the explanation we must inquire into the atomic vibrations of a crystal.

In considering atomic vibrations we are really concerned with the motions of the nuclei. The electrons act as a cement to hold the nuclei in their equilibrium positions and exert restoring forces on them when they are displaced. (We shall see below why the electrons do not partake of the thermal energy.) The nuclei are effectively mass points in this theory and for quantum mechanical reasons, which we shall not discuss, they are incapable of acquiring thermal energy of rotation; hence so far as the crystal vibrations are concerned, we need consider only their translational or rectilinear motions. A crystal containing N atoms has $3N$ degrees of freedom since each nucleus can move in three dimensions. In order to find the specific heat of a crystal we must find the normal modes of vibration. The system of coupled oscillators in Fig. 11 represents reasonably well the normal modes of vibration for a one dimensional crystal whose atoms have only one degree of freedom. There is a similar set of normal modes for

a three dimensional array of atoms and, once the forces between the atoms are known, the frequency of vibration of each of the modes can be found. This means that so far as thermal vibrations are concerned, we can consider the crystal as equivalent to a set of $3N$ oscillators whose frequencies are those of the normal modes. We must next discuss the specific heat of a single oscillator.

According to classical statistical mechanics, a harmonic oscillator in a temperature bath at absolute temperature T will have an average thermal energy equal to kT, where k is Boltzmann's constant. The value kT is only an average value, we emphasize, and the oscillator will have other energies some of the time, the probability of each energy being given by known equations. The probability is very small, however, that the oscillator acquires more than two or three times kT of thermal energy. In a very large system of oscillators, the fluctuations of energy of the oscillators tend to cancel out and the probability of any appreciable fractional deviation of the total energy from its mean value is very small. If N is the number of molecules in a gram molecule ($N = 6.06 \times 10^{23}$), then $Nk = R$, the gas constant, $= 1.99$ cal. per gm. molecule per degree C. Hence the energy of $3N$ oscillators is $E = 3NkT = 3RT$ and the specific heat is $C = dE/dT = 3R$; this classical result that the specific heat of one gram atom of solid is $3R$ is known as the DuLong-Petit Law.

According to quantum mechanics, an oscillator of frequency ν has a set of quantum states whose energies are $\frac{1}{2}h\nu$, $(1 + \frac{1}{2})h\nu$, $(2 + \frac{1}{2})h\nu$, etc. The oscillator can take on only these energies. If it is in a heat bath of temperature T, however, it will sometimes have one allowed energy and sometimes another and as for the classical case we shall be concerned with its average energy. At absolute zero, the average energy is, of course, $\frac{1}{2}h\nu$. Now the probability of the oscillator gaining much more than kT of thermal energy is very slight. Hence the average energy of the oscillator remains at $\frac{1}{2}h\nu$ until thermal energy becomes large enough to excite it to the next state which is $h\nu$ higher, and consequently so long as kT is much less than $h\nu$ the quantum oscillator acquires much less thermal energy than would a classical oscillator. For kT much greater than $h\nu$, the oscillator will spend an appreciable fraction of its time in many of the quantum states and, as may be shown mathematically, the quantum restriction is no longer of importance so far as the average energy is concerned and the value kT is obtained just as in the classical case. In Fig. 20 the dependence upon temperature of the average energy and the specific heat for a quantum oscillator are shown.

The specific heat of the crystal is just the sum of the specific heats of its oscillators. Since the oscillators have different frequencies they have different specific heats and in order to add up the specific heats of all of them it is necessary to know how the various frequencies of the

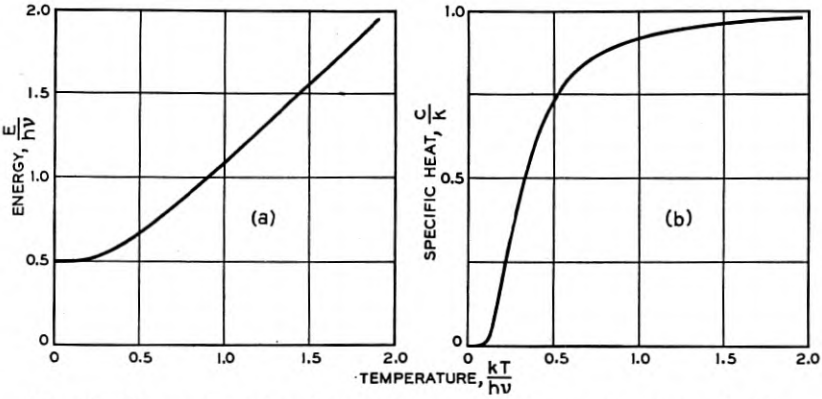

Fig. 20—Thermal behavior of an oscillator according to quantum mechanics.
(a) Energy versus temperature.
(b) Specific heat versus temperature.

oscillators are distributed. Once this distribution in frequencies is known it is merely a matter of summation to find total specific heat. The problem of finding the distribution in frequency of the oscillators was first solved by Debye. The low-frequency vibrations are very simply found for they are merely the acoustic vibrations of the crystal; they are very similar to the normal modes shown for the square membrane of Fig. 7*g*. For these low-frequency vibrations it can be shown by a straightforward argument, which is too long to give here, that the number dN of oscillators whose frequencies lie between ν and $\nu + d\nu$ is

$$dN = V4\pi \left(\frac{2}{C_T{}^3} + \frac{1}{C_L{}^3} \right) \nu^2 d\nu, \tag{7}$$

where C_T and C_L are the velocities of transverse and longitudinal waves in the solid and V is its volume.[26] Debye assumed that this distribution held for all the normal modes. There is of course a highest frequency of vibration, ν_{max}, and the total number of normal modes must be $3N$; hence Debye concluded that

$$\begin{aligned} 3N &= \int_0^{\nu_{max}} V4\pi \left\{ \frac{2}{C_T{}^3} + \frac{1}{C_L{}^3} \right\} \nu^2 d\nu \\ &= V4\pi \left\{ \frac{2}{C_T{}^3} + \frac{1}{C_L{}^3} \right\} \frac{\nu^3{}_{max}}{3}. \end{aligned} \tag{8}$$

[26] For a derivation see P. Debye, *Ann. d. Physik*, *39*, 789 (1912).

From this equation ν_{max} can be found if N/V and the velocities C_T and C_L are known. Knowing ν_{max} and the distribution in frequency, Debye summed the specific heats of all the oscillators and obtained the specific heat of the solid. According to this theory the specific heat vanishes at $T = 0$ and is proportional to T^3 near $T = 0$. At high temperatures it approaches the classical value of $3R$. A measure of the temperature at which the classical value is closely approached is given by the maximum frequency of atomic vibration ν_{max}; when kT is greater than $h\nu_{max}$, all the modes of vibration including the highest make substantial contributions to the specific heat. The temperature at which this occurs is known as the Debye temperature and denoted by the symbol θ_D; obviously $\theta_D = h\nu_{max}/k$. The specific heat given by Debye's equation is a function of T/θ_D only and can thus be represented by the expression $C(T/\theta_D)$; so that by this theory all crystals should have the same curve for specific heat versus temperature except for changes in the temperature scale corresponding to the different values of their Debye temperatures.

TABLE III

DEBYE TEMPERATURES IN DEGREES KELVIN USED IN FIGURE 21

Pb	Tl	Hg	J	Cd	Na	KBr
88	96	97	106	168	172	177
Ag	Ca	KCl	Zn	NaCl	Cu	Al
215	226	230	235	281	315	398
Fe	CaF_2	FeS_2	C			
453	474	645	1860			

In Fig. 21 is shown a compilation of specific heat data.[27] For each substance a value of θ_D (given in Table III) has been chosen so as to obtain the best agreement with experiment and the values of the specific heat have then been plotted as a function of T/θ_D. The Debye theory relates to specific heat at constant volume and in it no allowance is made for the energy due to thermal expansion. The experimental points are derived from measurements of specific heat at constant pressure which have been transformed by using a thermodynamical relationship so as to give specific heat at constant volume.

For these curves θ_D was chosen so as to obtain the best fit. It is, however, possible to calculate θ_D from theory by using the elastic constants of the material to evaluate C_T and C_L and then substituting in Eq. (8). For sodium the elastic constants have been calculated entirely from theory by the methods described in the section on

[27] Taken from E. Schroedinger, *Handbuch der Physik*, Vol. X, p. 307 (1926).

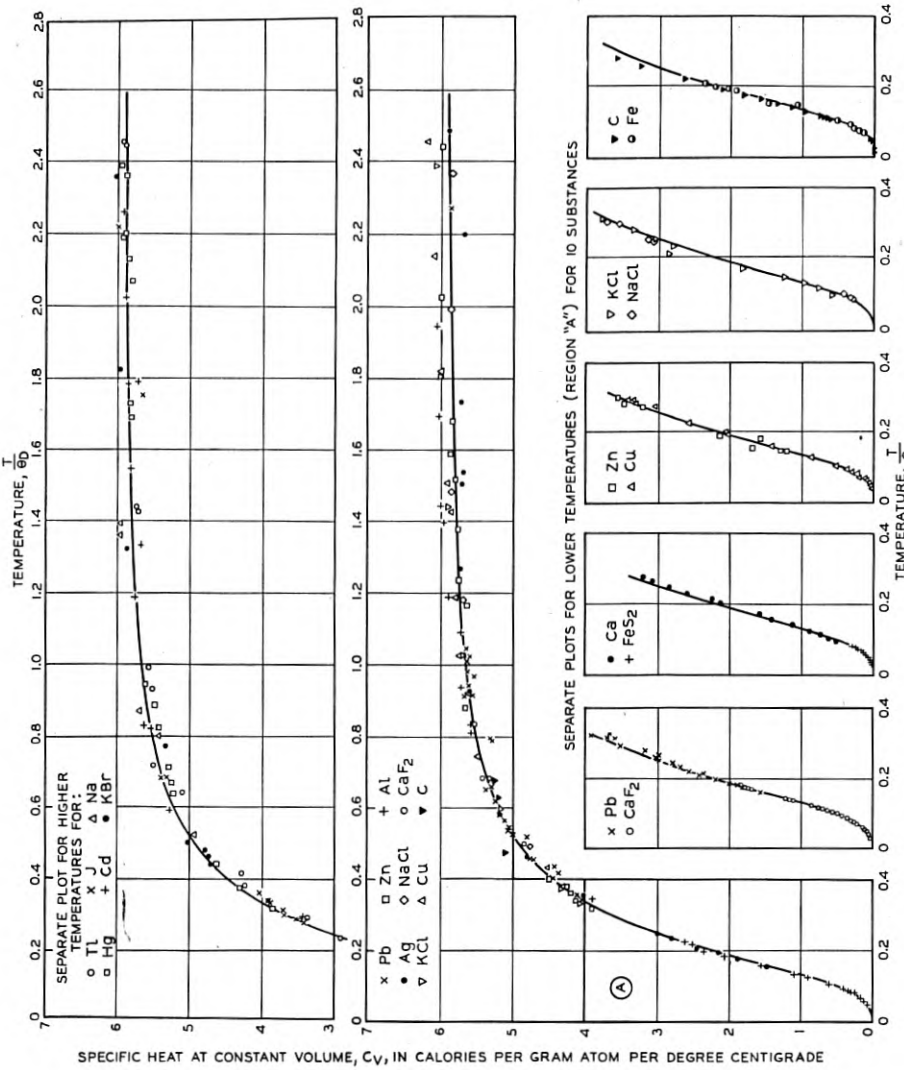

Fig. 21—Comparison of the Debye specific heat curve with experiment.

"Electrons in Crystals" and extensions of them* to be discussed in the third paper of the series. Using the theoretical values one obtains a value of 143° K for θ_D, whereas the value that fits experiment best is 172° K.

Recently calculations have been made from a model of the crystal as an assemblage of atoms rather than as a continuum as postulated in deriving Eq. (7)—that is, a model like the coupled oscillators, rather than like the stretched membrane, is used. These calculations, principally by Blackman, have explained some discrepancies between the Debye theory and experiment.

The Specific Heat of the Electrons

We must now see why the electrons contribute only slightly to the specific heat. Let us consider a case like that of sodium where we have a partially filled band. At the absolute zero of temperature, the electrons will fill all the levels below a certain energy E_1 and all the higher levels in the band will be empty (Fig. 22*a*). Now at temperature T some of the electrons will be excited to higher states; since, however, an electron cannot gain more than about kT of energy thermally, only those electrons whose energies lie in a range kT below E_1 can be excited. Electrons occupying states farther down in the band cannot acquire kT of thermal energy for, if they did so, they would have to move to states already occupied and such an act is forbidden by Pauli's principle. In order to demonstrate what a small fraction of the electrons can gain energy thermally, we point out that the width of the energy band is usually 4 or 5 ev while the value of kT in electron volts is $T/11{,}600$ and room temperature corresponds to a kT of about .03 ev. The electrons which do gain thermal energy have a normal value for the specific heat but constitute only about one per cent of all the valence electrons.

It might be maintained that the above argument is specious and that the electrons could all gain energy kT; this would not violate Pauli's principle because the electrons would move upward in the band as a unit, each moving into a state vacated by another electron. This contention is found to be wrong; one finds by using the statistical mechanics appropriate to electrons that the distribution of the electrons among the energy levels is given by the Fermi-Dirac distribution function.[28] According to this, the distribution of the electrons among the levels would be as indicated in Fig. 22*b*. The proba-

* K. Fuchs, *Proc. Roy. Soc. 157*, 444 (1936).

[28] For a discussion of the Fermi-Dirac statistics see K. K. Darrow, *The Bell System Technical Journal*, Vol. VIII, p. 672, 1929, or *The Physical Review Supplement*, Vol. I, p. 90, 1929.

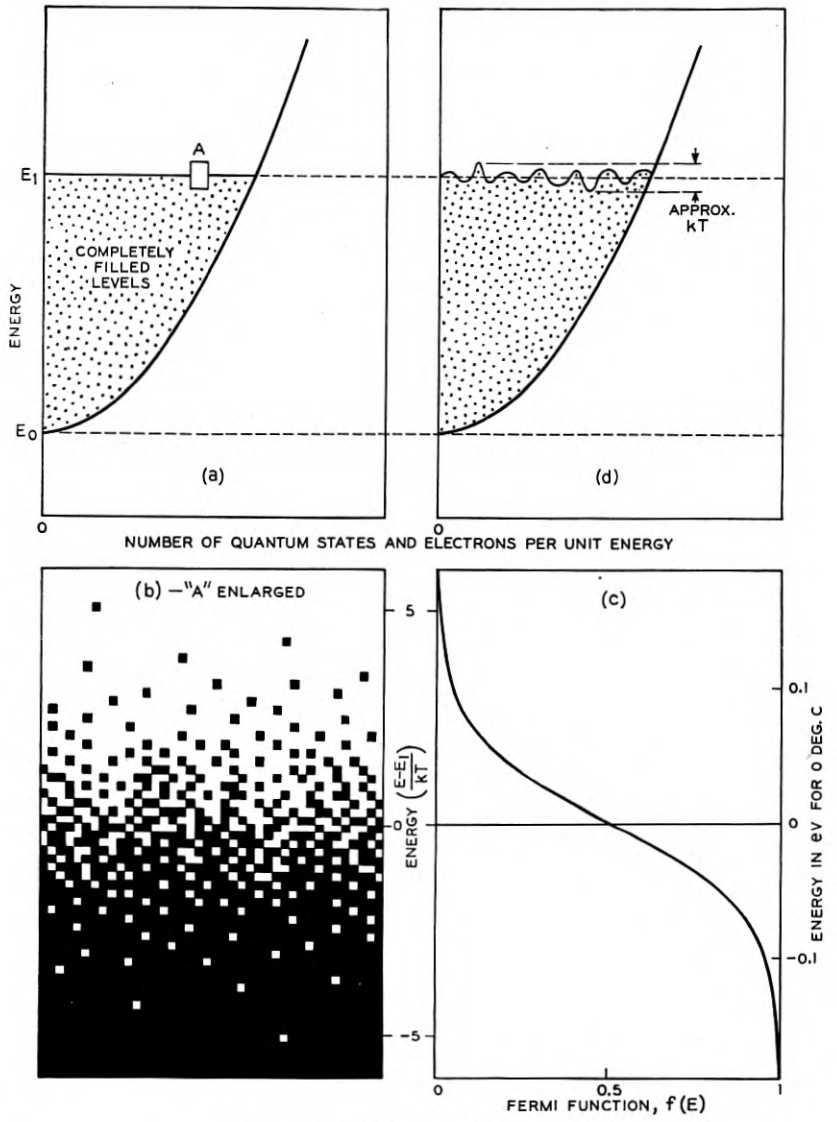

Fig. 22—Specific heat of the electrons.

