Analysis of the Energy Distribution in Speech[1]

By I. B. CRANDALL and D. MacKENZIE

SYNOPSIS: *The frequency distribution of energy in speech* has been determined for six speakers, four men and two women, for a 50-syllable sentence of connected speech, and also for a list of 50 disconnected syllables. The speech was received by a condenser transmitter whose voltage output, amplified 3,000 fold, was impressed on the grids of twin single stage amplifiers. The unmodified output of one of these amplifiers was measured by a thermocouple and was a known function of the total energy received by the transmitter, corrections being made for the slight variation with frequency of the response of the circuit. The output of the other amplifier was limited by a series resonant circuit to a narrow band of frequencies, the energy in this band being measured by a second thermocouple. The damping of the resonant circuit was so chosen that sufficient resolving power and sufficient energy, sensitiveness were obtained over the range from 75 to 5,000 cycles per second; and 23 frequency settings were made to cover this range. For each syllable simultaneous readings were recorded on the two thermocouples at each frequency setting. The consecutive syllables were pronounced deliberately by each speaker, maintaining as nearly as possible the normal modulation of the voice. Corrections were applied to offset the unavoidable variations in total energy incidental to repetition of a given syllable. 13,800 observations were made for all speakers. *The energy distribution curves* obtained are essentially the same for connected as for disconnected speech, and indicate that differences between individuals are more important than variations due to the particular test material chosen. A composite curve drawn from the individual curves shows a great concentration of speech energy in the low frequencies, a result which would not be expected from data previously published by others. The actual results contain a factor due to standing waves between the speaker's mouth and the transmitter, a complication always present in telephoning; this could not be eliminated.

The rate of energy output in speech for the normally modulated voice, was determined from the readings for total energy and was found to be about 125 ergs per second.

IN the study of speech and its reproduction by mechanical apparatus it is necessary to consider its composition from several different points of view. We desire first of all to know the actual frequency distribution of the total energy in speech, as well as the separate distributions for each individual sound. We also desire to know the apparent distribution of energy, that is, the distribution as perceived by the ear. Finally, we wish to know the importance of each frequency, that is, the contribution to "articulation" or "quality" in the exact reproduction of speech which can be traced to the energy of each elementary band of frequencies in the speech range. In all three cases certain frequency functions are used to represent these distributions. The advantage of considering these different frequency distribution functions separately has already been indicated by one of the present writers.[2]

[1] Reprinted from THE PHYSICAL REVIEW, N.S., Vol. XIX. No. 3, March, 1922.

[2] "The Composition of Speech," PHYS. REV., X, p. 74, 1917.

In our judgment the most important of these data of speech study is the *actual energy distribution*, considering speech as "a continuous flow of distributed energy," in accordance with the ideas expressed in the earlier paper. The present paper offers a determination of this fundamental factor.

To determine the energy distribution in speech to a high degree of accuracy it would be desirable to analyze a certain amount of connected speech and take a time average of the energy distribution of the whole. This is not feasible at the present time, but a very close approach to this result has been made. The method consists in analyzing the speech waves as impressed on a condense transmitter, using a tuned circuit to transmit narrow frequency bands of energy and pronouncing the separate syllables of the connected speech so slowly that the kick of a direct current galvanometer connected to an A. C. thermocouple can be separately read for each syllable. Using a suitable calibration for the whole apparatus, the magnitude of this kick can be interpreted in terms of the time integral of the energy at a particular frequency setting for each syllable. A mean of the readings for all the syllables in the "speech" at any frequency setting gives the relative energy at that frequency.

The present method is a modification of an earlier method in which approximate analyses of speech sounds were made, using a condenser transmitter, tuned circuit, an amplifying-rectifying circuit, and ballistic galvanometer. The method is, however, much improved as we now have very accurately calibrated condenser transmitters of better design,[3] and a great deal of care has been taken to calibrate the successive elements of the train of apparatus, and increase the resolving power.

Experimental Procedure

Sound waves emitted from the mouth of the speaker are allowed to fall upon the diaphragm of a condenser transmitter, connected in the conventional manner to the input of a three-stage amplifier. The output of this is impressed upon the input circuits of twin single stage amplifiers, potentiometers being interposed to permit regulation of the grid voltages of the twin amplifier tubes.

The output circuits of the fourth stage consist of the high windings of two step down ironclad transformers. These step down transformers have a voltage ratio of 11: 1 and are designed to work between impedances of 6,000 and 50 ohms. The low impedance winding of one of these transformers operates into a thermocouple heater of,

[3] The present design of the condenser transmitter and its calibration are fully treated in a paper by Dr. E. C. Wente which will appear shortly in this Journal.

roughly, 40 ohms resistance. The low side of the other transformer operates through a tuned circuit into a similar thermocouple heater.

