Speech Power and Energy

By C. F. SACIA

Introduction

IN the past, much research has been devoted to the determination of the relative magnitudes of the frequency components of speech, and the results of these explorations are useful and well known. Thus the communication engineer is apprised of the frequency range over which his apparatus should respond uniformly in order that the transmitted speech suffer no frequency distortion. But to provide against load distortion, he requires the knowledge of a different kind of data: numerical values of the magnitude of power involved in speech waves as a whole. This investigation deals with the magnitudes and forms of speech waves primarily in terms of power, and is not concerned with frequency as the argument.

Although the subject matter is not fundamentally new, this treatment of it is somewhat of a venture. The broad classification of power is a convenience here, but its future value will be dependent upon engineering usage. I have also introduced the use of the peak factor, which, being a simple index of the wave form, may perhaps find application in vowel study and phonetics as well as in the technical field. A condensed table of peak factors was incorporated in Mr. Fletcher's compilation in the preceding issue of this Journal.

Derivation

The nature of power in a syllable of speech may be most easily comprehended by reference to an illustration such as that shown in Fig. 1. The representation of the instantaneous power (P_i) is an enlarged copy of a power oscillogram of the word "quite." Because of its extreme jaggedness, the curve had to be represented by a profile rather than by an outline. Although this is a quickly spoken syllable it plainly displays a cyclic repetition; the cyclic interval (for example, from *a* to *b* in the figure) is ordinarily called the vocal period and its reciprocal, the vocal frequency[1]).

One feature of interest may be noted here: the irregularity in the growth and decay of the peaks. This is evidence of a slight vocal

[1] The power due to any periodic force, containing only odd harmonics, fluctuates with double the frequency of the fundamental; but in the case of any periodic force containing even harmonics also, the power fluctuations have the same fundamental frequency as the force. Although speech sounds are not periodic an analogous relation exists for them.

tremolo. Tremolos usually occur in singing voices and vary widely in their character. They constitute modulations which in actual singing sometimes occur as slowly as two per second. The slower modulations affect the ear as beats or pulses, while the most rapid ones affect the quality by the resulting sidebands of overtones. Those

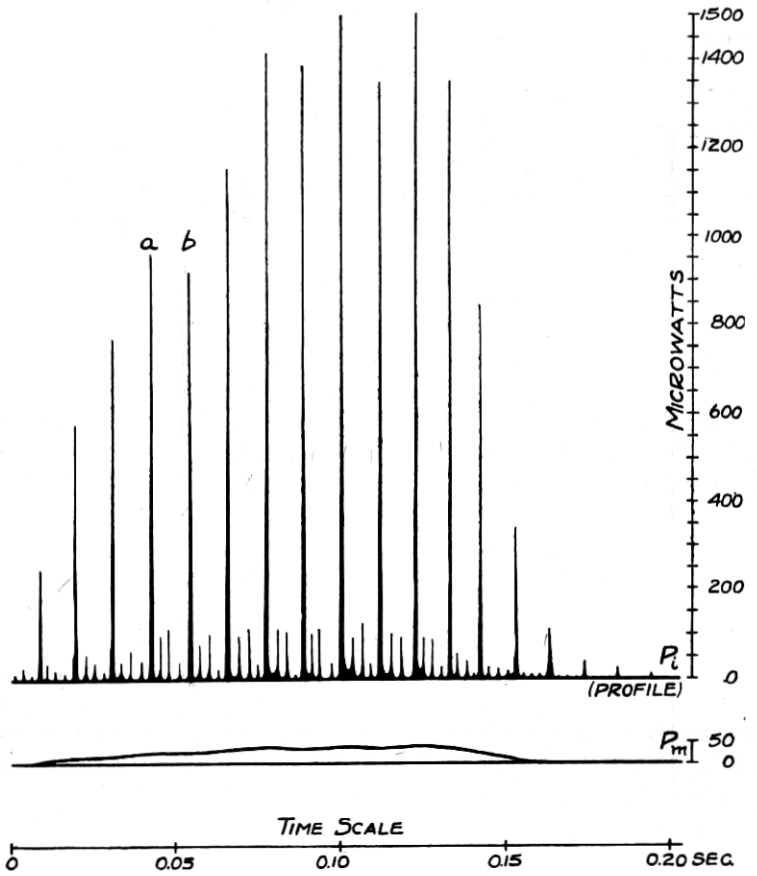

Fig. 1—Instantaneous and mean power. Enlarged copy of original oscillogram of the word "QUITE"

shown in the figure are of the latter types, their modulating frequency being about 50 per second.