(a) Distribution of electrons in energy for the absolute zero of temperature.
(b) Enlargement of part of (a) but for room temperature; each unit of area represents a quantum state.
(c) The Fermi distribution function.
(d) A water tank analogue.

bility that any particular energy state be occupied is given by the Fermi-Dirac factor f

$$f = \frac{1}{e^{\frac{E-E_1}{kT}} - 1}. \tag{9}$$

This factor is shown in Fig. 22*c* and the corresponding filling of energy levels is shown schematically in 22*b*. A physical picture which is helpful in understanding this result may be obtained by considering the distribution of energy levels, Fig. 22*a*, to be the cross-section of a trough or tank. If we pour water into this tank it will fill to a certain level, E_1. If we let each molecule of water in the tank represent an electron in the crystal, then the distribution in energy of the electrons is correctly represented by the distribution in height of the molecules. Thermal agitation is represented by shaking the tank; this will produce surface ripples as in Fig. 22*d* which represent crudely the Fermi-Dirac distribution.

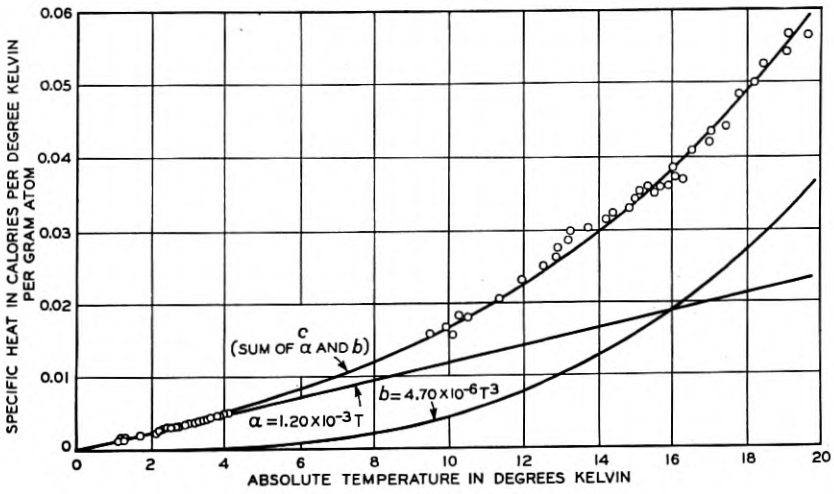

Fig. 23—Specific heat of iron at low temperature.

Under certain conditions, however, the electronic specific heat is not negligible. We have seen that the number of electrons participating in specific heat is proportional to kT and that these have a more or less normal specific heat. Hence the electronic specific heat is proportional to T. On the other hand, at low temperatures the Debye specific heat is proportional to T^3. Hence for sufficiently low temperatures the electronic specific heat is the larger. In Fig. 23 we give the specific

heat of iron near absolute zero.[29] The theoretical curve c, which is seen to represent the experimental data quite well, is the sum of two terms represented by curves a and b. a is linear in the temperature and represents the electronic specific heat while b is cubic and represents that due to lattice vibrations. Numerical calculations from theory of the slope of curve a which could be compared with the observed slope are not available. Curve b, we have said, is just the Debye curve and is drawn as if the Debye temperature were 462°, a value which is in good agreement with 453°, the value deduced from the specific heat at higher temperatures in connection with Fig. 21. At very high temperatures the electronic specific heat will again be of importance. But at high temperatures it is necessary to apply corrections to the Debye theory and the writer is not acquainted with any unambiguous evidence for electronic specific heat in that case.

Thus we see that only a very small fraction of the electrons of a partially filled band contribute to the specific heat. It is the Pauli principle which restrains the remainder. We shall see in the next paper why the Pauli principle does not interfere with the conduction of electricity. For the case of an insulator—that is, a crystal each of whose bands is either wholly filled or wholly empty—it is still harder for electrons to arrive at empty states and the electronic specific heat is quite negligible. Hence all of the specific heat for an insulator is of the atomic vibration type discussed in the Debye theory.

The Theory of Thermal Expansion

In order to understand the theory of thermal expansion we must study the curve representing energy versus lattice constant for the solid. This is shown qualitatively in Fig. 24. We note that the energy curve is unsymmetrical about its minimum. We may describe its behavior by saying that it is harder to compress than to expand the solid. This statement is illustrated by a comparison of the expansion and the compression which can be produced by a given energy E; it is seen that the asymmetry of the curve causes the expansion produced by this energy to be greater than the compression. Now the origin of thermal expansion is as follows: owing to thermal agitation—that is, atomic vibrations—regions of the crystal are alternately expanding and contracting; since the expansions occur more readily than the contractions, there is on the average a net expansion. The greater the temperature the greater this net expansion; hence we find that the size of the solid increases with increasing temperature. This explanation of thermal expansion can be made clearer by considering, not a solid,

[29] W. H. Keesom and B. Kurrelmyer, *Physica 6*, 633 (1939).

but a diatomic molecule. Suppose Fig. 24 gives the dependence of the energy of a molecule upon the internuclear distance. Suppose the molecule is given vibrational energy corresponding to E on the figure.

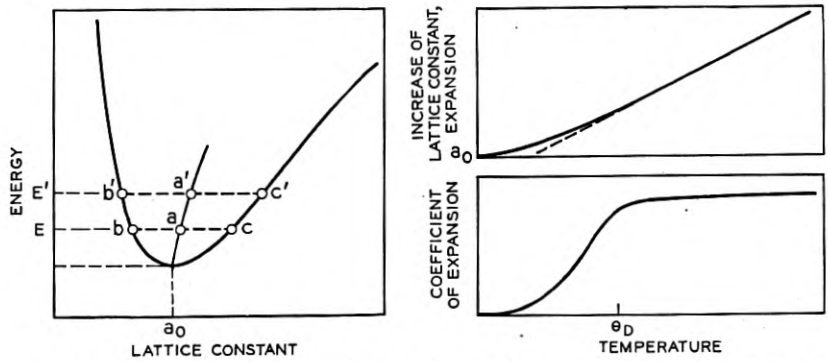

Fig. 24—The theory of thermal expansion. The asymmetry of the curve for the energy of a crystal versus the lattice constant is responsible for the thermal expansion.

Then the nuclei will vibrate between positions b and c on the figure. Since c lies more to the right of the equilibrium position than b does to the left, the mean distance of separation, a, lies to the right of a_0. Increasing the vibrational energy to E' increases the mean separation to a'. This shows that the asymmetry of the potential curve results in a continuous increase in mean internuclear separation with increasing energy of vibration. A crystal is, in a sense, an assemblage of diatomic molecules, each pair of nearest neighbors having a potential energy curve like that of Fig. 24, and its expansion is explained in the same way.

The theory outlined above can be made quantitative. From it we obtain the interesting result that the thermal expansion coefficient is proportional to the specific heat. This is a rather natural result: we have seen that the total expansion is proportional to the thermal energy; hence the rate of expansion with increasing temperature, i.e. the thermal expansion coefficient, should be proportional to the rate of increase in thermal energy with increasing temperature, i.e. to the specific heat. The relationship embodying this statement is known as Grüneisen's law and is expressed by the equation

$$\alpha = \gamma \frac{K}{V} C_V, \tag{10}$$

where α is the volume coefficient of thermal expansion (three times the linear coefficient), K is the compressibility, V the volume of one gram

atom, and C_V the specific heat per gram atom at constant volume. γ is a parameter which measures the asymmetry of the curve and is defined as follows: if we think of the solid as being compressed by an external pressure, the forces between the atoms will change and the Debye temperature will increase. If the curve were a parabola—that is, perfectly symmetrical—the Debye temperature would not change. γ is defined by the relationship

$$\gamma = -\frac{\partial \ln \theta_D}{\partial \ln V}. \tag{11}$$

The γ, K, and V are nearly constant for a given substance. Hence the thermal expansion curve is practically the same as the specific heat curve except for a constant factor. In Fig. 25 we give the thermal

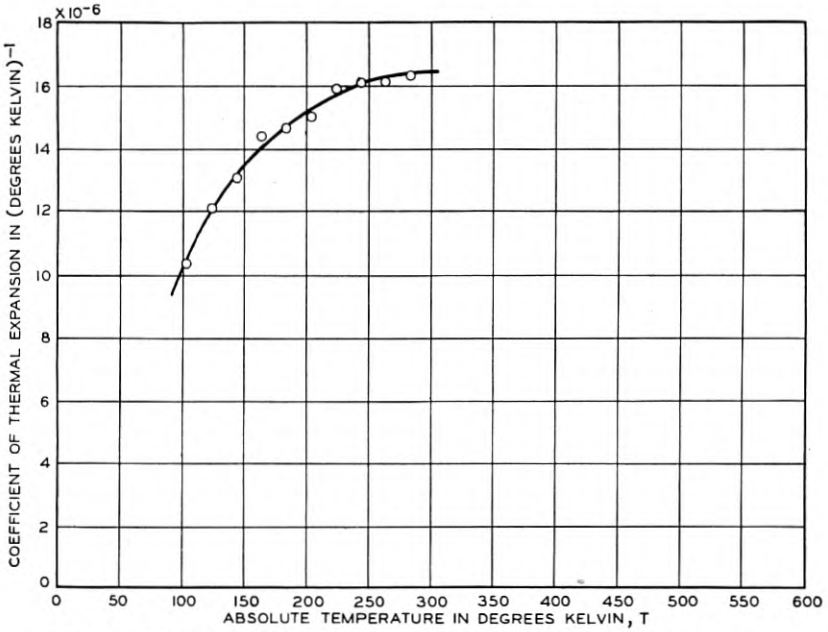

Fig. 25—Coefficient of thermal expansion versus temperature for copper.

expansion of copper.[30] The Debye temperature was chosen to give the best fit. We see that the theory of thermal expansion gives as good agreement with experiment as does the theory of specific heat. If Grüneisen's law were perfectly satisfied, the same Debye temperature would be found for both the thermal expansion and the specific heat curves. The relatively small difference between the two values, 325 for expansion and 315 for specific heat, is a measure of the validity of Grüneisen's law.

[30] E. Grüneisen, *Handbuch der Physik*, X, p. 43 (1926).

Grüneisen's law applies only to simple crystals; we shall see in the next section that it is not applicable to the anomalous expansion associated with ferromagnetic transformations nor is it applicable to the abnormal expansions of the order-disorder transformations in alloys.

Magnetic Effects

In this section we return to a discussion of the energy band theory and this time introduce the magnetic moment associated with the spin of the electron. It is the spin magnetic moment which when added to the concept of energy bands leads to explanations of para and ferromagnetism.

When a body is placed in a magnetic field it becomes magnetized; in other words it acquires a magnetic moment. Ferromagnetic materials become very easily magnetized in the field with their magnetic moments parallel to the field and they may remain magnetized after the field is removed. Paramagnetic materials are also magnetized in the direction of the field but only very weakly compared to ferromagnetic materials and only while they remain in the field. Diamagnetic materials are magnetized in a direction opposite to the field and, like paramagnetic substances, only weakly and while in the field. These magnetic effects are produced by the electrons in two distinct ways. In the first place, the motion of the electron as a whole produces a current and this current, like the ordinary macroscopic currents in a wire, produces a magnetic field. Conversely, an externally applied magnetic field affects the motions of the electrons in a body and can thereby magnetize it; this process accounts for the diamagnetism of diamagnetic bodies but it may contribute to the paramagnetism as well. It is not with this first way in which electrons can behave magnetically but rather with the second way, described below, that we shall be concerned. The first way, which is mentioned for completeness, involves a theory too complicated for treatment in this article. In the second place, an electron can behave magnetically by virtue of its spin: the rotation of the electron about its own axis produces a magnetic moment which is anti-parallel—because the charge of the electron is negative—to the angular momentum due to the spin. A magnetic field tends to align the spin magnetic moments of the electrons and to make them contribute to the paramagnetism. We shall see below that this process accounts for the paramagnetism of nonferromagnetic metals. We shall see also that the magnetism of ferromagnetic bodies is due to the magnetic moment of the electron spin but that the energy involved in the theory of ferromagnetism is not an interaction between the magnetic dipoles of the electrons but is

instead an electrostatic exchange energy like that discussed for atoms in connection with Figs. 4 and 6.

Paramagnetism and Diamagnetism

Let us consider first the so-called "weak spin paramagnetism." This occurs in metals, since they have partially filled bands. In the presence of a magnetic field the spin of the electron is quantized so that the component of its angular momentum in the field direction is either $+\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$ where $\hbar = h/2\pi$ (h = Planck's constant) is the quantum mechanical unit of angular momentum. The corresponding components of magnetic moment along the field are $-\mu_\beta$ and $+\mu_\beta$ where μ_β is the quantum mechanical unit of magnetic moment known as the Bohr magneton. Letting $-e$ be the charge and m the mass of the electron and c be the speed of light, we have from the quantum theory

$$\mu_\beta = e\hbar/2mc. \tag{12}$$

The ratio of mechanical moment (i.e. angular momentum) to magnetic moment, taken without regard to sign, is called the "gyro-magnetic ratio." For the spin of the electron its value is mc/e, but for the motion of the electron as a whole, its value is $2mc/e$. Because of the difference between these two values, experimental measurements of the gyromagnetic effect play a decisive role in the experimental verification of the electron spin theory of ferromagnetism in a way which we shall describe below.

Half the quantum states in an energy band of a crystal have angular momentum components along the magnetic field of $\frac{1}{2}\hbar$ and the other half of $-\frac{1}{2}\hbar$. In Fig. 26*a*, we have divided the states in the band into two groups, corresponding to the two spins. We shall refer to one of these as "the band with plus spin" and to the other as "the band with minus spin." When a magnetic field is applied, the energies of the electrons are changed. Thus if an electron in the lowest state of the band with minus spin has an energy E_0 before the field is applied, it has an energy of $E_0 - \mu_\beta H$ afterwards; the second term represents, of course, the energy of the magnetic dipole μ_β when parallel, as distinguished from anti-parallel, to the field—the situation for minus spin. All the states in the band with minus spin will be thus altered in energy. Similarly all the states in the band with plus spin are displaced upwards in energy by $\mu_\beta H$. This is the situation represented in Fig. 26*b*. After the displacement we find that some of the electrons in the band with plus spin have higher energies than empty states in the band with minus spin; such an arrangement is not stable and the electrons will change their quantum states so as to produce the lowest energy possible consistent with the distribution of energy levels shown

in Fig. 26*b* and with Pauli's principle. The arrangement of lowest energy is shown in Fig. 26*c*; electrons have shifted from the band of plus spin to states of lower total energy in the band of minus spin until the two bands are filled to the same energy level, indicated by the solid horizontal line. As the figure shows, the number of electrons shifted will be the number lying in the energy range $\delta E = \mu_\beta H$.[31]

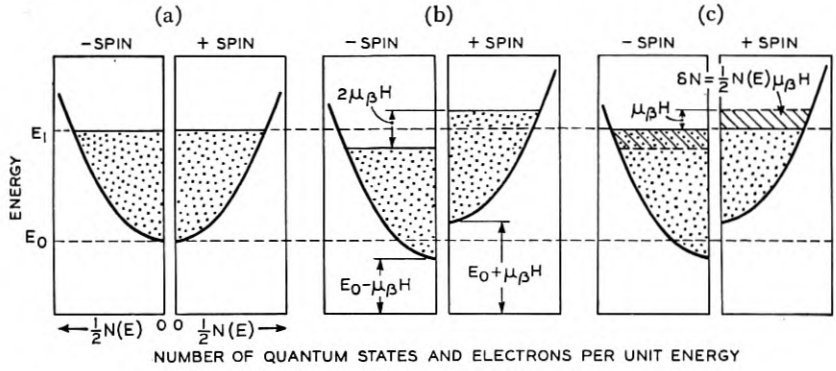

Fig. 26—The paramagnetism of free electrons.
(a) Distribution of electrons in energy.
(b) Displacement of levels by a magnetic field.
(c) Distribution of electrons in energy in a magnetic field.