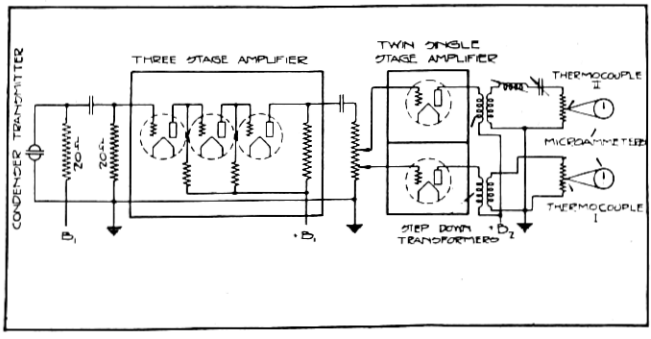

Fig. 1—Circuit Used for the Analysis of Speech. (The Usual Details of the Three-Stage Amplifier Are Not Shown)

The diagram of Fig. 1 exhibits the essential features of the electrical circuits just described.

When the diaphragm of the condenser transmitter is set in vibration by speech a current made up of a range of frequencies flows in the heater of thermocouple I., while the heater of thermocouple II is traversed only by such a band of frequencies as the resonant circuit allows. Fig. 2 shows a number of typical resonance curves obtained

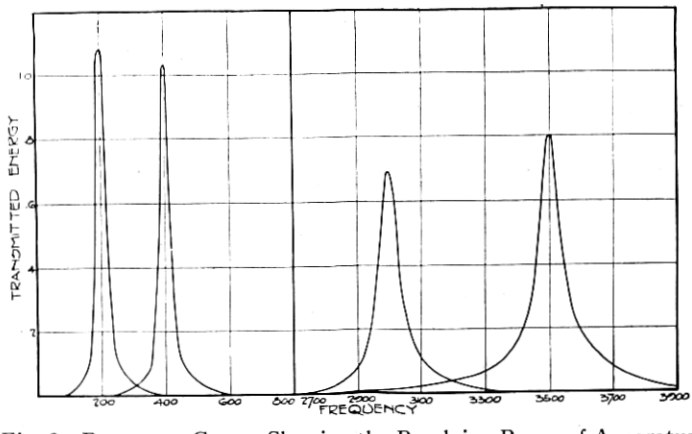

Fig. 2—Resonance Curves Showing the Resolving Power of Apparatus

in the course of calibrating this apparatus. These curves are such that the tuned circuit functions as a filter transmitter only a narrow region of frequencies. One side of the twin amplifier transmits the

entire electrical response of the system; the other side suppresses all save a band of frequencies, the center of this band being shifted by resetting the condenser and inductometer.

Having chosen for analysis a piece of connected discourse, the speaker utters the successive syllables separately but as nearly as may be with the same inflection and volume as if the syllables were continuously spoken. Two observers record the readings of microammeters in the couple circuits of the thermocouples. One of these instruments gives a deflection corresponding to the total energy of the syllable uttered; the deflection of the other instrument corresponds to the energy of the syllable lying within the limits of transmission of the tuned circuit.

Preliminary experiments were carried out to determine the relation between momentary deflection read on the microammeter, and the current momentarily flowing in the thermocouple heater. Currents of different values were caused to flow for intervals of time varying from 0.2 second to 1.2 seconds, and the deflections were found nearly proportional to the product of current squared and time interval; this proportionality was most nearly exact when the current was weak and the time intervals short. For all cases likely to be duplicated in the speech analysis work the error might be taken as about 5 per cent, a quantity small in comparison with the inevitable uncertainties due to other causes.

Quite low damping is attained in the resonant circuit. The values of inductance used ranged from 0.20 to 0.66 henry and the total resistance of the circuit—transformer winding, inductometer coil, thermocouple heater—is of the order of 100 ohms. The damping thus ranges from 75 to 250.

The circuit is calibrated in the following manner:

A switch is so introduced that it is possible to include in series with the thermocouple the resonant circuit, or replace it by a non-inductive resistance whose value is approximately that of the A. C. resistance of the inductometer winding. With the tuned circuit excluded, an alternating current of suitable magnitude is caused to flow in the thermocouple heater; the tuned circuit is then substituted and the new value of the current observed, the input voltage remaining constant. The ratio of current squared "tuned circuit in" to current squared "tuned circuit out" is plotted against frequency, yielding a curve for energy transmission.