From the instantaneous power we derive the mean power, P_m, whose chief significance lies in the fact that it is the kind of power that would be read by a quickly acting wattmeter; it is likewise proportional to the deflection shown by the ordinary a.c. voltmeter or

ammeter, or by the volume indicator. A graph of the mean power may be obtained by drawing the average power in each vocal cycle and then drawing a smooth curve through the resulting broken line. This would be an impracticable way of obtaining curves of mean power; actually they have been obtained independently of the P_i curves in this work, in a manner described later.

Vowel sounds carry by far the most of the power and energy of speech, and it was to them that the above considerations were tacitly applied; but the definition of the mean power is similarly applicable to the semi-vowels, voiced consonants, and fricative consonants.

The peak factor is the square root of the ratio of a peak value of P_i to the corresponding value of P_m.

Still another commonly used interpretation of power is made in terms of its average over an entire syllable, word or speech. Such an average, although the same for instantaneous and mean power, is most easily determined by means of the latter: it is the total energy divided by the time involved. Graphically it is the area of the P_i or P_m curve divided by the base. If the base includes the silent intervals between syllables the result will be called the long average; if the silent intervals are excluded from the base, the result will be called the short average.

Thus it is seen that the word "power" when applied to speech has a variety of meanings and always needs to be qualified. For example, the speech of a certain person may have shown a long average power of 10 microwatts while the instantaneous power frequently rose to 2,000 microwatts.

In obtaining the power, we obtain indirectly the pressure on the condenser transmitter, which is located 9 cm. from the speaker's lips. In the treatises on acoustics, the power of a simple-harmonic wave is derived in terms of the pressure,[2] the numerical result being at 20° C,

$$P = \frac{p^2}{415} \tag{1}$$

where P is the power in microwatts across 1 sq. cm. of wave front, and where either mean or peak value is taken for both power and pressure. Here we are not concerned with simple harmonic waves, but the same result holds for instantaneous, mean, or average values in any kind of wave, since

$$P = \frac{1}{10} p \frac{d\xi}{dt} \text{ microwatts across 1 sq. cm.,}$$

[2] See, for example, Rayleigh: Theory of Sound, Vol. 2, page 16.

and the air particle displacement,

$$\xi = \frac{1}{41.5}\int p\,dt \qquad \text{(41.5 is a resistance factor)}$$

for a wave travelling in the positive direction.

From the power intensity thus found at the transmitter we can obtain an estimate of the power developed by the speaker. With the transmitter surrounded by a plane reflecting surface so as to give reflection for speech frequencies, the pressure is doubled and the power intensity quadrupled over the values they would have in free air, hence the observed intensity is divided by 4. The usual assumption is made that this same intensity is distributed over a hemisphere whose center is at the speaker's lips. Hence the required estimate of the speaker's power is obtained by multiplying the measured power intensity at the transmitter by the factor $\frac{\pi 9^2}{2} \equiv 127$. For the sake of convenience, these two values are always given together in the accompanying tabulated results.

Instantaneous and Mean Power

In dealing with the power in a syllable, the matter of greatest interest is the maximum values attained by P_i and P_m throughout the entire syllable. These maxima will be denoted by $\overline{P}_i$ and $\overline{P}_m$, respectively. Table I shows their approximate ranges in the case of accented syllables.