The number of states, δN, lying in this energy range in the band of plus spin, which contains of course half the states in the band, is according to equation (3)

$$\delta N = \tfrac{1}{2}N(E_1)\delta E = \tfrac{1}{2}N(E_1)\mu_\beta H. \tag{13}$$

The magnetic moment of these states is

$$\delta M_+ = -\mu_\beta \delta N = -\tfrac{1}{2}N(E_1)\mu_\beta^2 H. \tag{14}$$

The minus sign occurs because the angular momentum and the magnetic moment of an electron are in opposite directions; the states of plus spin have minus moments in Fig. 26.

The electrons that occupied these states before the field was applied now occupy states with minus spin and produce a magnetic moment of

$$\delta M_- = \tfrac{1}{2}N(E_1)\mu_\beta^2 H. \tag{15}$$

Hence the minus band gains a plus moment and the plus band loses a

[31] We have here assumed that the fractional change in $N(E)$ in the interval $\mu_\beta H$ is negligible; this assumption is reasonable. For a field of 10,000 gauss, $\mu_\beta H$ is only 5.77×10^{-5} ev while $E_1 - E_0$ is of the order of several ev.

minus moment and, since the net moment of Fig. 26*a* is obviously zero, the net moment produced by the magnetic field is

$$\delta M = \delta M_- - \delta M_+ = N(E_1)\mu_\beta^2 H. \tag{16}$$

The susceptibility of a material, denoted by χ, is defined as the magnetic moment produced per unit volume per unit field:

$$\chi_s = \frac{\delta M}{VH} = \frac{N(E_1)}{V}\mu_\beta^2. \tag{17}$$

The subscript "s" is a reminder that this susceptibility was produced by the spin magnetic moment of the electron.

Since the moment produced is in the direction of the field, χ_s is positive; the susceptibility is of the paramagnetic type. As for its magnitude: in the monovalent metals, as we have said before, the distribution of levels in the bands is well approximated by the free electron formula (4). Using this, we find

$$\chi_s = \frac{4\pi}{h^3}(2m)^{3/2}(E_{\max})^{1/2}\mu_\beta^2. \tag{18}$$

where $E_{\max}$ $(= E_1 - E_0)$ is the maximum kinetic energy in the band.

Before comparing susceptibilities calculated from this expression with experimental values, we must discuss diamagnetism. The electrons in the partially filled band of Fig. 26 give formula (18) because of their spin magnetic moments. They give a susceptibility also because of their motion through the crystal. For the case of free electrons, this susceptibility is negative—that is, it is a diamagnetic susceptibility, and, according to a theory we cannot discuss here, in magnitude it is one third of χ_s. Denoting it by χ_m ("m" for motion of the electron as a whole), we have

$$\chi_m = -(1/3)\chi_s. \tag{19}$$

The electrons in the filled bands, corresponding to electrons in closed shells in the ionic cores of the metal, also give rise to diamagnetism. They can give no spin paramagnetism because there is no possibility of transferring electrons from a *filled* band of one spin to a *filled* band of the other spin—this would require putting more electrons in the band of one spin than it has quantum states, a violation of Pauli's principle. Denoting by χ_i the susceptibility of the ionic cores of the metal, we have for the net susceptibility χ the equation

$$\chi = \chi_s + \chi_m + \chi_i. \tag{20}$$

Specializing this for the case of free electrons in the valence electron

band gives

$$\chi = (2/3)\chi_s + \chi_i = \frac{8\pi}{3}\left(\frac{2m}{h^2}\right)^{3/2}\mu_\beta^2(E_{max})^{1/2} + \chi_i. \qquad (21)$$

In Table IV we give theoretical and experimental values for the susceptibilities of the simple metals. The values of χ_i are obtained from theory for lithium and by experiment for the other metals.

TABLE IV

MAGNETIC SUSCEPTIBILITIES *

	Li	Na	K	Rb	Cs
χ_s	1.5	0.68	0.60	0.32	0.24
χ_i	−0.1	−0.26	−0.34	−0.33	−0.29
$\chi = \frac{2}{3}\chi_s + \chi_i$	0.9	0.2	0.06	−0.12	−0.15
χ observed †	0.5	0.51	0.40	0.07	−0.10

* This Table is taken from N. F. Mott and H. Jones, "The Theory of the Properties of Metals and Alloys," Oxford 1936, p. 188.

† K. Honda, Ann. d. Physik *32*, 1027 (1910) and M. Owen, Ann. d. Physik *37*, 657 (1912).

Although equation (17) for the spin susceptibility χ_s in terms of $N(E_1)$ is generally true, the relationship that $\chi_m = -\chi_s/3$ is true only for the case when $N(E)$ is the free electron distribution.[32] For some metals $N(E)$ differs greatly from that for free electrons and then larger values of χ_m may occur. The high diamagnetism of bismuth is explained in this way. In the next paper, we shall discuss the meaning of the freeness of electrons; however, a discussion of electron diamagnetism lies beyond the scope of this paper.[33]

Ferromagnetism

The shift of electrons from one band to another for the paramagnetic behavior shown in Fig. 26 persists only so long as the magnetic field is applied. When the magnetic field is removed, the stable arrangement is as shown in Fig. 26*a*, equal numbers of electrons having each spin. The situation is quite different in ferromagnetic materials and, for reasons discussed below, in the stable arrangement there are many more electrons of one spin than of the other.

Two things are important for the occurrence of ferromagnetism: the exchange effect as illustrated in Figs. 4 and 6 and the structure of the bands arising from the 3*d* levels. The 3*d* levels, as is shown in

[32] An even more stringent condition is actually required.

[33] For the diamagnetism of electrons in closed shells the reader is referred to K. K. Darrow's article, "The Theory of Magnetism," *Bell System Technical Journal*, Vol. XV, 1936, and in particular to Page 247.

Fig. 6, are only partially filled for free atoms of the ferromagnetic elements iron, cobalt, and nickel. We shall show first how the partially filled 3*d* bands together with exchange forces can produce ferromagnetism and later discuss the theory of why only the last three of the eight transition elements are ferromagnetic.

The splitting of the atomic energy levels into bands is shown in

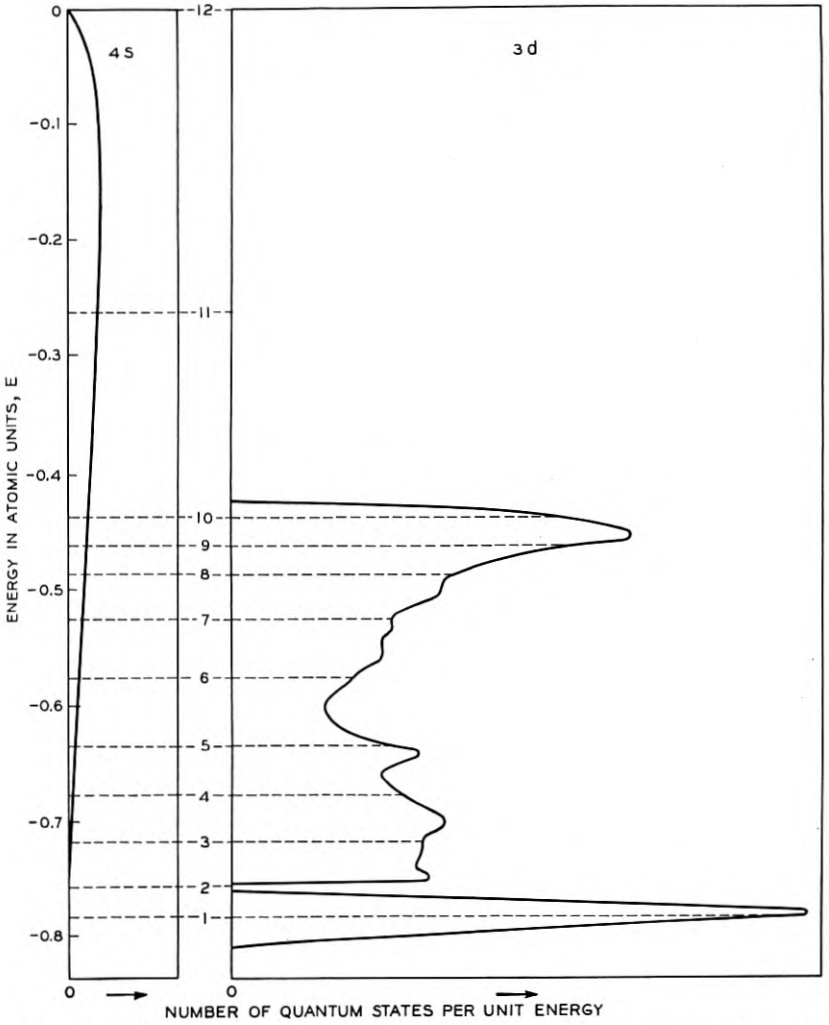

Fig. 27—Distribution of states in energy for copper. The distribution is probably quite similar for iron, cobalt, and nickel and in the absence of calculations for these other metals, this figure will be used for them. The total number of quantum states per atom in the 4*s* and 3*d* bands having energies less than the ordinates of the dashed lines are given by the corresponding integers.

Fig. 16. The 3*d* levels give a band capable of containing ten electrons per atom, five with each spin; and the 4*s* band can hold two electrons per atom, one with each spin. Curves representing $N(E)$ for these bands, calculated for the case of copper by Slater and Krutter, are shown in Fig. 27. We see that the 4*s* band is much wider in energy than the 3*d* and that it contains only one-fifth as many electronic states. The band structure will be similar for all the transition elements; the energy scales, however, will be different. As is shown in Fig. 6 the 3*d* electrons are more tightly bound for copper than for nickel or chromium. Corresponding to this tighter binding, the 3*d* wave functions of copper extend less in space than those of nickel and chromium and consequently they overlap less between atoms and the 3*d* band is narrower for copper. Progressing towards decreasing atomic number in the sequence of elements from copper to scandium, the 3*d* band will continually widen; and this widening, as we shall see later, can help account for the absence of ferromagnetism for the elements before iron in the periodic table.

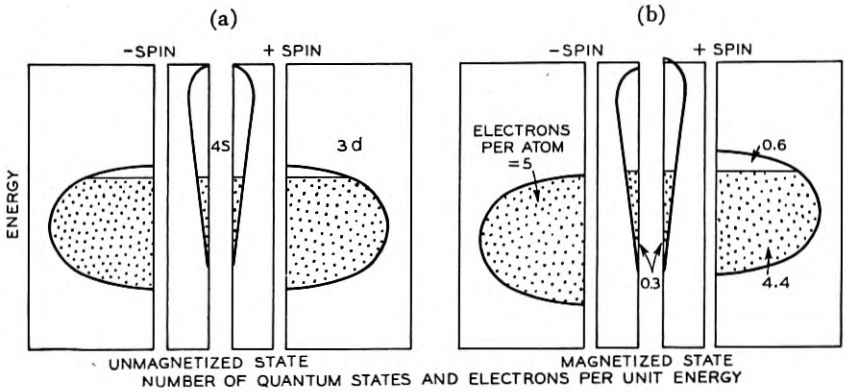

Fig. 28—The ferromagnetism of nickel.

In Fig. 28, we give a simplified representation of the 4*s* and 3*d* bands split into two sets according to the spin. (We may, if we wish, suppose that a magnetic field is applied along which the spin is quantized, but that the field is so weak that the displacement of the energy bands produced by it is negligible; this supposition is not necessary, however, for regarding the spin we shall need only the fact that all the electrons in the + spin band have parallel spins which are anti-parallel to those in the − spin band.) For the element nickel there are 28 electrons, 10 of which are in the 3*d* and 4*s* bands. They can fill the bands as indicated in Fig. 28*a*. Let us compare this distribution with the electron

configuration of the atom, Fig. 6; we see there that there are unequal numbers of electrons of the two spins. This inequality is produced by the exchange effect which lowers the more occupied set of 3*d* levels in respect to the less occupied set and produces a stable arrangement with the 3*d* levels of one set completely filled. This exchange effect operates in the same way in metallic nickel. In Fig. 28*b* we show the distribution which results when electrons are shifted from the 3*d* band of plus spin to that of minus spin until the latter is filled. The exchange effect produces the displacements of the bands as shown. The arrangement in Fig. 28*b* is stable; in order for electrons to be transferred from the filled minus 3*d* band to the plus band, they would have to increase their energy, a fact which is expressed by drawing the diagram so that the lowest vacant quantum states are appreciably above the highest energy state of the filled 3*d* band. Thus for nickel an unbalanced distribution of spins prevails both for the free atom and the metal.

The Energy of Magnetization

The argument presented above for the stability of the magnetized state shown in Fig. 28*b* is not really rigorous. We saw that if one electron was transferred from the filled 3*d* band to one of the vacant states, its energy and, therefore, the energy of the crystal would be raised. In other words, the magnetized state has less energy than a state which is slightly less magnetized. This fact in itself does not prove that the magnetized state is stable; it proves only that it is metastable—i.e., that its energy is less than the energy of other states which differ from it slightly; in order to establish the stability of the magnetized state, it is necessary to prove that its energy is less than the energy of any other state including that of the unmagnetized state shown in Fig. 28*a*. We may illustrate this necessity by considering the following hypothetical behavior: as the magnetization is reduced from that of Fig. 28*b* to zero (the value for Fig. 28*a*), the energy might at first increase and then decrease—decreasing so much finally that the energy would be lower for the unmagnetized than for the fully magnetized state. We shall, therefore, discuss the difference in energy between the fully magnetized and unmagnetized states; theory shows that this quantity is the fundamental one whose value determines whether or not ferromagnetism occurs.

Let us consider the change in energy in going from the unmagnetized state to the magnetized state in Fig. 28. This change in energy can be separated into two contributions, one positive and one negative. The positive contribution comes from an increase in "Fermi energy" or

"energy of motion," which was discussed in connection with the binding energy of metals. This energy is positive because after the shift to the magnetized state, electrons have moved from states in the band of plus spin to states which lie higher—in respect to the bottom of the bands in both cases—in the band of the minus spin; that is, the electrons which have moved from one band to the other have all gained "energy of motion." The negative contribution to the energy comes from the exchange effect. This causes the lowering of the filled band and the raising of the unfilled band; since there are more electrons in the lowered band than in the raised band, there is a net decrease in energy due to this exchange effect. Thus we have a positive change in Fermi energy and a negative change in exchange energy in going from the unmagnetized to the magnetized state. If the exchange energy has a greater change than the Fermi energy, the energy of the magnetized state is lower and the metal is ferromagnetic.

No satisfactory calculations have as yet been made for these energy differences. In order to calculate them, accurate values for the distribution of states in the $3d$ band are needed, and the mathematical methods available for computing this distribution are not as yet very satisfactory. Next the exchange effect energy must be found; this is also difficult to calculate accurately. Finally, the description given here is over-simplified; in particular another energy term, known as the correlation energy, must be included; this energy acts somewhat like an exchange energy but between the bands of different spins and it tends to cancel out the exchange energy. Although these difficulties greatly mar the usefulness of the theory of ferromagnetism represented in Fig. 28, this theory is able to correlate a large amount of experimental material in a very natural way; and since it is the theory based on the concepts of energy bands, it is the one that we shall discuss in this paper. In passing, however, we must state that there are other theories of ferromagnetism which in some ways are more successful and in other ways less successful than the band theory. Some of these are atomic rather than band theories. An example of this type of difference in method of attack was given in the discussion of the binding energy of sodium chloride; two treatments were given: for one the basis being the ions and for the other the energy bands. In the case of sodium chloride, however, the theoretical equivalence of the two methods is easily demonstrated. In the case of ferromagnetism, the two theories are not equivalent and are both simplifications of a more complex and as yet unsatisfactorily explored intermediate case.

Although no satisfactory calculations of the energy difference between the magnetized and unmagnetized states of metals exist, the

theory must be regarded as representing great progress over non-wave-mechanical theories. The reason is this: in older theories of ferromagnetism the energy was supposed to come from the magnetic interaction between the magnetic dipoles, and it turned out that the energies calculated in this way were at least a thousandfold too small. The energies calculated in the new theory are adequate in magnitude but have nothing to do with the magnetic moment of the electron; they arise from the exchange energy, which is, as we have said before, an electrostatic energy resulting from the wave-mechanical treatment of Pauli's principle. It is the laws governing the spin quantum number of the electron, not the magnetic moment, which are responsible for the energy of magnetization; the externally observed magnetic field of a ferromagnetic material is merely a superficial indication of more fundamental electrostatic forces.