Twenty-three bands in all were considered adequate for the analysis of energy distribution in speech; the centers of these were at 75, 100, 200, 300 cycles, 400 to 3,200 cycles by steps of 200; 3,500, 4,000, 4,500,

5,000 cycles per second. Beyond 5,000 cycles per second, the energy is so low as to be impossible of measurement with the apparatus used. A Weston Type 322 microammeter recorded the couple current for the tuned circuit side of the twin single stage amplifier. With this instrument and the thermocouple used, 0.2 microampere in the couple circuit corresponds to one-quarter of a milliampere in the heater, and this is the lowest readable deflection of the Weston instrument.

Reduction of Observations

Three corrections have to be made, the first being the correction for varying volume.

Simultaneous observations are made, at each setting of the tuned circuit, of the filtered and the unfiltered energy of each syllable. It is not possible to utter a given syllable with the same intensity and at the same distance from the transmitter for every one of twenty-

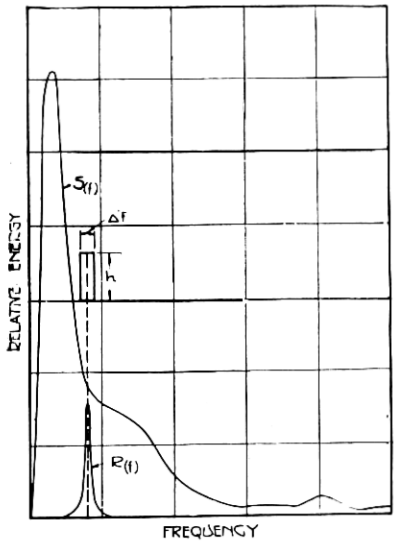

Fig. 3—Illustrating Correction of Observations, Necessary Because of Variation in Resolving Power with Frequency Setting

three times. Accordingly, the "unfiltered" readings are averaged and each of the filtered readings for each syllable reduced from the value actually observed to the value that would have been read had the volume and distance been such as to give the average "unfiltered" reading. This procedure is quite legitimate if it be granted possible to maintain a definite composition of the syllable in question throughout the changes of the tuned circuit setting.

A second correction was made for the varying area of tuned circuit curves.

In Fig. 3 let $S(f)$ be the speech spectrum determined by ideal methods; "R" the transmission curve of the tuned circuit, set for a resonant frequency f. An ideal transmission curve would be a rectangle when plotted in this figure, of height "h" and transmission range Δf.

The true amount of energy $S(f)$ associated with frequency f, and the experimentally determined value which we may call $(\overline{S}f)$ are connected by the relation

$$h\overline{S}(f)\Delta f = \int_f^{f+\Delta f} S(f)R(f)df$$

and if we make

$$h\Delta f = \int_f^{+\Delta f} R(f)df$$

we may take for all practical purposes $S(f) = \overline{S}(f)$, considering the narrowness of the transmission range. We must therefore find the factor $h\Delta f$, proportional to the area of each tuned circuit curve and divide the energy received through the filtered side by $h\Delta f$, in order to obtain $S(f)$. This treatment may be gone through for each syllable individually, but it is more convenient to sum the tuned circuit readings for all the syllables used, corrected one at a time for varying volume, and then apply the curve area correction to this sum.

A third correction was made for the varying frequency-sensitivity of the whole apparatus. Thus far we have discussed only the electrical energy in the output circuit of the fourth stage. It remains to show in what way this is related to the mechanical energy of the diaphragm, and this in turn to the incident sound energy.

The calibration of the circuit as a whole was made by introducing a small resistance carrying alternating current in series with the condenser transmitter, thus introducing a known potential drop in the undisturbed input mesh of the circuit.

An amplification curve is appended (A, Fig. 4) which gives to an arbitrary scale the ratio of volts output to volts input as a function of frequency, for the system as actually operated. The calibration of the condenser transmitter, shown in Fig. 4, Curve C, gives the open circuit voltage of the transmitter per unit pressure on the diaphragm as a function of frequency. The product of these curves is the volts output per unit alternating pressure on the diaphragm, and the square of this product, curve E is proportional to the electrical energy output per unit sound energy incident on the diaphragm, if we assume that the sound energy is proportional to the square of the alternating pressure. This point, however, requires some further discussion, which will be given later on.