TABLE I

Instantaneous and Mean Power

Typical Maximum Values for an Accented Syllable

	Speaker's Power Microwatts	Power Per Cm.² at Transmitter
$\overline{P}_i$	1000 to 2000	8 to 16
$\overline{P}_m$	60 to 120	0.5 to 1.0

At this point it is worth while to consider an application of the foregoing. A salient characteristic of speech waves is the generally high ratio of peak value to mean square value (peak factor), as can be inferred from Fig. 1. Failure to take this into account frequently causes load distortion in speech transmitting amplifiers. It sometimes happens that the effective output voltage or current has been measured, and the assumption of an equivalent sine wave (i.e., one having the same effective value) is made; but this leads to a large error in the estimate of the peak value. Thus with an insufficient allowance made for the peak voltage impressed upon the grid of the tube, there is the possibility of the grid becoming momentarily positive due to insufficient negative bias or still worse, the plate may be over-

loaded by the peaks. The resulting suppression of the peaks in the sound output can readily be detected by an accustomed ear, provided that the whole system is reasonably free from frequency distortion.

Average Power

In Tables II and III are summarized the observations made upon the two speeches which were used in this work. There are two reasons for showing them separately: the two speeches were not spoken in immediate succession; and they differ somewhat in character, the first being declamatory while the second is of a more conversational nature. This difference is not very great, but should account nevertheless, for the slightly higher values in Table II. By taking the weighted mean of the first number in both tables, we obtain 7.4 microwatts as the long average power in normal speech.[3]

TABLE II

First Speech, 50 Syllables

Average Power in Microwatts

	Long Average		Short Average	
	Speaker's Power	Per cm² at Trans.	Speaker's Power	Per cm² at Trans.
Composite of 16	8.6	0.067	13.1	0.102
Composite of 8 male	8.2	0.064	12.7	0.099
Composite of 8 female	9.0	0.070	13.5	0.105
Maximum male	10.6	0.082	17.1	0.133
Maximum female	17.0	0.131	21.8	0.169
Minimum male	7.0	0.055	10.8	0.084
Minimum female	5.7	0.044	8.8	0.069

TABLE III

Second Speech, 72 Syllables

Average Power in Microwatts

	Long Average		Short Average	
	Speaker's Power	Per cm² at Trans.	Speaker's Power	Per cm² at Trans.
Composite of 16	6.6	0.054	9.9	0.080
Composite of 8 male	6.2	0.050	8.9	0.072
Composite of 8 female	7.1	0.057	10.8	0.087
Maximum male	8.1	0.065	13.0	0.105
Maximum female	9.8	0.079	15.1	0.122
Minimum male	3.9	0.032	5.7	0.046
Minimum female	4.0	0.033	6.0	0.048

NOTE: The average ratio of the total time in the silent gaps to that consumed by the syllables is 0.55; the syllables average 0.16 sec.

[3] Crandall and MacKenzie gave an estimate of 12.5; B. S. T. J., Vol. 1, No. 1; Phys. Rev., Mar. 1922.

Stress

Since our observations have shown qualitatively that the louder syllables have the greater rise of mean power, means are available for calibrating the stress modulation of the voices under test. To form a discriminant for each speaker we proceed in the following way:

(1) Measure the $\overline{P}_m$ of each syllable;
(2) Find the ratio of each $\overline{P}_m$ to the greatest $\overline{P}_m$ occurring in the speech; call this ratio ϵ;
(3) Find the proportional number, $s/\underline{s}$, of syllables for which ϵ is greater than the magnitude n, where n may vary between 0 and 1;
(4) Plot the variables $s/\underline{s}$ and n against each other to give the required curve.