Intrinsic Magnetization

According to our theory, the low energy state and therefore the stable state of metallic nickel is a magnetized one. If one picks up a piece of nickel at random, however, it may not appear to be magnetized. This apparent absence of magnetism is due to the presence of "domains." According to the domain theory—which is a very well established branch of magnetic theory—a block of nickel will consist of a number of microscopic domains, each highly magnetized, but having their magnetic moments pointing at random in a number of directions so that on the average there is no magnetism. The application of a magnetic field aligns the magnetic moments of these domains and, since they are then all parallel, one can measure the total magnetization of the sample. A field strong enough to line up all the domains is said to produce "saturation" because a further increase in field will give no further increase in magnetization. It is customary and convenient to divide the total or saturation magnetic moment of the material by the total number of atoms, thus finding the average magnetic moment per atom, and to express this value in Bohr magnetons. The resultant value is called the intrinsic magnetization per atom and is denoted by β.[34] For example, if a crystal had one electron per atom and all the electrons had their spins parallel, then all their magnetic moments would be parallel, too, and the intrinsic magnetization would be unity, $\beta = 1$.

For nickel the intrinsic magnetization is 0.6 Bohr magnetons per atom. The following argument shows how easily such a fractional number can be accounted for by our theory. Nickel has 10 elec-

[34] The "intrinsic magnetization" is customarily defined as the magnetic moment per unit volume when the moments of the domains are parallel.

trons per atom in the 3*d* and 4*s* bands. The 3*d* band with minus spin is supposed full, containing five electrons per atom. The 4*s* band (both spins) can contain two electrons per atom, and from Fig. 27 we see that it is about one-fourth full; suppose it contains 0.6 electrons per atom; the remaining electrons go to the 3*d* band with plus spin which is not quite full but has a "hole" in it of 0.6 electrons per atom. There are equal numbers of electrons of each spin in the 4*s* band and their magnetic moments cancel.[35] The net magnetic moment arises from the unbalance of 0.6 electrons per atom between the two parts of the 3*d* band. This unbalance will correspond to a magnetization of 0.6 Bohr magnetons. The theory is not capable of predicting the number 0.6 exactly; however, this number is entirely consistent with what can be said about the distribution of levels in the band. In the "atomic" theories of magnetism, it is supposed that each atom has a certain magnetic moment. From the results of the gyromagnetic experiments,[36] one concludes that the magnetization is due to electron spin. Since an atom whose magnetism is due to electron spin must have a magnetic moment equal to an integral multiple of the Bohr magneton,[37] the "atomic" theory is forced to assume that 40 per cent of the nickel atoms are unmagnetized and that 60 per cent have one Bohr magneton, or else that 70 per cent are unmagnetized and 30 per cent have two Bohr magnetons or at any rate that there are at least two kinds of atoms. These rather awkward assumptions are not required in the band theory, the reason being, as is suggested in Fig. 28, that the electrons are not thought of as belonging to the atoms individually but to the crystal as a whole.

The intrinsic magnetization of ferromagnetic material decreases with increasing temperature. In the band theory this is explained as follows: at a temperature T some of the electrons are excited from the filled band to the partially filled band; as the temperature is increased more are shifted. Furthermore, if we compare the effects of two equal increments of temperature, one occurring at a higher temperature than the other, the one at the higher temperature will have the greater effect. This is because at the higher temperature more electrons have been shifted; hence the exchange effect displacement of the band of one spin in respect to the band of the other spin is less and electrons need

[35] Actually there will be a slight exchange effect in the 4*s* band; however, it will be so slight that the magnetic moment produced can be neglected

[36] If a piece of iron is suspended so that it can rotate and then is magnetized, it will acquire an angular momentum. The ratio of angular momentum to magnetic moment should be mc/e if the magnetization arises from electron spin and $2mc/e$ if it arises from motion of the electron as a whole. Experiment gives the following fractions of the former value: for iron 1.03, for cobalt 1.23, for nickel 1.05.

[37] For a discussion of this theorem see K. K. Darrow's article, "Spinning Atoms and Spinning Electrons," *Bell Sys. Tech. Jour.*, XVI, 319 (1937).

not gain so much energy to shift from the more full to the less full band. Hence at the higher temperature there is more decrease in magnetization per degree rise in temperature than at the lower temperature. A logical consequence of this reasoning is that the magnetization decreases more and more rapidly as the temperature increases

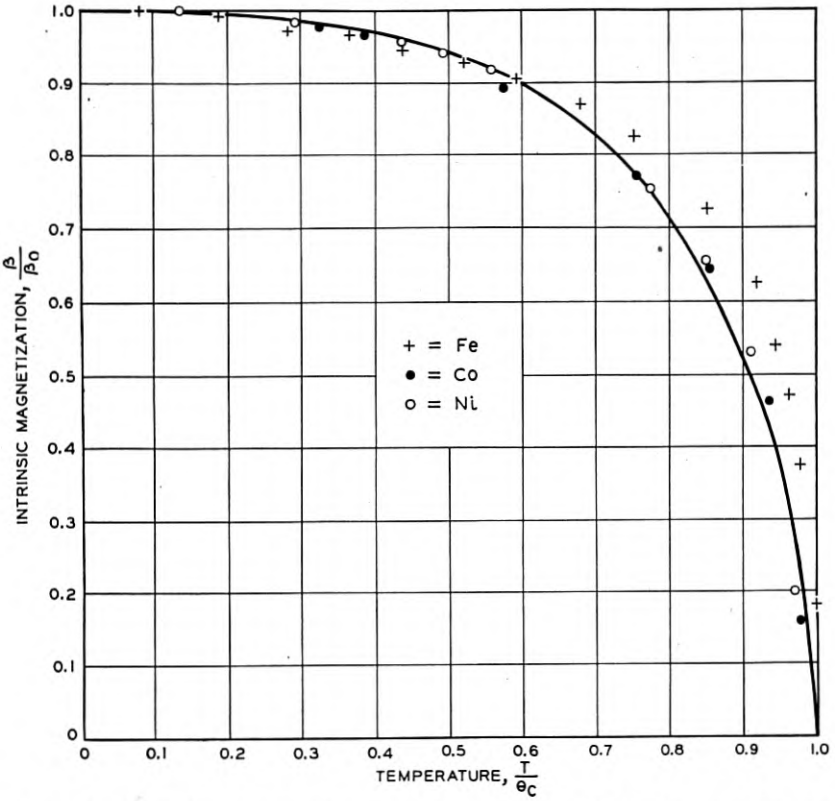

Fig. 29—Intrinsic magnetization versus temperature. The horizontal scale represents the temperature divided by the Curie temperature and the vertical scale, the intrinsic magnetization divided by the intrinsic magnetization at absolute zero. The theoretical curve is derived from quantum mechanics.

and becomes zero at a certain critical temperature, which is known as the Curie temperature and denoted by θ_c. A more complete discussion of the theory of the temperature dependence of magnetism would belong in a paper devoted solely to the theory of magnetism.[38] In

[38] See, for example, K. K. Darrow, *Bell Sys. Tech. Jour.* XV, 224 (1936), R. M. Bozorth, "The Present Status of Ferromagnetic Theory," *Bell Sys. Tech. Jour.*, XV, 63 (1936) and texts such as J. H. Van Vleck, "The Theory of Electric and Magnetic Susceptibilities," Oxford, 1932, E. C. Stoner "Magnetism and Matter," Methuen and Company, Ltd., London, 1934, and F. Bitter "Introduction to Ferromagnetism," McGraw-Hill Book Co., New York, 1937.

this paper we shall use the fact that the magnetism changes with the temperature to explain the anomalous expansion of ferromagnetic materials. In Fig. 29 we show the variation in intrinsic magnetization with temperature as observed for iron, cobalt and nickel.

Variation of Intrinsic Magnetization with Composition

Let us consider how the intrinsic magnetization should vary from element to element in the transition series, supposing always that the temperature is so low that thermal effects can be neglected. The element next to nickel is cobalt; cobalt has one less electron than nickel so that the $3d$ band and partially filled $4s$ band for it will have one less electron in them. Because of the relatively small number of quantum of states in the $4s$ as compared to the $3d$ band, this deficit will be made up mainly by the $3d$ band which will therefore contain not 4.4 as for nickel but instead 3.4 electrons leading to an unbalance of 1.6 Bohr magnetons per atom. The observed β for cobalt is 1.7 in good agreement with this.

One can obtain electron atom ratios intermediate between cobalt and nickel by forming alloys. We shall speak of the electron concentration, C, of these and other alloys, meaning by this term the total number of electrons available for the $3d$ and $4s$ bands divided by the total number of atoms. So long as the minus spin half of the $3d$ band remains full and so long as the number of electrons in the $4s$ band does not vary much, the value of β will be a linear function of the electron concentration varying from ~ 1.6 to ~ 0.6 as the concentration varies from 9 for cobalt to 10 for nickel. In Fig. 30 are given the intrinsic magnetizations plotted against electron concentration for a series of alloys. It is seen that from cobalt to about halfway between nickel and copper, an increase in C produces, very nearly, a numerically equal decrease in β. This means that the increase in C goes toward filling up the holes in the $3d$ band and reducing the unbalance and hence β. Some alloys are included in Fig. 30 for which the two elements are not adjacent in the periodic table; their values of β also conform to the values predicted from their electron concentrations.

The very natural way in which the band theory accounts for the results shown on Fig. 30 is its principal success in the theory of ferromagnetism.

The bend in the curve between iron and cobalt is not very satisfactorily explained at present. One theory is that for iron neither $3d$ band is entirely full; but this explanation is said to be inconsistent with the observed dependence of magnetization upon temperature at low

temperatures. Another theory is that there are only 0.2 electrons in the 4*s* band, thus leaving the remaining 7.8 electrons of iron distributed 5 to one 3*d* band and 2.8 to the other, leaving an unbalance of 2.2; this theory is unsatisfactory because it would require an inexplicable displacement upwards of the 4*s* band compared to the 3*d* in going from

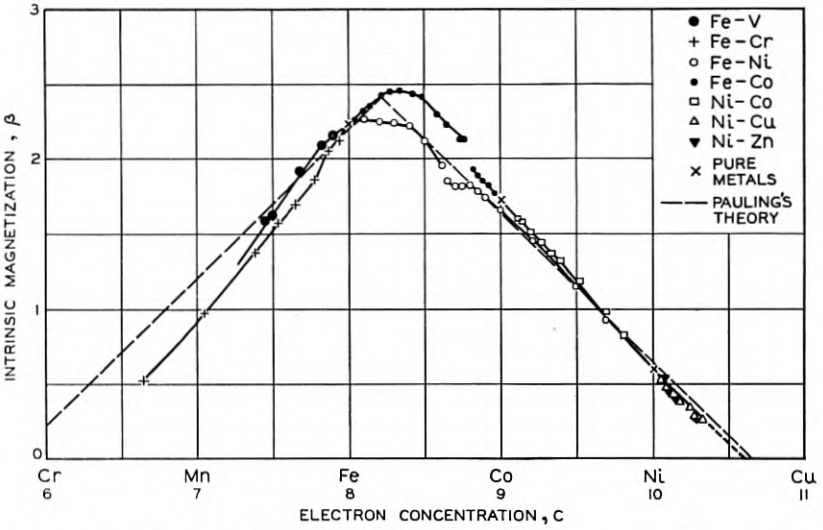

Fig. 30—Intrinsic magnetization versus electron concentration.

The data for this figure were obtained from the following sources:
Fe-V and Fe-Cr M. Fallot Ann. de Physique 6, 305–387 (1936).
Fe-Co R. Forrer J. de Physique et le Radium, 1, 325–339 (1930).
Fe-Ni M. Peschard Comptes Rendus 180, 1836 (1925).
Ni-Co P. Weiss, R. Forrer, and F. Birch Comptes Rendus 189, 789–791 (1929).
Ni-Cu and Ni-Zn V. Marian Ann. de Physique 7, 459–527 (1937).

cobalt to iron. Another theory has been proposed by Pauling [39]; he has stated it in the "atomic" language but it can be translated into the band language as follows: the 3*d* band is broken into two parts, an upper part containing 4.88 levels per atom, 2.44 for each spin, and a lower part separated from the upper by an energy gap and containing 5.12 levels per atom, 2.56 for each spin. A number of electrons per atom varying from 0.6 for nickel to 0.7 for cobalt are in the 4*s* band; for simplicity we shall suppose that this number has a constant value of 0.65 electrons per atom. According to this simplification, one of the upper parts of the 3*d* band has 0.65 holes for nickel. This band will become empty if the electron concentration is decreased by 1.79 (= 2.44 − 0.65)—that is, for a concentration of 10 − 1.79 = 8.21.

[39] L. Pauling, *Phys. Rev.*, *54*, 899 (1938).

If the concentration is decreased below 8.21, electrons will be removed from the upper part with the other spin; this will result in a decrease in the unbalance and hence in β, which has for $C = 8.21$, a value of 2.44—corresponding to one filled and one empty upper part; and this decrease will be numerically equal to the decrease in C. Accordingly, the value of β for iron, $C = 8$, is $2.44 - 0.21 = 2.23$. The numbers 2.44 and 2.56 were, of course, chosen so as to obtain this agreement for iron. This theory of Pauling expresses reasonably well the variations in β for all the alloys of Fig. 30.

Criterion for Ferromagnetism

We must now see how the theory explains the absence of ferromagnetism for the remaining transition elements. We have seen that the exchange energy lowers and the Fermi energy raises the energy of the magnetized state compared to the unmagnetized state. These two effects very nearly cancel even for the magnetic elements iron, cobalt, and nickel. For the other elements in the transition series, which are not ferromagnetic, the Fermi term apparently exceeds the exchange term. We shall give a theoretical reason for expecting this result.

In the first place we must indicate how nearly the effects cancel. Let us take cobalt, which has nine electrons in the $3d$ and $4s$ bands, as an example. From Fig. 27 we see that for cobalt in the unmagnetized state both $3d$ bands are filled to about -0.46 atomic units. In the magnetized state one band is filled by electrons which have come from levels with less energy of motion in the other band. Since the top of the $3d$ band comes at about -0.42 units on Fig. 27, the average gain in energy for each transferred electron is about 0.04 units. Since the number of electrons transferred is 1.7 per atom, the increase in Fermi energy is 0.068 atomic units or 0.9 ev per atom. From an analysis of thermal measurements the value for the actual energy of magnetization is found to be about 0.2 ev per atom; a value which is only about one fourth of the predicted increase in the Fermi energy. Hence the exchange energy exceeds the Fermi energy by only 25 per cent and the two energies nearly cancel.

The variation in the structure of the $3d$ band from element to element was discussed in connection with Fig. 27; we concluded then that the bands become wider as we recede in the periodic table from nickel towards scandium. Greater band width means greater Fermi energy in the magnetized state and this effect opposes the occurrence of ferromagnetism. The exchange energy can also change. Calculations by Slater,[40] which unfortunately are too over-simplified to bear much

[40] J. C. Slater, *Phys. Rev.*, *49*, 537, 931 (1936).

weight, show that for manganese the Fermi energy outweighs the exchange energy so that manganese is not ferromagnetic at all. In Fig. 31 we represent the state of affairs predicted for chromium; the exchange

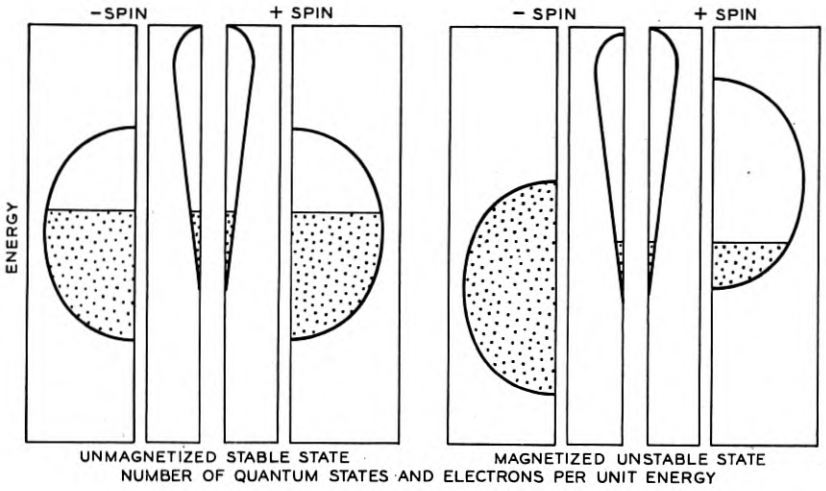

Fig. 31—The absence of ferromagnetism for chromium.

energy is over-balanced by the Fermi energy and for this metal the unmagnetized arrangement has the least energy and is the stable state.