It is plain from curve E that the response of the system is a maximum of frequencies in the neighborhood of 2,250 cycles. If, now, the observations already corrected for varying volume and for area of resonance curves, are subjected to further correction for the exaggeration of these frequencies, it is possible to draw a curve which shall

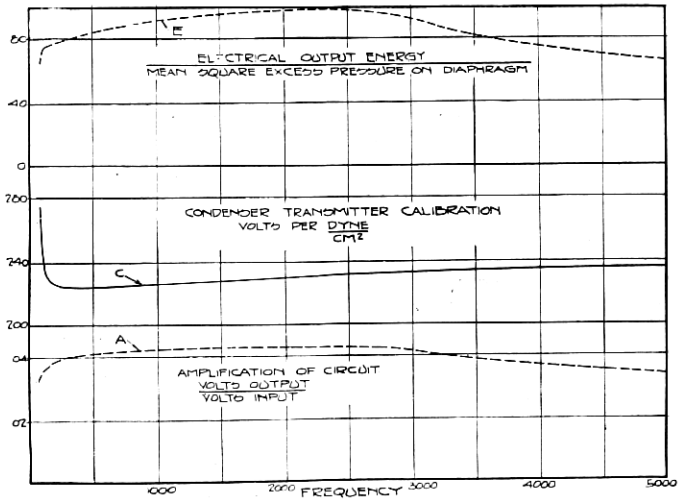

Fig. 4—Energy-Frequency Characteristics of the Apparatus

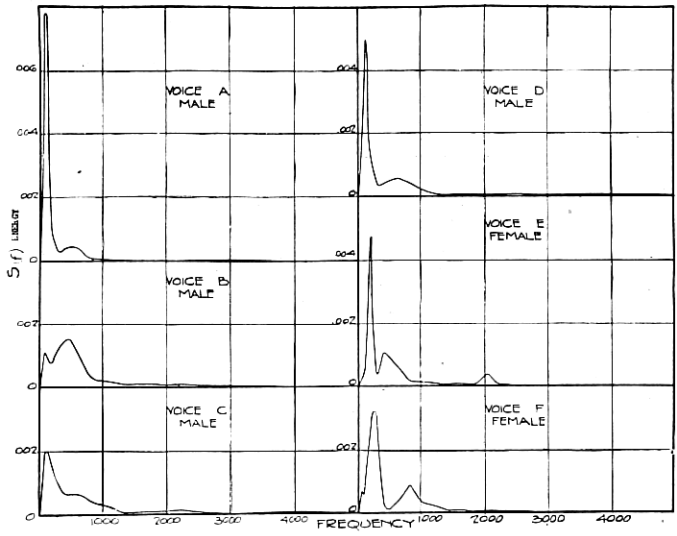

Fig. 5—Analysis of Individual Voices

exhibit the mean square of the excess pressure on the diaphragm, as a function of frequency in the voice exciting the vibration. We obtain this corrected curve by dividing the results, after the first and second corrections above have been made, by the ordinates of curve *E*.

OBSERVATIONS

In order to investigate the possibilities of this method it was decided to work with a rather short piece of connected speech, and to use a limited number of observers, on account of the large number of observations which are required for each separate syllable. With six speakers (four men and two women) each pronouncing the test sentence of fifty syllables for each of the twenty-three frequency settings, 6,900 separate observations were required. It is believed that representative results have been obtained from these observations, but if this is not the case then some method of graphical registration of the energy-time curve of speech for the different frequency settings must be applied in order to handle the vast amount of data involved in work on an appreciably larger scale.

The test sentence used was as follows:

"*Quite* four score and seven years ago our father brought forth on this continent, a *nice* new nation, conceived in liberty, and dedicated to the proposition that all men are created equal."

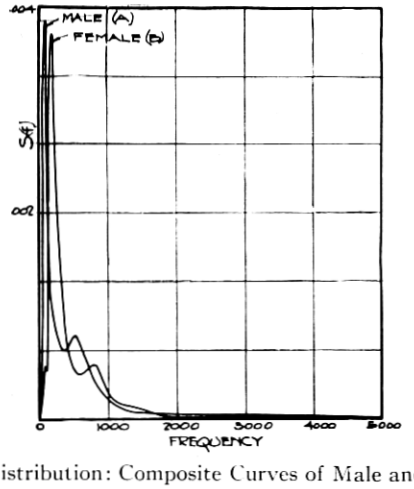

Fig. 6—Energy Distribution: Composite Curves of Male and Female Voices

The two *italicized* words were added to the first sentence of the "Gettysburg Address" in order to bring the total up to fifty syllables, and improve the balance between the vowel sounds.