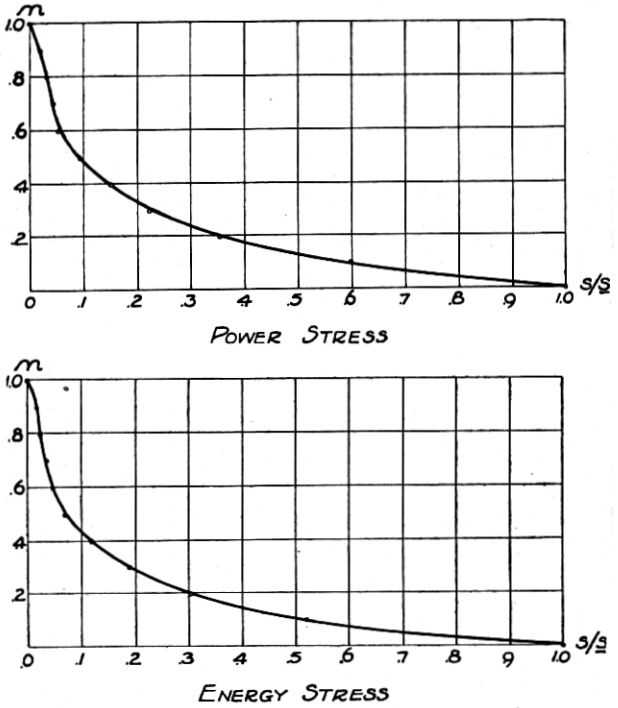

Fig. 2a—Composite stress curves of 16 voices

The analogous relation between syllabic energy and stress is found by using the total energy of each syllable instead of P_m in the above.

A large number of these curves has been so obtained, but it will suffice to consider here a few of the representative types. Fig. 2a

shows composite curves and Fig. 2b gives a series of each kind of curves for four speakers. Note the changing mode of stress which is shown in the sequence from top to bottom: in the first case the syllables of weaker stress greatly predominate while in the last case there is a more nearly uniform distribution of the syllables with respect to the degree of stress. It is evident from a comparison of the two series

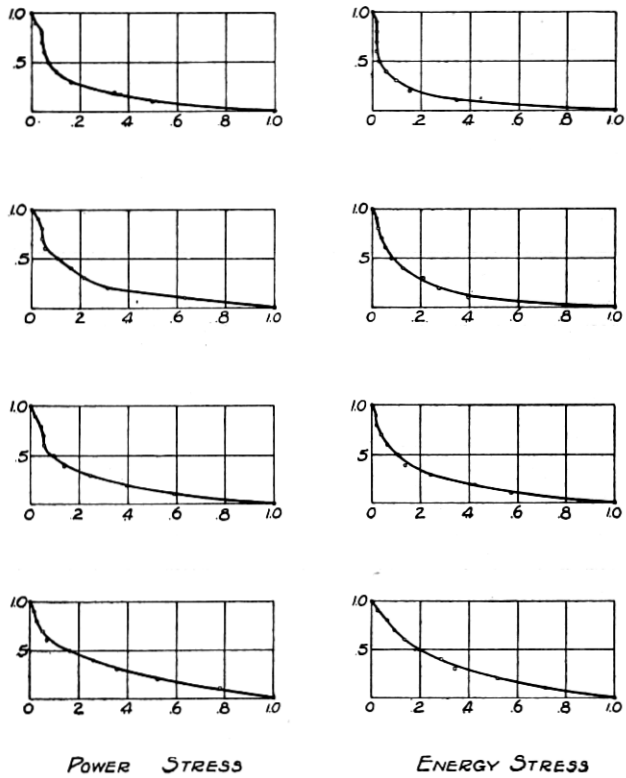

Fig. 2b—Types of stress curves

that the speaker's type is much the same whether judged by the power or energy standard. An exceptional case might arise, however, if one should put emphasis on a syllable by prolonging the time of utterance, for here the increased energy of the syllable would not necessarily mean a greater stress. But from the point of view of phonetics, the energy method should be useful in calibrating emphasis, which can be taken as a function of time of duration as well as of mean power.

Relative Power of Vowels

One test which was made on the speakers was for them to utter disconnectedly and without accent eleven monosyllables, each of which contained a fundamental vowel sound. The results of this test give a general indication of the inherent power, $\overline{P}_m$, in unaccented (but unslighted) vowels relative to each other. The difference between the