A very instructive curve can be drawn to illustrate the criterion for the occurrence of ferromagnetism. It is shown in Fig. 32. The vertical scale is the energy of the unmagnetized state, E_U, minus the energy of the magnetized state, E_M. When $E_U - E_M$ is positive, the magnetized state has the lower energy and will be the stable state,

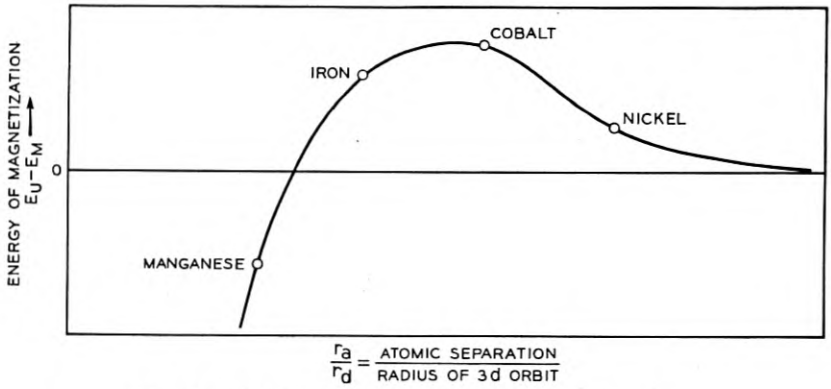

Fig. 32—Criterion for the occurrence of ferromagnetism.

and when $E_U - E_M$ is negative, the reverse is true. Hence a positive value for $E_U - E_M$ is a necessary and sufficient condition for ferromagnetism. The variable on the horizontal scale is r_a (the distance between nearest neighboring atoms in the crystal) divided by r_d (the average radius for the $3d$ wave function). Small values of r_a/r_d mean crowding together of the atoms, large values of the Fermi energy, and no ferromagnetism. Certain values of r_a/r_d, such as are found for iron, cobalt, and nickel, favor ferromagnetism. Very large values of r_a/r_d mean widely separated atoms and low Fermi energy and, consequently, ferromagnetism; however, for very widely separated atoms, the energy of interaction between them is small and so is the energy of magnetization. The curve shown in Fig. 32 is only qualitative. The theory that the curve should have this form was first worked out by Bethe using the "atomic" rather than the band theory of magnetism; for the reasons discussed above, however, no quantitative theoretical curve is available. Ratios of r_a/r_d have been calculated by Slater * and occur as indicated for several elements. This curve can be considered from either of two viewpoints. We may imagine that r_a remains constant, as it does approximately for the transition elements, and that r_d varies from element to element; we then get the result shown in Fig. 32. On the other hand we may consider a definite chemical element thus fixing r_d; then Fig. 32 tells us how the energy of magnetization depends on the lattice constant or volume of the sample. We shall use this in the following paragraphs to explain the effects of magnetism upon thermal expansion.

Magnetism and Thermal Expansion

In Fig. 33*a* we show a solid curve which represents for iron in the magnetized state the dependence of the energy E_M upon the lattice constant a. In Fig. 33*b* is shown, on a relatively enlarged energy scale, the value of $E_U - E_M$ as taken from Fig. 32 with r_d thought of as fixed, and a the lattice constant in place of r_a. The position of curve (*b*) has been adjusted so that the point marked O, corresponding to iron in Fig. 32, comes at the equilibrium distance or minimum of the E_M curve. Adding the solid curves of (*a*) and (*b*) (adjusting the energy scales, of course) gives the dashed curve representing the energy E_U of the unmagnetized state shown in Fig. 33*a*. We are now in a position to make predictions about the thermal expansion of iron.

Let us imagine that the iron is somehow made to stay in the magnetized state. Then its expansion curve, lattice constant versus temperature, will be shown as in Fig. 33*c* by the solid heavy line. Next

* J. C. Slater, *Phys. Rev.*, *36*, 57 (1930).

imagine it maintained in the unmagnetized state; in this state the equilibrium lattice constant is smaller than for the magnetized case and the expansion curve is shown dashed. The curves for fixed

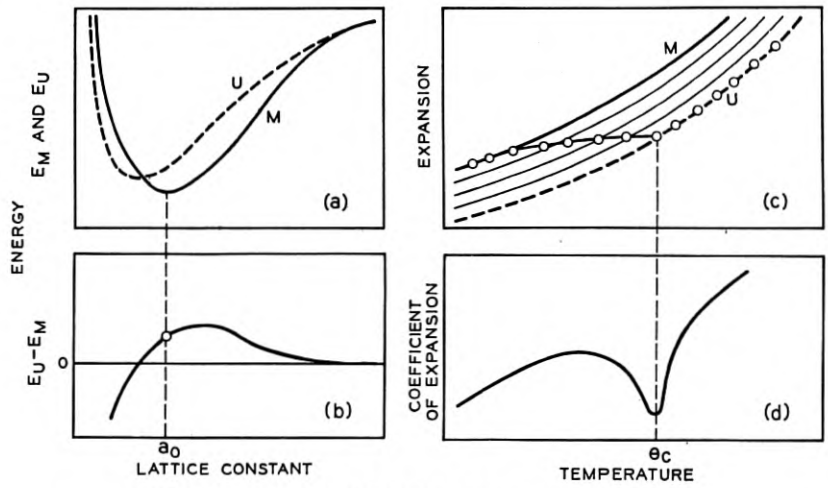

Fig. 33—Theory of the thermal expansion of iron.

(a) Energy in magnetized (M) and unmagnetized (U) states versus lattice constant.
(b) Difference in energies versus lattice constant.
(c) Lattice constant versus temperature.
(d) Thermal expansion coefficient versus temperature.

intermediate degrees of magnetization are shown as light lines. Now as the iron is heated the magnetization does not stay constant but decreases with temperature and becomes zero at the Curie temperature θ_c. In Fig. 33*c* this corresponds to a continuous shifting from the line of higher magnetization to the lines of lesser magnetization with increasing temperature as indicated by the curve with circles. We see that the rate of expansion—that is, the thermal expansion coefficient, which is defined as the derivative of the curve divided by a—should have an irregular form as shown in Fig. 33*d*.

In Fig. 34 we show observed thermal expansion curves for a series of iron nickel alloys,[41] showing that the expansion for iron rich alloys agrees with that predicted from Fig. 33. The reader may verify that had the curve of Fig. 33*b* been adjusted to correspond to nickel, the anomalous expansion would have been in the opposite direction, as is found experimentally for the nickel rich alloys.

The more rapid the transition from the magnetized to the unmag-

[41] Figures 34 and 35 are taken in a modified form from J. S. Marsh, "Alloys of Iron and Nickel," Vol. I, Special-Purpose Alloys, 1938, McGraw-Hill Book Co.

netized curve, the greater will be the anomaly in expansion. For the alloy invar the Curie point occurs at about 200° C. and the transition is so rapid that the magnetic effect nearly cancels the normal expansion.

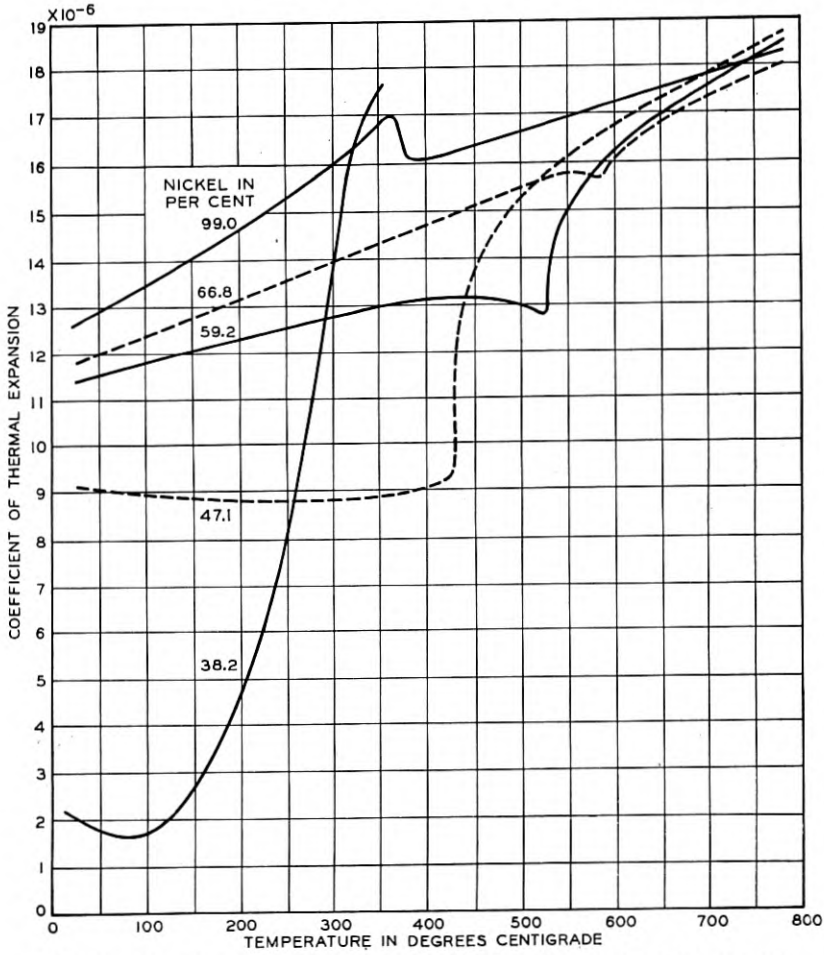

Fig. 34—Coefficients of expansion for iron-nickel alloys versus temperature.

Figure 35 shows a curve for the thermal expansion of an iron-nickel alloy containing 36.5 per cent Ni, corresponding to Fig. 33*c*. The flat region implies an expansion coefficient of nearly zero.

Grüneisen's law is definitely violated by metals having expansion effects of the sort associated with ferromagnetic changes. Grüneisen's law, it will be recalled, states that the thermal expansion coefficient is proportional to the specific heat. For all ferromagnetic transforma-

tions, the specific heat has a peak at the Curie temperature. For invar, however, the thermal expansion suffers a dip at the Curie temperature. Hence the proportionality between specific heat and thermal expansion coefficient does not hold. Even for cases where the expansion coeffi-

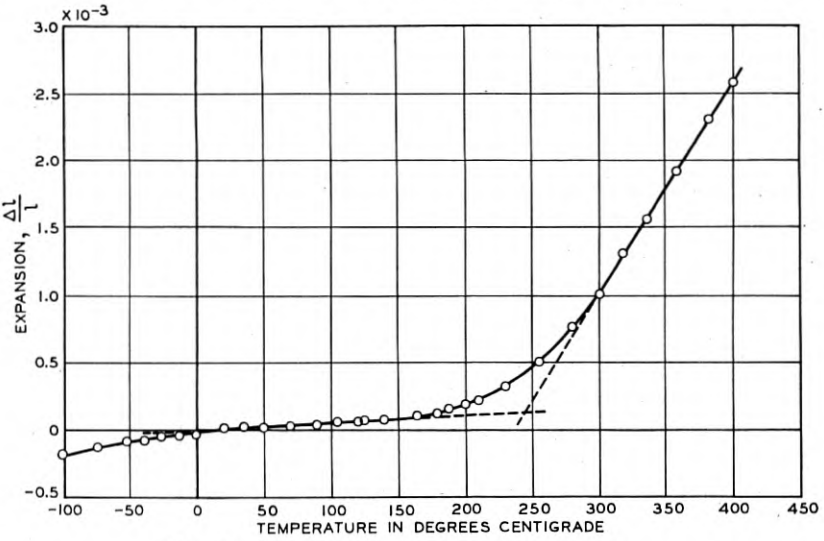

Fig. 35—Expansion of invar versus temperature.

cient has a peak, as in nickel for example, the proportionality does not hold. The reason for the failure of Grüneisen's law is easily found and reflects in no way upon validity of the law for the cases to which it is intended to apply. Grüneisen's law is derived by assuming that the crystal has a single definite energy versus volume curve. For ferromagnetic materials this is not true as is evinced by the two curves of Fig. 33*a*.

In this paper we have been concerned with the important but inactive attributes of electrons associated with their energies. We have seen how the variations of the electronic energy levels can be used to explain a number of the important properties of solids. In the next paper, we shall discuss the more dynamic subjects of electron velocities and accelerations.

ACKNOWLEDGMENTS

The writer would like to express his gratitude to Dr. R. M. Bozorth for discussions of the section on magnetism, to Dr. K. K. Darrow for criticisms and suggestions, to Mr. A. N. Holden for many valuable comments on the manuscript and for the preparation of the crystal of Fig. 1, and to Mr. B. A. Clarke for much valuable advice and assistance with the figures.

Dial Clutch of the Spring Type*

By C. F. WIEBUSCH

The mathematical theory is developed for the spring clutch which consists of two coaxial cylinders placed end to end and coupled torsionally by a coil spring fitted over them. Relations are derived whereby it is possible to design spring clutches in terms of the requirements and the constants of the spring material. Experimental verification of the relations is given. The theory of residual and active stresses as applied to the springs is discussed.

THE operation of all present day machine switching telephone systems depends on the use of the telephone dial. The dial originates the current pulses required to operate the step-by-step, panel, or crossbar switching equipment and for the reliable functioning of this equipment the pulses must occur within a closely limited frequency range. The stepping pulses are produced during the unwinding of the dial from the position to which it has been wound by the subscriber and it is this unwinding which must occur at a constant speed. To accomplish the speed control a governor depending on centrifugal force is used. It is not desirable that the governor come into action on the windup of the dial as this would put an extra load on the user's finger and slow up the operation of dialing. A clutch which holds in one direction of rotation and is free in the other direction is therefore interposed between the governor and the finger wheel with its associated circuit interrupting mechanism. In the past, the most commonly used clutch consisted of a pawl and ratchet, but this has now been replaced by the spring clutch because of its quietness and lower cost. A partially assembled dial using a spring clutch is shown in Fig. 1.

The ideal clutch for a dial governor would be one offering zero coupling torque during dial windup and an infinite positive coupling in the other direction. In practice the free torque in the windup direction need only be small compared to the torque of the main spring, while the holding torque in the other direction need only be great enough to withstand the main spring torque plus any helping torque that a

* Essentially the same material was presented at National Meeting of Applied Mechanics Division of The American Society of Mechanical Engineers, New York, N. Y., June 14–15, 1939, and published in *Journal of Applied Mechanics*, September 1939, under the title of "The Spring Clutch."

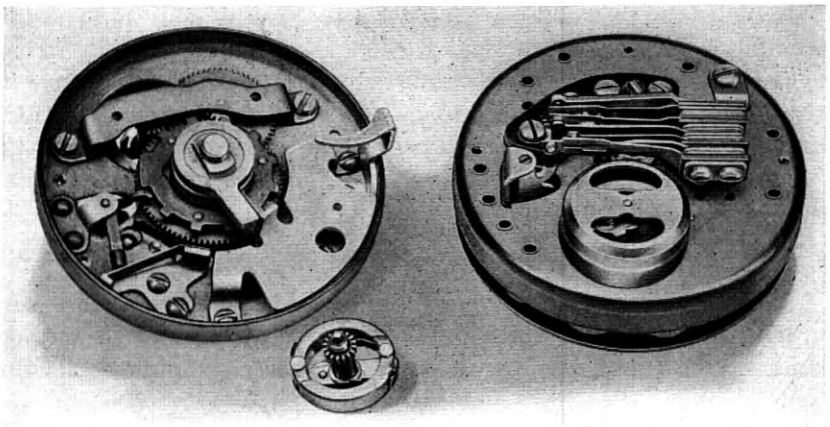

Fig. 1—Dial governor and front and back view of partially assembled dial.

subscriber may impatiently exert. The first limit is easy to set up on the basis of the mechanical constants of the dial; the fixing of the latter limit required measurements on the strength of a considerable number of persons. It was found that the maximum force that an ordinary man can exert with any finger at the finger hole of a dial is about six pounds. When the proper factors of safety have been added to these limits it is possible to specify exactly the requirements on the dial clutch. The remainder of this paper is devoted to the development of the relations and a discussion of the problems involved in designing a spring clutch to meet a given set of requirements.

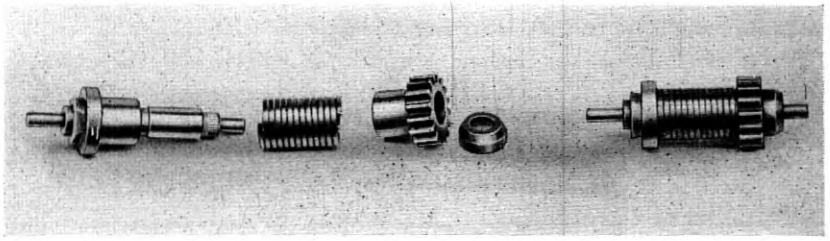

Fig. 2—Telephone-dial clutch.

The type of spring clutch to be discussed here consists of two cylinders placed end to end, rotating on a common axis, and torsionally coupled by the friction between the cylinders and a coil spring fitted over the cylinders. A photograph of such a clutch, assembled and apart, from a telephone dial is shown in Fig. 2. In spite of widespread use there seems to be little theoretical discussion of this device in the literature.

It is obvious that if the driving drum be rotated in the direction to wind up the spring and decrease the diameter, the spring will grip the cylinders and will be capable of exerting more torque than it would in the direction of rotation which tends to unwind the spring. Equations are to be developed which will permit the calculation of these two torque values in terms of the physical dimensions and the material constants of the clutch.

Torque of Spring Clutch in the Free Direction

In Fig. 3 assume that the spring is fastened to the left-hand arbor in order that any slipping which may take place must occur on the driving drum on the right. Assume also that the spring is so formed that the

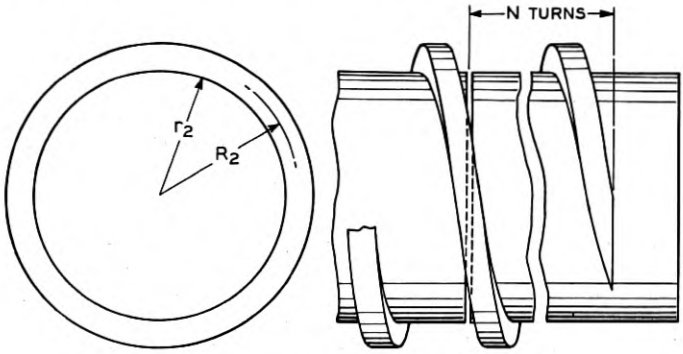

Fig. 3—Diagram of spring clutch.

inward radial force on the drum per unit length of the material is constant when no torque is applied.

Notation

l = length along the line of contact of the spring on the arbor, measured from the free end to any point, in.

μ = coefficient of friction between the spring and the arbor

r_2 = radius of the arbor, in.

R_2 = radius to the neutral bending axis of the spring when on the arbor, in.

N = number of turns on the right-hand arbor

P = compression in the spring wire at any point due to the applied torque; this is not the stress in the material but the resultant force acting across the entire cross section of the wire, lb.

f_0 = radial force of spring on arbor when no torque is applied, lb. per in. of contact line.

As compression exists in the spring wire at any point when the arbor is turned to make the spring unwind, there will be a radial force subtracting from f_0 at every point. This subtracting force is P/r_2. The increase of compression in the wire along the length of the line of contact due to friction is

$$dP = \mu(f_0 - P/r_2)dl, \tag{1}$$

which upon integration gives

$$l = -(r_2/\mu) \ln [f_0 - (P/r_2)]C,$$

where C is a constant of integration equal to $1/f_0$ since $P = 0$ at $l = 0$. Hence

$$P = r_2 f_0(1 - e^{-\mu l/r_2}). \tag{2}$$

Since $l = 2\pi r_2 N$,

$$P = r_2 f_0(1 - e^{-2\pi N\mu}). \tag{3}$$

Since the torque is equal to Pr_2

$$T = r_2^2 f_0(1 - e^{-2\pi N\mu}) \text{ (in.-lb.)}. \tag{4}$$

It will be observed that for any but fractional values of $N\mu$ the exponential term becomes very small and

$$T = r_2^2 f_0 \qquad (N\mu > 1). \tag{5}$$

If $N\mu = 1.0$ this expression is in error by only 0.2 per cent. It can thus be seen that provided $N\mu$ does not become too small, variations in N or μ do not affect the torque exerted. The torque will depend only on the radius of the arbor and on the force f_0 which is controlled entirely by the dimensions and the elastic properties of the spring.

Torque of Spring Clutch in the Gripping Direction

If the torque is applied to the clutch in the direction to wind up the spring, instead of unwind it as in the previous case, the force P'/r_2 due to the tension P' in the spring wire adds to the inward force f_0 and the relation corresponding to equation (1) is

$$dP' = \mu(f_0 + P'/r_2)dl. \tag{6}$$

From which by the same method as before

$$P' = r_2 f_0(e^{2\pi N\mu} - 1). \tag{7}$$

The corresponding torque is

$$T' = r_2^2 f_0(e^{2\pi N\mu} - 1) \text{ (in.-lb.)}. \tag{8}$$

In this case the torque increases rapidly with an increase in either N or μ especially for large values of $N\mu$. The coefficient of friction is in general a rather variable factor. It is to be expected that the slipping torque will also be variable, but since a lower limit can in general be set for this coefficient it will always be possible to make the number of turns of the spring such as to give any desired lower limit of torque.

The Radial Force on the Arbor

One method of evaluating the force f_0 occurring in the torque relations depends on equating the potential energy of strain per unit length of the wire when on the arbor to the work done in expanding the spring from its free diameter to the diameter of the arbor.

Let

E = Young's modulus for the spring material, psi
I = the area moment of the wire section, in.^{4}
h = the radial thickness of the wire, in.
R_1 = free radius to the neutral axis, in.
r_1 = free inner radius of the spring, in.

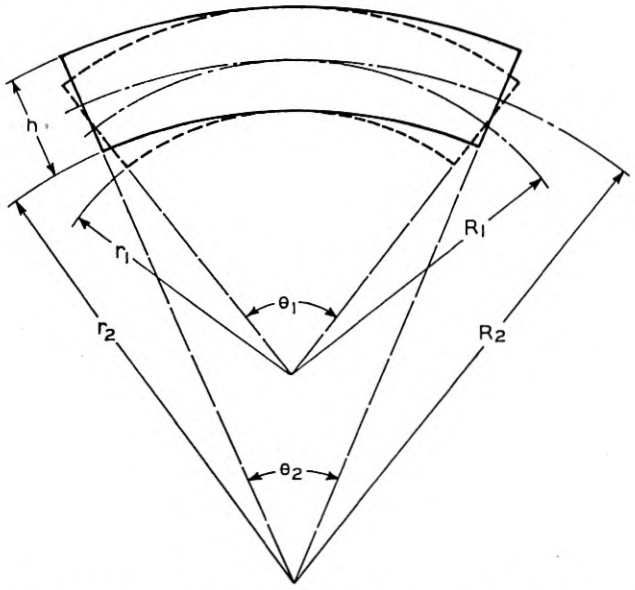

Fig. 4—Element of spring in initial and expanded condition.

Consider the portion of spring wire shown in Fig. 4 straightened out from the initial radius of curvature R_1 to the radius R_2. The

fibers above the neutral axis will be compressed while those below the neutral axis will be stretched. Let y be the distance of any given fiber from the neutral axis. The length of the undistorted fiber will be $L = (R_1 + y)\theta_1$ while after bending this same fiber will have the length $L' = (R_2 + y)\theta_2$. The strain or the change in length per unit length is

$$\frac{L - L'}{L} = 1 - \frac{(R_2 + y)\theta_2}{(R_1 + y)\theta_1}. \tag{9}$$

Since along the neutral axis there is no change in length, $\theta_2 = L_0/R_2$ and $\theta_1 = L_0/R_1$. Substituting these values in equation (9) gives

$$\text{Strain} = (y/R_2)(R_2 - R_1)/(R_1 + y). \tag{10}$$

The potential energy per unit volume in a material strained in tension or compression is

$$W/V = (E/2)(\text{Strain})^2. \tag{11}$$

Substituting the value of strain from equation (10) in equation (11) the energy density at any point of the deflected wire will be

$$\frac{W}{V} = \frac{E}{2}\left[\frac{y(R_2 - R_1)}{R_2(R_1 + y)}\right]^2. \tag{12}$$

Let b represent the width of the wire at the point y. Then for a wire symmetrical about the neutral axis the strain energy per unit length of the wire becomes

$$\frac{W}{l} = \int_{-h/2}^{h/2} \frac{E}{2}\left[\frac{y(R_2 - R_1)}{R_2(R_1 + y)}\right]^2 \frac{R_1 + y}{R_1} b\,dy, \tag{13}$$

where the factor $(R_1 + y)/R_1$ represents the ratio of the length of the fiber at the point y to the length along the neutral axis. If all values of y, and hence also $h/2$, are small compared to R_1 there will be little error made in neglecting the y, which carries both positive and negative values, in the expression $R_1 + y$. Hence

$$\frac{W}{l} = \int_{-h/2}^{h/2} \frac{E}{2}\left(\frac{R_2 - R_1}{R_2R_1}\right)^2 by^2dy. \tag{14}$$

The integral of by^2dy is the area moment I of the section and therefore

$$\frac{W}{l} = \frac{1}{2}EI\left(\frac{R_2 - R_1}{R_2R_1}\right)^2. \tag{15}$$

This must be equal to the work done per unit length of the neutral

axis, by the force per unit length $F(\Delta R)$ working through the distance ΔR where $\Delta R = R_2 - R_1$. That is

$$\int_0^{\Delta R} F(\Delta R)d\Delta R = \frac{1}{2} EI \left[\frac{\Delta R}{R_1(R_1 + \Delta R)} \right]^2. \tag{16}$$

Differentiating both sides with respect to ΔR gives $F(\Delta R)$ for the left-hand side, and after simplification

$$F(\Delta R) = \frac{EI}{R_1} \frac{\Delta R}{(R_1 + \Delta R)^3}. \tag{17}$$

Substituting for ΔR its value $R_2 - R_1$ gives

$$F(\Delta R) = EI(R_2 - R_1)/R_1R_2^3. \tag{18}$$

The equivalent force per unit length measured along the surface of the arbor must be larger than this value in the ratio of R_2/r_2 since the same total force is here distributed over a shorter length. This latter force is f_0; hence

$$f_0 = \frac{R_2}{r_2} F(\Delta R) = EI \frac{R_2 - R_1}{R_1 r_2 R_2^2} \text{ lb. per in.} \tag{19}$$

The value of f_0 calculated from this equation in terms of the constants of the spring material and the dimensions of the spring may be used in equations (4) and (8) to calculate the free torque and the slipping torque of the clutch.

Experimental Check of the Free-Torque Relation

In order to check the validity of the relation for the free torque, equation (4), and that for the radial force on the arbor, equation (19), the free torque of a given spring on arbors of various diameters as well as the torque for different numbers of turns on the same arbor was measured. The spring of 0.0085 × 0.022-in. phosphor-bronze ribbon was attached to a short vertical shaft suspended by a torsion fiber of measured torsional stiffness. The free end of the spring was placed over a vertical arbor capable of rotation. The arbor was rotated and the angle of twist of the torsion fiber was measured thus giving a measure of the slipping torque. The precision of the measurements of torque was about 0.5 per cent although the sensitivity to small changes was about 0.2 per cent. No measurable increase of torque occurred by increasing the number of turns on the arbor beyond six. This is to be expected if the coefficient of friction exceeds about 0.12.

For all succeeding measurements seven to eight turns were used. As a further check on the independence of the torque and the coefficient

of friction for this number of turns a measurement was made before and after oiling the arbor and spring with a light lubricating oil. There was a decrease in torque of approximately 0.25 per cent.

The inside diameter of the spring as measured by a taper gage was 0.180 in. ± 0.001 in. The torque for arbors ranging in size from 0.182 to 0.193 in. was determined and is shown in Fig. 5. This curve ex-

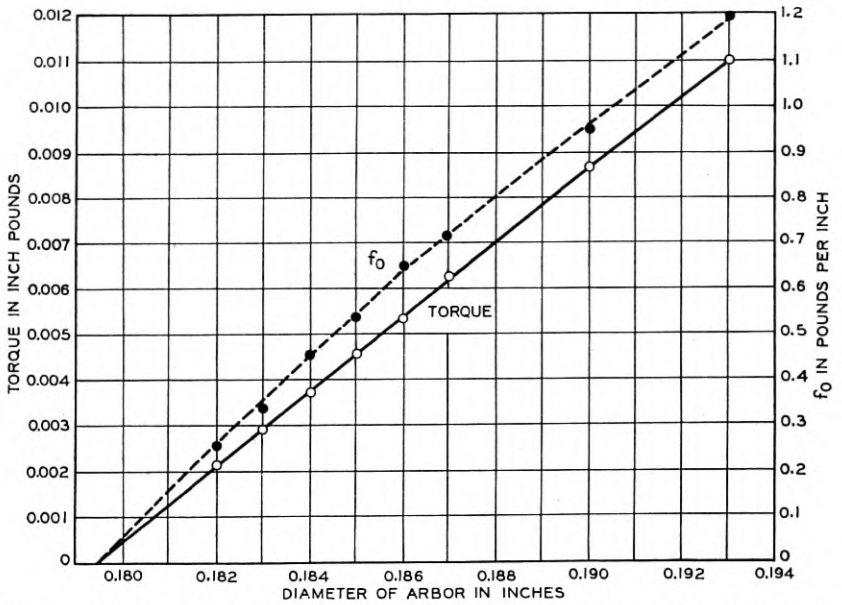

Fig. 5—The free torque and the radial force on the arbor for a phosphor-bronze spring on arbors of different diameters.

trapolated to zero torque gives, for an accurate measure of the inside diameter of the spring, 0.1794 in. Using this value and the quantity EI, determined by obtaining the resonant frequency of a straight short length of the ribbon, of which the spring was made, vibrating as a fixed free reed, the radial force on the arbor as calculated by equation (19) is shown by the dotted curve as a function of the arbor diameter $2r_2$. The points indicate values of f_0 obtained from the measured values of torque by the use of equation (5). The two sets of values agree within about 2 per cent.

As a further check on the validity of the calculations under practical conditions, the free torque of a phosphor-bronze spring on a dial-governor arbor was measured for various numbers of turns of the spring engaging on the slipping arbor. The torque due to bearing friction alone with no spring in place was also measured and found to

be approximately 0.001 in.-lb. This constant value was subtracted from the other measured values of torque. The resulting values of free spring torque are plotted as a function of the number of turns in Fig. 6. The torque values calculated by equations (4) and (19) and

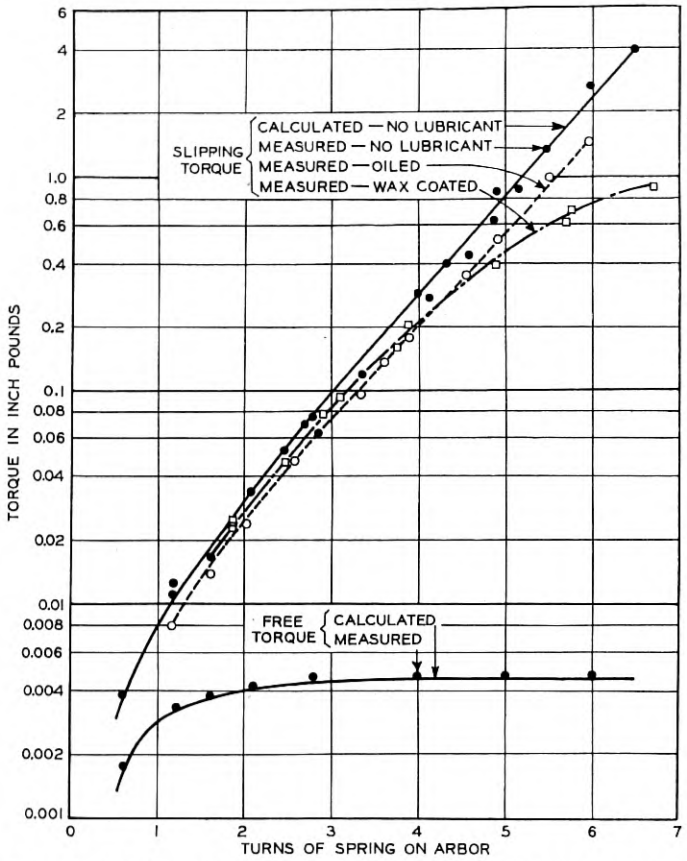

Fig. 6—Dependence of torque on the number of turns of the spring engaging the arbor.

using the value of $\mu = 0.165$ obtained as described in the next section, are shown on the curve to indicate the agreement.