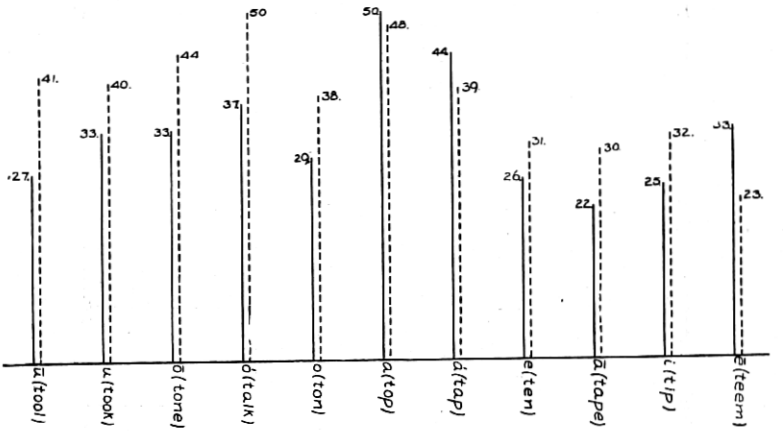

Fig. 3—Inherent relative power

——— Indicates Male Voices ------ Indicates Female Voices

Numbers indicate approximate power from voice (in microwatts)

male and female voices in this respect warrants separate charting of these characteristics. Fig. 3 shows the chart in which the vowels are arranged in the sequence[4]) the first half of which accompanies an increase in the angle of the speaker's jaws, and the succeeding half accompanies an increase in the elevation of the tongue.

It might have been anticipated that the more open vowels have more power; but there is apparantly one irregularity in this tendency in the case of the vowel o (as in ton). Furthermore, the vowel ē (as in teem) looks somewhat different for the two voices, when compared with the vowels immediately preceding it in the series. There is some difficulty in uttering it so as to make it carry, in the case of female voices—a fact which I have previously encountered when recording them. The male voice, on the other hand, shows a decided rise in this direction. The advantage in the case of ū (tool) is reversed: here the male voice begins to fall off while the female voice stays about the same. These results suggest a difference in the resonant structure

[4] This arrangement is based upon the well known vowel triangle of Vietor.

between the male and female voices, which, however, does not affect the higher frequencies enough to alter the vowel characteristics.

Peak Factor

The tests just described were also used to obtain the peak factors of the vowels. These were determined by measurement of the maximum P_i and P_m of each syllable and are charted in Fig. 4. Here again there

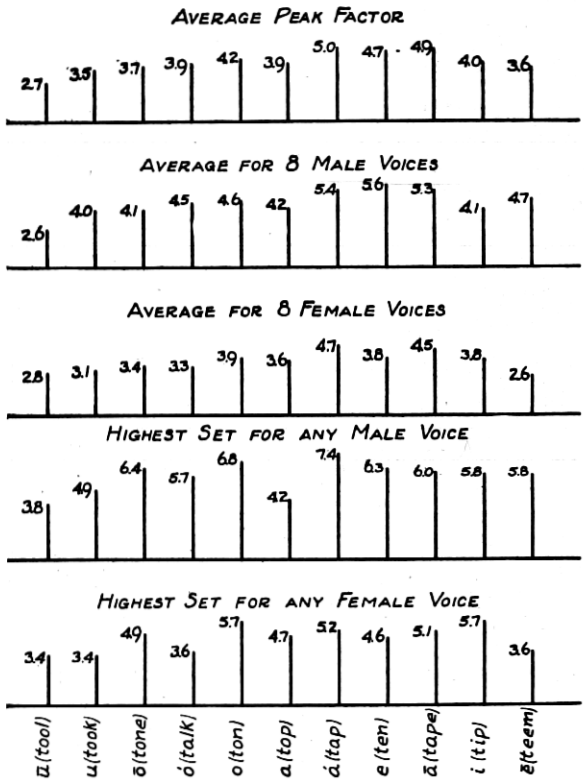

Fig. 4—Peak factors of vowels

are differences between the sets for the male and female voices, the former being somewhat higher, especially for the vowel ē. In both cases such rasping vowels as á (tap), e (ten), ā (tape) have sharp waves and high peak factors. Having listened attentively to all these voices under test, I have become able to associate peak factors with vocal qualities in the following way: the voices with the higher peak factors are those which in the ordinary terminology are said to be "resonant"

or "vibrant"; they have the greater carrying power, especially over the telephone; they are rich in the musical sense and are therefore well suited to singing, although many such voices, unfortunately, are never applied to the art.