Experimental Check of the Holding-Torque Relation

It is to be expected that any relation for which the coefficient of friction is a controlling factor will be difficult to check accurately. Equation (8) for the slipping torque is of this type. It is possible however by taking a large number of measurements to establish the

validity of the equation and then by determining limiting values for the coefficient of friction, to use this equation as a design relation, especially if only a minimum torque limit is set. Measurements of the slipping torque, as a function of the number of turns, were made on the same dial clutch that was used for checking the free torque.

This slipping torque was not steady as was the case in the free direction but varied as much as $\pm$ 20 per cent. An average value was taken in each case. An uncertainty also existed regarding the number of turns engaging the rotating arbor. Since the crossover from one arbor to the other requires practically one whole turn, slight differences in the arbor diameters may result in a gain or loss of almost half a turn. In addition to these factors there was an end effect due to the fact that the free end of the spring wire was cut off square rather than beveled but this factor although calculable was neglected in view of the other uncertainties. The results of these measurements are shown in Fig. 6.

From equation (8) it can be seen that for large values of N the plot of T' versus N will be a straight line provided μ is independent of the force between the spring and the arbor. The slope of this straight line when multiplied by the proper constant, which can be shown to be 0.368, gives the coefficient of friction. For the experimental points shown in Fig. 6 this value of μ is 0.165. Using this value of μ in equation (6) the calculated curve was plotted. Considering the uncertainties involved the calculated and measured curves are in good agreement.

The dotted curve of Fig. 6 shows the effect of lubricating the clutch with a light machine oil. This resulted in only a small decrease of the coefficient of friction. The curve shown by the dashes illustrates the effect of lubricating a clutch with spermaceti. The coefficient was no longer a constant but decreased with an increase in load.

Spring Stresses

In determining the load that a spring clutch will withstand, first without stretching which will result in backlash, and second without breaking, initial as well as load stresses must be considered. The initial stresses are made up of the residual stresses due to forming the spring, plus the stresses due to expanding the spring to fit the arbor, that is from an inner radius r_1 to an inner radius r_2.

Of these limiting load values the easiest to calculate is the torque required to break the spring. This is given by the product of the radius of the spring and the breaking strength of the spring wire. Loads much smaller than this value would stretch some of the fibers of the spring, especially those in which a high initial stress already

existed. As will be shown, such stretching will cause the radius to increase at those portions of the spring where the applied stresses are highest, that is, for those turns near the dividing line of the arbor.

Substituting for y, in equation (10), the distances from the neutral axis to the extreme inner and outer fibers will give the strain in these fibers. Provided R_1 is considerably larger than $h/2$, half the thickness of the material, the distance to these extreme fibers becomes $h/2$ and the y in the denominator can be neglected in comparison to R_1. The maximum fiber stress due to placing the spring on the arbor is this value of strain multiplied by Young's modulus. Then

$$S_0 = \frac{h}{2R_2} \frac{R_2 - R_1}{R_1} E. \tag{20}$$

This stress is in the form of compression for the outer fibers and tension for the inner. Since a load on the clutch results in a tension in the spring, the stress given by equation (20) must be added to the load tension stress to get the total stress on the inner fibers of the spring. This initial stress therefore reduces the load-carrying capacity of the clutch.

Residual Spring Stresses

When a straight wire is wound upon an arbor to form a coil spring the strain on the inner and outer fibers must exceed that corresponding to the yield-point and plastic-flow results. To simplify the discussion assume the idealized stress-strain curve shown by the heavy lines of Fig. 7(a). The stress distribution across any section of the wire while wound on the winding arbor will be as shown by the heavy lines of Fig. 7(b) where S_{YP} is the yield-point stress, the maximum stress that the material will sustain. The moment, across the section, required to produce this bending is

$$M = \int_{-h/2}^{h/2} Sbydy \tag{21}$$

where b is the width of the wire at any point of y distance from the neutral axis and S is the stress at the same point. If the spring is released, it expands to a radius R_1 in which condition the external and the internal moments are both zero. It is now possible by applying the same moment as was given by equation (21) to reduce the radius of curvature again to R_0 but without causing additional plastic flow. The added stress distribution produced by this second bending must therefore follow a straight line as shown by S_2-S_2, which together with the residual stresses in the relaxed condition (radius R_1) must

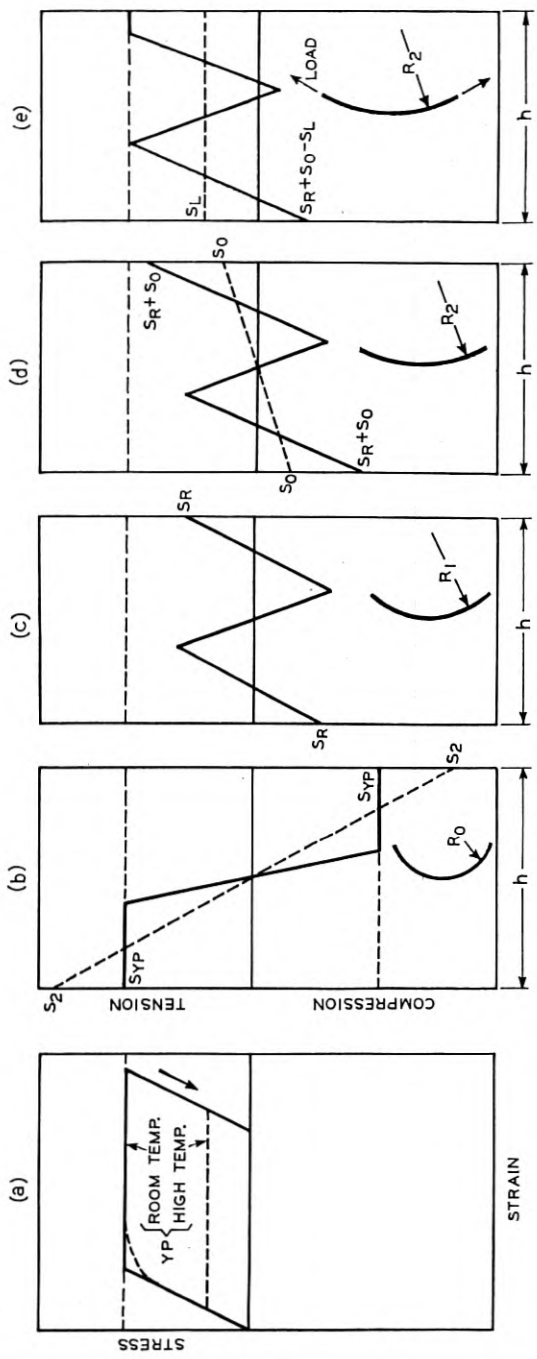

Fig. 7—Idealized stress conditions in clutch springs.

equal the distribution S_{YP}-S_{YP} resulting from the original forming operation. Therefore the stress distribution in the relaxed condition (radius R_1) must be the difference between S_{YP}-S_{YP} and S_2-S_2 or as indicated in Fig. 7(c). The value of S_2 is of course so determined that the moment as specified by equation (21) is the same for the dotted-line as for the solid-line stress distribution.

The value of $S_R = S_2 - S_{YP}$ is relatively easy to determine for rectangular and round wire on the basis of the straight-line stress-strain characteristic if the bending has been sufficiently severe to have caused plastic flow almost to the neutral axis. The moment given by the actual stress distribution will then differ but little from that obtained by equation (21) with S replaced by S_{YP}, a constant. The values of S corresponding to the S_2-S_2 distribution are given by

$$S = 2S_2y/h. \tag{22}$$

For a rectangular wire b is a constant and equating the moments corresponding to the two stress distributions gives

$$\int_{-h/2}^{h/2} S_{YP}bydy = \int_{-h/2}^{h/2} 2\frac{S_2}{h}by^2dy,$$

from which $S_2 = (3/2)S_{YP}$ or

$$S_R = S_2 - S_{YP} = (1/2)S_{YP} \text{ (rectangular wire)}. \tag{23}$$

Similar integrations in the case of a round wire of radius $h/2$ for which $b = 2\sqrt{[(h/2)^2 - y^2]}$ gives

$$S_R = 0.7S_{YP} \text{ (round wire)}. \tag{24}$$

These equations give the residual fiber stresses in the extreme inner and outer fibers under the assumed conditions. In any case where the stress-strain characteristic is known a correct value for the residual stress can be obtained by graphical integration. The values given by equations (23) and (24) will, however, be fair approximations even in cases where the stress-strain characteristic is curved provided it does not exhibit strain hardening to a decided extent. Since a limited amount of plastic flow takes place for any stress above the proportional limit, which is generally far below the yield point, the residual stress may be sufficient to cause a small amount of creep. Any additional stresses will then cause permanent deformation of the spring.

The analysis will be continued on the basis of the idealized straight-line characteristic. Figure 7(d) shows the result of placing the spring

on the clutch arbor. The line S_0-S_0 shows the stresses added by this expansion where S_0 is given by equation (20). The sum of these stresses and those shown in Fig. 7(c) gives the total stresses as shown by the solid line of Fig. 7(d). If now a load be put on the clutch a uniform stress S_L will be added but this stress for even relatively light loads may be sufficient to cause the total stress on the inner fibers to exceed the yield point as is indicated in Fig. 7(e). The inner fibers are consequently stretched and when the load is released and the spring taken from the arbor it will be found that the center turns of the spring have expanded. Even with the spring on the arbor if the clutch is turned in the free direction it will be noticed that these center turns raise off the arbor. It was shown in the paragraphs on the clutch torque in the free direction that the torque did not increase appreciably after the first few turns. This can be explained by the fact that as soon as the outward radial force due to the compression along the wire is equal to the initial inward radial force of the spring on the arbor the friction on these turns vanishes. The value of the compression will be fixed by a relatively few end turns. Hence if the inward force of some of the center turns decreases due to their stretching this compression will be sufficient to expand the turns to clear the arbor. If S_L is still further increased the stretch will be sufficient to cause the diameter of the center turns to exceed the arbor diameter even when no torque is applied in the free direction.

Since the yield point of metals decreases at higher temperature it is possible to produce a spring having lower residual stresses by the proper heat-treatment. If the wire is wound on an arbor and then heated, additional plastic flow takes place since the maximum stress that can be sustained at the high temperature is that shown in Fig. 7(a) as the high-temperature yield point. If the spring is then cooled and released the expansion will not be as great as for the untreated spring. The residual stresses will again be given by equation (23) or (24) where S_{YP} is taken as the lowest yield point reached in the temperature cycle. In Fig. 8, (a), (b), and (c) show the stresses in the heat-treated specimen corresponding to those shown in Fig. 7, (c), (d), and (e), for the untreated spring. Figure 8(c) shows that for the same load stress as for Fig. 7(e) no permanent deformation has taken place. It is of course important that the strain-relieving temperature should not go high enough to lower permanently the strength of the material. This limit [1] for phosphor bronze is about 320° C.

To determine the stress-temperature characteristics of 18-8 stainless

[1] "Better Instrument Springs," Robert W. Carson, *Trans. A.I.E.E.*, vol. 52, September, 1933, p. 869.

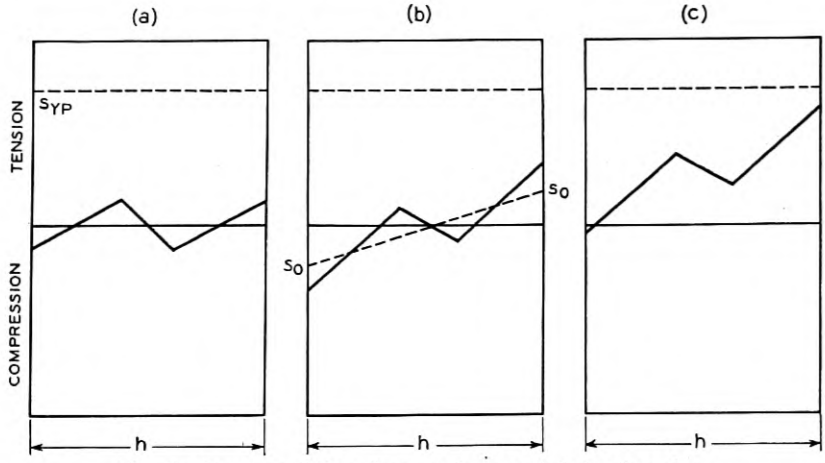

Fig. 8—Stress conditions in a strain-annealed clutch spring.

steel [2] a number of springs of 0.0068 × 0.021-in. ribbon were wound on 0.1486-in. arbors and given various heat-treatments. They were then cooled, released from the arbors, and measured for inside diameter. Table 1 gives the results of these measurements.

TABLE 1

EFFECT OF HEAT-TREATMENT ON SPRINGS

Ribbons of 18-8 stainless steel, 0.0068 × 0.021 in., heat-treated on 0.1486-in. winding mandrels

Heat-treatment temp., °C.	Time, hr.	Diam. after release, in.	Residual stress, psi
25	..	0.228	111000
100	4	0.206	88000
200	4	0.188	66000
300	4	0.182	58000
400	4	0.175	47000
470	4	0.169	37000
500	4	0.166	33000
400	¼	0.177	51000
400	½	0.176	49000
400	1	0.176	49000
400	2	0.175	47000
400	3	0.175	47000
400	4	0.175	47000

The residual stresses were calculated by noting from equation (23) that $S_R = S_2/3$ and then obtaining S_2 from equation (20) rewritten as

$$S_2 = \frac{h}{2R_1} \frac{R_1 - R_0}{R_0} E. \tag{25}$$

[2] 8 per cent nickel, 18 per cent chromium.

Straight pieces of the stainless-steel ribbon were given the same series of heat-treatments as the springs. Young's modulus was determined for each of these samples and a bending test was also applied to determine whether the wire had been permanently annealed. No appreciable effect was noted. The proportional limit and the ultimate tensile strength of this ribbon at room temperature as measured on a tensile-testing machine were 41,600 and 252,000 psi, respectively. It is thus seen that except with the high-temperature anneals the residual stress alone exceeds the proportional limit and any additional stress will cause a permanent deformation.

It is also possible to obtain a favorable residual-stress distribution, that is, an initial compression on the inner and a tension on the outer fibers. If the released spring having the stress distribution shown in Fig. 7(c) is expanded until considerable plastic flow takes place on the inner and outer fibers the stress distribution of Fig. 9(a) is obtained, which on release results in the residual-stress distribution as shown in Fig. 9(b).

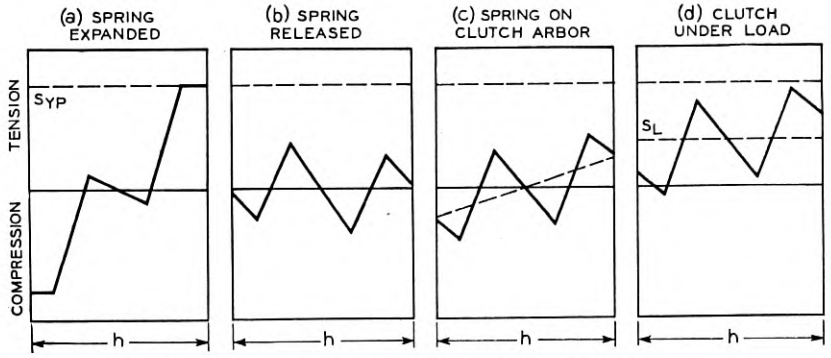

Fig. 9—Stress conditions in an expanded clutch spring.

To verify the validity of these arguments three springs of stainless-steel ribbon with different preliminary treatments and one of heat-treated phosphor bronze were tested for backlash as a function of previous loading. The backlash angle was measured from a point of slipping in the free direction to the point in the holding direction at which it would sustain a load of 0.05 in.-lb. An initial load of 0.5 in.-lb. was then applied and removed and the backlash measured as before. This was repeated for various loads up to the breaking point of the spring. The results are shown in Fig. 10. In the case of the untreated stainless steel the backlash began to increase immediately. Its higher initial value was probably due to the unavoidable stressing occasioned

by assembling the spring on the clutch arbor and to the 0.05-in.-lb. testing torque. The backlash of the other three samples remained constant to slightly above 0.5 in.-lb. and then began to rise. At low loads the phosphor bronze was better while at higher loads the heat-treated stainless steel had the advantage; the breaking load for the latter was also about 30 per cent higher.