To illustrate an application of the peak factor to engineering, we shall again take into consideration the speech amplifier whose mean effective output voltage is indicated by a suitable device such as a volume indicator. From this, the peak value of the instantaneous voltage is wanted; to find it necessitates a knowledge of the peak factor. Now since the latter differs somewhat for different sounds and speakers, it is necessary to use one factor which makes allowance for the worst cases (highest voltage peaks) which can occur often. For most purposes, the factor 5 will suffice, hence the rule is: the mean effective voltage should not exceed one-fifth the overload voltage of the system.

Apparatus

In order that the apparatus (see Fig. 5) be a faithful recorder, it was made with the following characteristics:

(1) A nearly distortionless reproduction of wave form by the condenser transmitter and amplifier.

(2) A full-wave parabolic rectification of the amplifier output.

(3) Load capacity sufficient to transmit the high sharp peaks of speech waves without cutoff.

(4) Uniform response, from 0 to 6000 cycles in the oscillograph vibrator recording instantaneous power.

The calibration of the amplifier and condenser transmitter is shown in Fig. 6. To make the overall characteristics so nearly uniform it was found necessary to use the resonant circuit in the output of the second N tube, this compensating for an irregularity due mostly to the 45 feet of cable which leads from the transmitter and first stage of amplification in the sound-proof room to the main part of the amplifier.

The oscillograph (see Fig. 5) was provided with two series connected vibrators one of which was sensitive to low frequencies only, and recorded the mean power. Although it did not completely suppress the fluctuations of vocal frequency, it reduced them to the order of small superimposed ripples through which the P_m curve could be drawn. The instantaneous power was recorded by the other vibrator whose characteristics are noted in item (4) above.