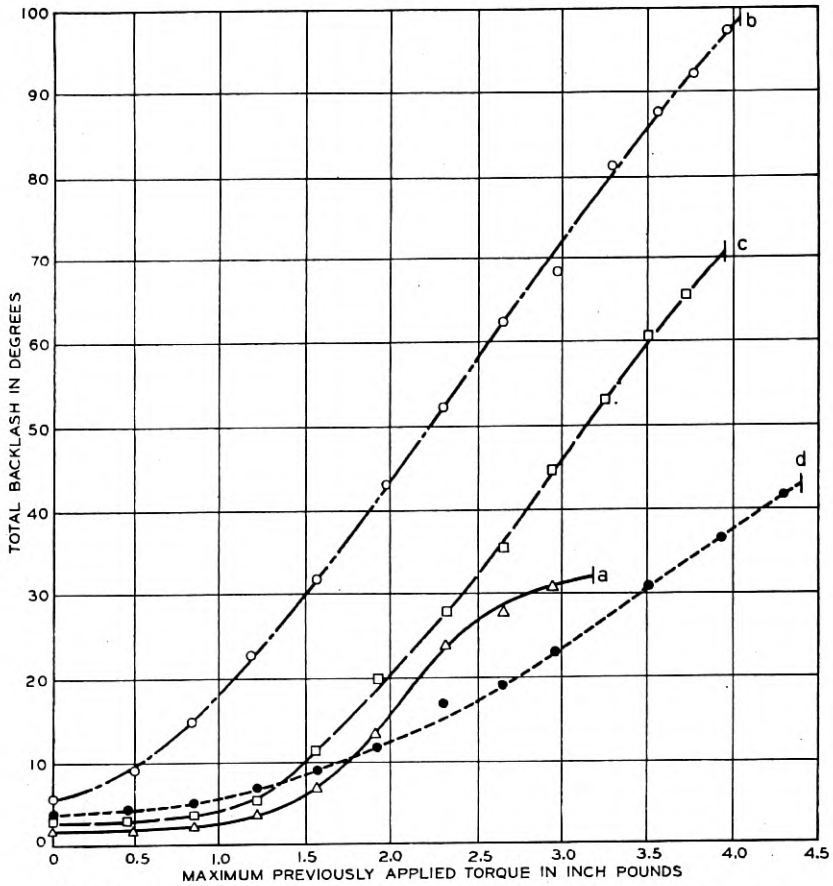

Fig. 10—Effect of overload on clutch backlash. Arbor diameter, 0.190 in.

(*a*) Phosphor-bronze-ribbon spring, 0.0085 × 0.022 in., heat-treated 4 hr. at 230° C., 0.1835 in. ID.

(*b*) 18-8 stainless-steel ribbon, 0.0068 × 0.021 in., wound on 0.127-in. mandrel and released, 0.183 in. ID.

(*c*) 18-8 stainless-steel ribbon, 0.0068 × 0.021 in., wound on 0.120-in. mandrel and released, 0.170 in. ID, mechanically expanded to 0.183 in. ID.

(*d*) 18-8 stainless-steel ribbon, 0.0068 × 0.021 in., wound on 0.157-in. mandrel, heated 4 hr. at 470° C., cooled, and released, 0.182 in. ID.

Conclusion

The relations developed in the preceding sections are sufficient to determine uniquely the correct spring dimensions for a spring clutch of specified free and gripping torque provided the material constants and the cross-sectional shape (round or rectangular) of the wire, as well as the length and diameter of the clutch arbor, are specified. Choice of values for the last two factors is based largely on permissible heating resulting from the slipping in the free direction. Furthermore, as would generally be the case, if only maximum values of the free torque and minimum values of the gripping torque are given a number of solutions can be obtained from which to choose the most convenient. By combining the relations derived, in a manner to permit step-by-step calculation, the design of spring clutches is reduced to a simple routine.

Abstracts of Technical Articles from Bell System Sources

Recent Observations on the Relation between Penetration, Infection and Decay in Creosoted Southern Pine Poles in Line.[1] C. H. AMADON. The relation between poor penetration and decay, and the necessity for rational and adequate penetration requirements in treating specifications, are now fairly well understood by producers and users of creosoted southern pine poles. The purpose of this brief paper is to supplement the information presented in the Proceedings for 1936 and 1937 on the behavior of these poles in line under actual service conditions.

Tarnish Studies. The Electrolytic Reduction Method for the Analysis of Films on Metal Surfaces.[2] W. E. CAMPBELL and U. B. THOMAS. A method is described for analysis of tarnish films on metals by electrolytic reduction at the cathode. Its suitability is demonstrated for the rapid and accurate measurement of oxide films on copper varying in average thickness from monomolecular layers to 1000 Å. It is shown to be useful for reduction of mixed oxide-sulfide films on copper and silver. The method is used to measure the oxide films on freshly reduced copper after one-half hour's exposure to oxygen or air. Such films are shown to be 10–20 Å thick. A thicker film, measuring 30–70 Å is found to be produced by abrasion of copper in air, water, benzene or toluene. Adaptations and modifications are discussed which give wide analytical application to the method.

An Electrochemical Study of the Corrosion of Painted Iron.[3] H. E. HARING and R. B. GIBNEY. The corrosion protective value of approximately 50 different paints was determined by means of an electrochemical method which has been previously described. This determination involved the measurement of the change in the potential of the painted iron with time when wet with water for 24 hr. or less. It was found that the interpretation of the time-potential curves which were automatically plotted by a recording vacuum tube electrometer, was facilitated if the test was conducted in a nitrogen atmosphere. The results obtained with the electrochemical or potentiometric method compared favorably with those obtained in a

[1] *Proc. American Wood-Preservers' Association*, 1939.
[2] *Electrochemical Society Preprint* 76–25.
[3] *Electrochemical Society Preprint* 76–24.

one-year outdoor exposure test. Such differences as were found were shown to be due either to deterioration or improvement in the paint film as the result of weathering.

Characteristics of Modern Microphones for Sound Recording.[4] F. L. HOPPER. Factors influencing the choice of a microphone for sound recording are considered. The characteristics of a new miniature condenser transmitter and amplifier, as well as a number of other types of microphones now in use, are included.

Cold-Cathode Gas-Filled Tubes as Circuit Elements.[5] S. B. INGRAM. The application of electronic devices to the local systems plant is still in its infancy. One of the first of these devices to receive extensive use is the cold-cathode gas-filled tube. As a sensitive relay it is beginning to make its appearance in a number of telephone control and signaling circuits, being best known for its use in the standard four-party subscriber set where its rectifying property enables it to discriminate between positive and negative polarity for selective ringing. Compared with other types of vacuum tubes the cold-cathode tube has the advantages that it operates without cathode heating power, has the ability to start immediately when a signal is applied, and does not deteriorate when not passing current. These advantages make it particularly suitable for use in telephone circuits where intermittent service is common and long life and economical operation are required.

The paper describes the structure and electrical characteristics of cold-cathode tubes. Their properties as circuit elements are then illustrated in a number of typical basic circuits.

Inductive Coordination with Series Sodium Highway Lighting Circuits.[6] H. E. KENT and P. W. BLYE. This paper describes the wave-shape characteristics of the sodium-vapor lamp and discusses the relative inductive influence of various series circuit arrangements in which such lamps are employed. A method is outlined by means of which the noise to be expected in an exposed telephone line may be estimated. Measures are described which may be applied in the telephone plant or in the lighting circuit to assist in the inductive coordination of the two systems. These measures need be considered only when a considerable number of lamps is involved, since noise induction is negligible

[4] *Jour. S.M.P.E.*, September 1939.
[5] *Elec. Engg.*, July 1939.
[6] *Elec. Engg.*, July 1939.

when there are only a few lamps as, for instance, at highway intersections.

A Cardioid Directional Microphone.[7] R. N. MARSHALL and W. R. HARRY. A microphone is described which has uniform directivity over a wide frequency range. This is made possible by placing in a single instrument a dynamic type pressure microphone element and a ribbon type "velocity" element, and electrically equalizing the outputs before combination. The resultant directional pattern is a heart-shaped curve or cardioid, giving a fairly wide pick-up zone in front and a substantial dead zone at the back of the instrument. Because of the unusually rugged ribbon employed, the new microphone is much less susceptible to wind noise than ordinary ribbon types. Housed in an aluminum case, the microphone weighs only $3\frac{1}{4}$ lbs. High output level, low impedance, and high quality, together with the excellent directivity, promise to make the cardioid microphone an important tool for the motion picture sound engineer.

Fractional-Frequency Generators Utilizing Regenerative Modulation.[8] R. L. MILLER. By the application of the principle of regeneration to certain modulation systems, a generator of submultiple or other fractional-frequency ratio may be obtained.

A simple example is obtained by considering a second-order modulator whose output is connected back to a conjugate input by means of a feedback loop including an amplifier and a selective network. If an input frequency f_0 is applied, it is found that a frequency component $f_0/2$ appearing in the feedback path will modulate with the applied frequency to produce sidebands of $f_0/2$ and $3f_0/2$. The network and amplifier, being especially efficient for the frequency $f_0/2$ and having a gain higher than the modulator loss, will reinforce this component causing it to build up to some steady-state value. Similar processes are possible by which greater submultiple ratios may be obtained.

Since the output wave is obtained by a modulation process involving the input wave, it will appear only when an input is applied and then bears a fixed frequency ratio with respect to it. Experiments show that the ability of the generator to produce a fractional frequency is independent of phase shift in the feedback path. Circuits are possible in which the amplitude of the fractional-frequency wave will bear a linear relation to the input wave over a reasonable range and at the

[7] *Jour. S.M.P.E.*, September 1939.
[8] *Proc. I.R.E.*, July 1939.

same time maintain a constant phase angle between the two waves. Typical circuits are discussed which make use of copper oxide as the modulator elements.

Seasonal Cosmic-Ray Effects at Sea Level.[9] R. A. MILLIKAN, H. V. NEHER and D. O. SMITH. By sending a Neher self-recording electroscope in a 10-cm lead shield repeatedly on a slow Norwegian steamer over the route Vancouver-Los Angeles, around South America and return to Los Angeles and Vancouver, we find (1) as heretofore an equatorial dip measured from Los Angeles of seven per cent on the western side of South America, eight per cent on the eastern side; (2) no measurable seasonal effect, or winter-summer differences, at all in the voyage from Los Angeles to the Straits of Magellan; (3) as heretofore constancy in cosmic-ray intensity in summer and fall, within the limits of uncertainty imposed by fluctuations estimated at not over one per cent, on the voyage between Los Angeles and Vancouver; (4) but in winter and spring an increase of as much as two or three per cent between Los Angeles and Vancouver. This is interpreted as the atmospheric-temperature effect earlier studied by Hess, Compton, and their respective collaborators.

Some Engineering Considerations in Loading Circuits.[10] J. A. PARROTT. This paper describes the various loading arrangements used on toll entrance and intermediate cable circuits and discusses the transmission benefits obtained by loading and some of the important problems in the consideration of loading railroad entrance and intermediate cables. In addition to voice frequency loading, loading for the lower frequency carrier systems such as the Type H is also discussed.

The Formation of Metallic Bridges between Separated Contacts.[11] G. L. PEARSON. Low resistance bridges were formed between gold, steel and carbon electrodes having separations of $2\text{–}70 \times 10^{-6}$ cm by applying voltages less than the minimum sparking potential. For a given pair of electrodes the field required to form the bridges is a constant and is $5\text{–}16 \times 10^{6}$ volts per centimeter. Measurements of the temperature coefficient of resistance of the bridges identify them as consisting of the material of the electrodes. A study of their resistance as a function of the displacement of one of the electrodes shows that they may be pulled out as well as crushed. At voltages

[9] *Phys. Rev.*, September 15, 1939.
[10] *Proc., Assoc. Amer. R.R., Telegraph and Telephone Section*, April 1939.
[11] *Phys. Rev.*, September 1, 1939.

less than those required to form the bridges, field currents exist. These increase rapidly as the field is raised and attain a value around 10^{-10} ampere before the bridges are formed. Calculation of the maximum electrostatic stress on the electrodes at the time of breakdown gives a value 0.05 to 0.0005 times the tensile strength of the electrode material at room temperature. The field is locally higher than that calculated because of surface roughness and the tensile strength is probably lowered by the local heating known to accompany field currents. The data therefore indicate that electrostatic force pulls material from the electrodes to bridge the gap.

Measuring Transmission Speed of the Coaxial Cable.[12] J. F. WENTZ. Time of transmission of carrier currents over high speed lines is discussed. A method of measuring this time delay as used on the 1000-kc system of the New York-Philadelphia coaxial cable is described and the results are given for the television band transmitted over it experimentally.

[12] *Bell Labs. Record*, June 1939.

Contributors to this Issue

M. J. AYKROYD, B.S. in Civil Engineering, Queen's University, 1913. Bitulithic and Contracting Company, Edmonton, 1913–14; Department of Public Works, Ottawa, 1915–16; Imperial Ministry of Munitions, Montreal and New York, 1916–18; Manager and Director, Export and Import Company, Montreal and London, England, 1919–23. Bell Telephone Company of Canada, Plant Department, Montreal and Toronto, 1923–34; Outside Plant Engineer, Bell Telephone Company of Canada, Toronto, 1934–.

D. G. GEIGER, Queen's University: B.S. in Electrical Engineering, 1922; B.S. in Mechanical Engineering, 1923; Department of Electrical Engineering, 1923–24. Bell Telephone Company of Canada, Montreal, Transmission Division of General Engineering Department, 1924–26 and 1928–29. Lecturer in Electrical Engineering, Queen's University, 1926–28. Transmission Engineer, Bell Telephone Company of Canada, Toronto, 1930–.

EGINHARD DIETZE, B.S. in Electrical Engineering, University of Michigan, 1917. American Telephone and Telegraph Company, 1917–34; Bell Telephone Laboratories, 1934–. Mr. Dietze has been engaged in transmission studies of telephone stations, including room noise conditions at telephone locations. He is co-author of "Indicating Meter for Measurement and Analysis of Noise."

WALTER D. GOODALE, JR., E.E., Lehigh University, 1928; M.E.E., Polytechnic Institute of Brooklyn, 1937. American Telephone and Telegraph Company, Department of Development and Research, 1928–34; Bell Telephone Laboratories, 1934–. Mr. Goodale has been engaged in studies of various factors affecting telephone station transmission.

W. B. BEDELL, Ohio Northern University, U. S. Army, 1917–1919 (Lieutenant, Infantry); American Telephone and Telegraph Company, Long Lines Engineering Department, 1919–. Mr. Bedell is engaged in equipment engineering work involving broad-band carrier telephone systems.

GLEN B. RANSOM, B.S. in Electrical Engineering, University of Minnesota, 1921. American Telephone and Telegraph Company, Long Lines Department, at Chicago, 1922–27; District Engineer at Indianapolis, 1927–28; Division Transmission Engineer at Cleveland, 1928–30; Long Lines Engineering Department, Transmission Branch, 1930–. Appointed to present position, Circuit Layout Engineer, 1932.

W. A. STEVENS, B.S. in Electrical Engineering, Bucknell University, 1925. New York Telephone Company, Engineering Department, 1925–28. American Telephone and Telegraph Company, Department of Operation and Engineering, Transmission Engineering Group, 1928–. Mr. Stevens' work has largely dealt with toll transmission matters.

B. D. HOLBROOK, A.B., Stanford University, 1924; A.M., 1925. University of Chicago, 1926–30. Bell Telephone Laboratories, 1930–. Mr. Holbrook has engaged in research on broad-band carrier telephony and on signaling systems.

J. T. DIXON, B.E. in Electrical Engineering, Johns Hopkins University, 1924; S.M. in Electrical Engineering, Massachusetts Institute of Technology, 1926. American Telephone and Telegraph Company, Department of Development and Research, 1926–34; Bell Telephone Laboratories, 1934–. Mr. Dixon has been engaged in development work on carrier systems.

W. SHOCKLEY, B.Sc., California Institute of Technology, 1932; Ph.D., Massachusetts Institute of Technology, 1936. Bell Telephone Laboratories, 1936–. Dr. Shockley's work in the Laboratories has been concerned with problems in electronics.

C. F. WIEBUSCH, B.A., 1924, M.A., 1925, University of Texas. Instructor, Department of Physics, University of Texas, 1925–1927. Bell Telephone Laboratories, 1927–. Mr. Wiebusch was involved for a number of years in the development of disc recording and reproducing apparatus and loud speakers and for the past few years has been engaged in the development of telephone signaling and registering apparatus.