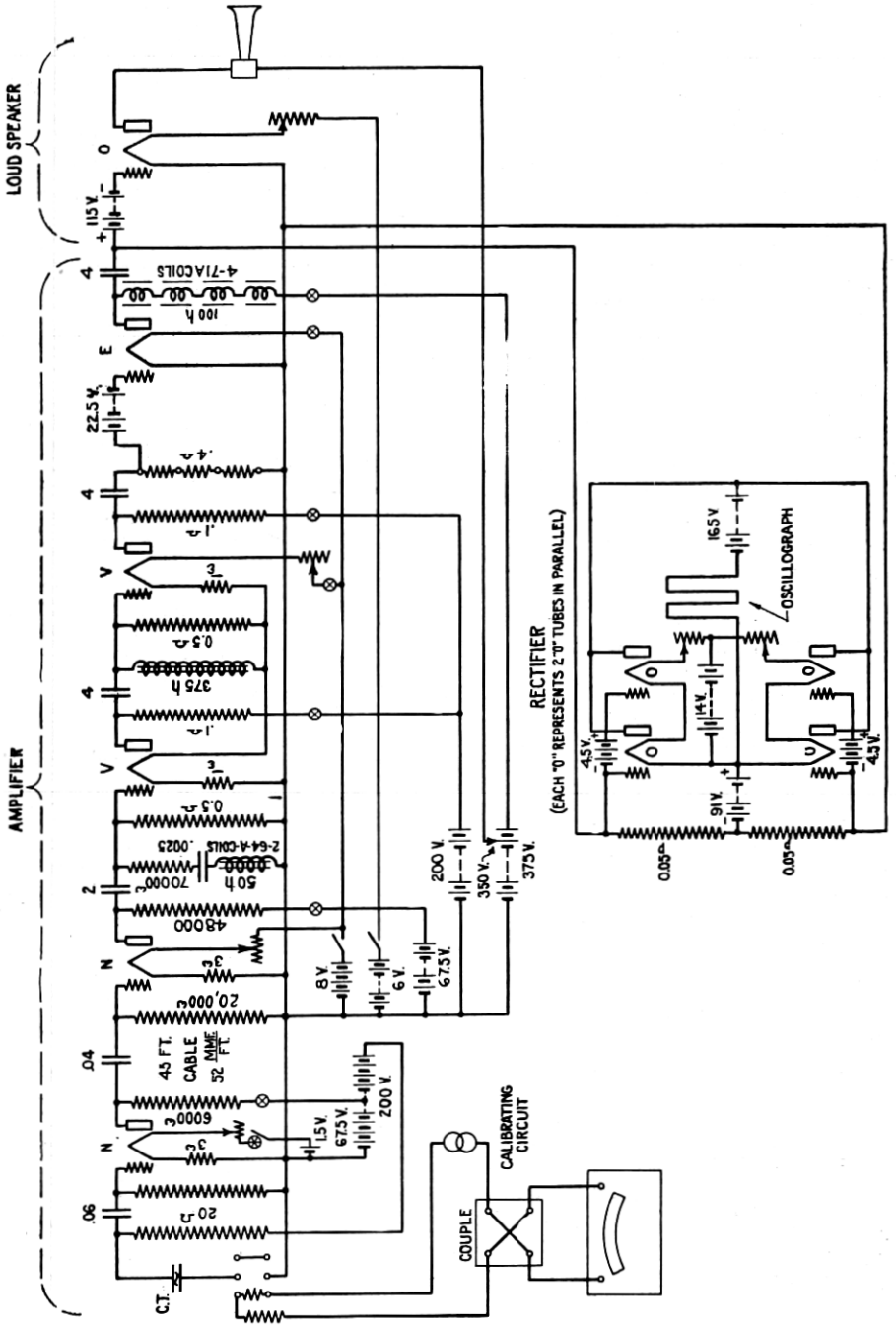

Fig. 5—Speech power recording circuit

TABLE IV

Calibration Constants

(a) Constants of Vibrators $I/D =$
  (1) Low frequency ..... 5 { milliamperes
  (2) Instantaneous power ..... 286 { per cm.
(b) Rectifier constant $E^2/I = /40$ (volts)^2/milliamp.
(c) Pressure on transmitter vs. amplifier output $p^2/E^2 = 1/2.95^2$ dynes^2/cm^4 volt^2.
(d) Power intensity at transmitter vs. pressure $P/p^2 = 1/415$ cm^2 microwatts/dynes^2.

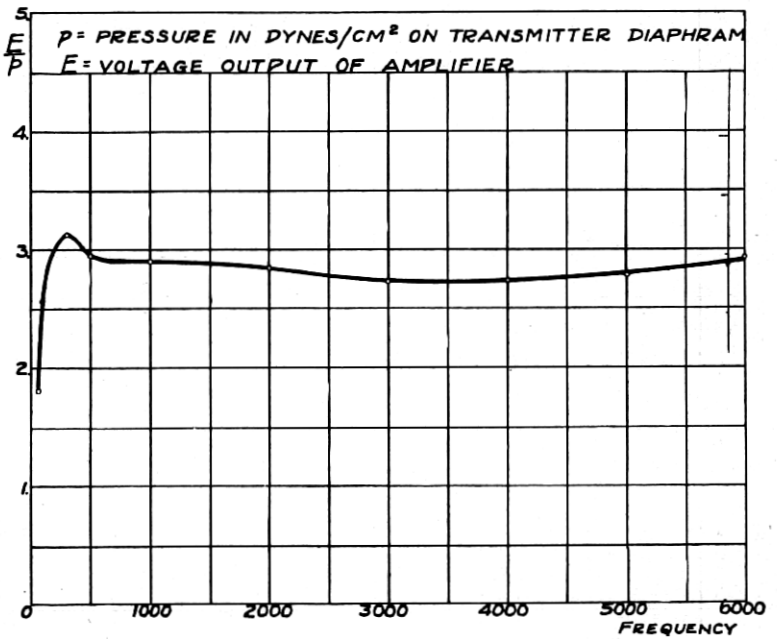

Fig. 6—Calibration of condenser transmitter with amplifier

The product a b c d gives $P_m/D_m = 0.192$ microwatts per sq. cm. of wave front as indicated by a deflection of 1 cm. of the oscillograph low frequency vibrator. Similarly $P_i/D_i = 11.1$ for the instantaneous power vibrator.

Method

Records were made on sensitized paper strips 6 cm. wide moving at a velocity of about 20 cm. per second. Three graphs were traced simultaneously, the instantaneous power, the mean power, and the timing wave of 100 cycles from an oscillator. When connected speech was being recorded, the oscillograph operator listened to the speech as reproduced by the loud speaker and punctuated the record at frequent

predetermined points by tapping a key which momentarily displaced the timing wave. By the aid of these punctuations we were enabled to identify the words and syllables on the records after development. The areas for computing average power were measured from the mean power curve, while the instantaneous power curve was measured only for its peak values.

Although chosen at random, the speakers used in these tests represent all sections of the United States. Their types range from soprano to bass-baritone, neither extreme type—high soprano and bass—being available; but this assortment is sufficiently representative for our purpose. Extraneous disturbances were to a large extent eliminated by the sound-proofing on the walls and ceiling. Lest the novelty of this situation be a distraction to the speaker, he was allowed to practice and become accustomed to the new condition.

Conclusion

One advantage in having speech data available in terms of its power rather than its amplitude is the fact that in most instruments used for making quantitative speech measurements, the force which operates the meter is proportional to the square of the wave amplitude. Common examples of such instruments are the dynamometer and the ordinary a.c. meters.

To summarize, the power is classified into:

1. Instantaneous power, P_i.
2. Mean power, P_m.
3. Long average power.
4. Short average power.

Stress calibrations are here derived from the maximum values of P_i and P_m ($\bar{P}_i$ and $\bar{P}_m$, respectively) in each syllable, while the use of the total energy of the syllable for calibrating emphasis also shows possibilities. The peak factor is the square root of P_i/P_m and is a useful index of the wave form.

The measuring apparatus—excluding the rectifier and oscillograph—is essentially a good quality speech-transmitting system. In view of the fact that good quality systems are now used commercially as well as in the laboratory the data naturally fall into two classes:

(1) Measurements which characterize the speech solely from the standpoint of the transmitting apparatus;

(2) Estimates or approximations concerning the total power from the voice.

Regarding (1) we note that the divergence of waves causes some frequency distortion which is greater, the nearer the source, and becomes negligible as the distance increases (see the appendix). We should accordingly expect the peak factors to be different at the speaker's lips. The estimates of total power, however, are as close as their importance necessitates.

When the data are applied to a case in which the speaker's distance is other than 9 cm., the required power intensity is found by the law of inverse squares and the pressure by the law of inverse distance.

APPENDIX

Frequency Distortion in Spherical Waves

A spherically diverging sound wave (see H. Lamb: "Dynamical Theory of Sound," page 206) is represented by

$$r\phi = f(v_o t - r)$$

where r = radius of the wave front

ϕ = velocity potential

t = time

v_o = velocity of sound

ρ_o = mean density of air

The pressure

$$p = -\rho_o v_o \partial\phi/\partial r$$

$$= \rho_o v_o \left[\frac{1}{r} f'(v_o t - r) + \frac{1}{r^2} f(v_o t - r)\right]$$

Let $f(v_o t - r) \equiv \sin \omega \left(t - \frac{r}{v_o}\right)$,

so that

$$p = \frac{\rho_o v_o}{r}\left(\frac{\omega}{v_o} \cos \omega \left(t - \frac{r}{v_o}\right) + \frac{1}{r} \sin \omega \left(t - \frac{r}{v_o}\right)\right).$$

When a wave composed of any number of such components (each having a different pair of values for ω and α) diverges from one radius to a larger one, it not only changes in size, due to the factor $\frac{\rho_o v_o}{r}$ but also in shape, due to the factor $\frac{1}{r}$ in the second term. When r

is large compared with $\frac{v_o}{\omega}$, this change in shape becomes negligible. In the case of speech, since the source is of finite size the effective radius is somewhat greater that that measured from the speaker's lips, and the wave front is not exactly hemispherical, so the comparison is only qualitative. Nevertheless, a difference in quality of transmitted speech can be detected when the speaker's lips are within 2 cm. of the transmitter diaphragm.