The resulting speech-energy curves are shown in Figs. 5, 6 and 7,

plotted so that $\int_0^\infty S(f)\,df = 1$ in each case. In Fig. 5 the individual curves for each of the six speakers are shown on a small scale; in Fig. 6 the composite curve for the men and the composite curve for the women, drawn separately, and in Fig. 7 the composite curve for all six speakers, giving the data of curves 6*A* and 6*B* equal weight.

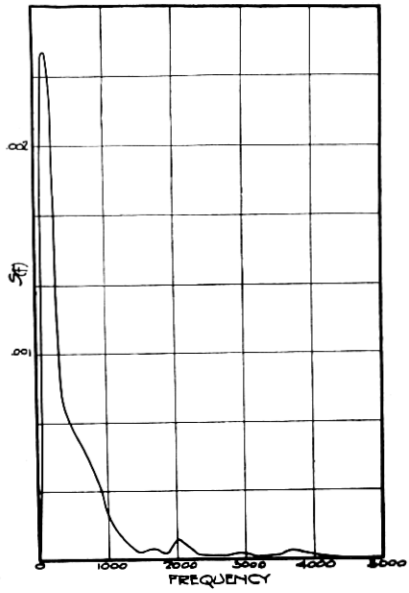

Fig. 7—Energy Distribution: Composite Curve for All Voices

These curves are very similar to a curve obtained by Dr. Fletcher of this laboratory, using block filters and based on the simple calling sentence "Now we're off on one." A general consideration of this fact and of the data shown leads us to believe that the differences between curves of this sort, made by the method described are due rather more to differences between the voices of the individual speakers than to the particular piece of connected speech which is chosen, provided the speech is of reasonable length. The differences between the different voices are so marked that we should expect them to remain even though we used as test material a connected speech ten or fifty times as long as the sentence used.

The Energy Distribution in Speech

An interesting comparison may be made between the curves shown for the energy distribution of "continuous speech" and certain speculative curves previously constructed to indicate the energy distribu-

tion. One of these curves is shown in Fig. 8. Curve *A* was constructed by one of the writers in 1916 in an attempt to synthesize the energy curve from the energy distributions of the vowel sounds, using the vowel analyses of Dr. Dayton C. Miller. Curve *C* is the composite "continuous speech" curve of Fig. 7. The vowel sounds analyzed by

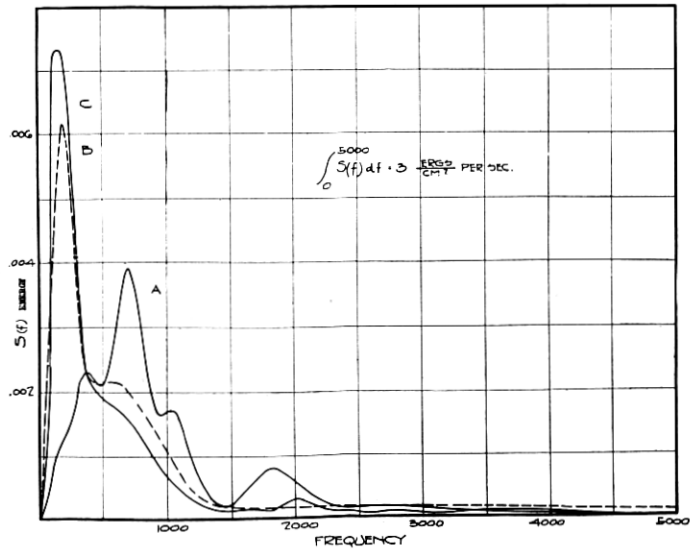

Fig. 8—Energy Distribution: *A*. Synthesized from Vowel Records of D. C. Miller (1916). *B*. Disconnected Speech Analysis of this Paper. *C*. Connected Speech Analysis of this Paper (from Fig. 7)

Miller were intoned and the vowel sounds analyzed by us were spoken, but Miller's work seemed to show that there was no essential difference between intoned and spoken vowel sounds. There is, however, a very noticeable difference between Curve *A* and Curve *C*, the energy in the fundamental tone of the speaker's voice coming out much more strongly in Curve *C*. We should expect that our improved apparatus would record the energy in the lower frequencies more correctly than the apparatus heretofore used but as we used different test material (connected speech instead of disconnected syllables or vowel sounds) it is not immediately evident which of these two factors is responsible for the differences between the *A* and the *C* curves.

In order to investigate this point more fully the testing routine for all six speakers was repeated, using instead of the fifty-syllable sentence, the fifty disconnected syllables of one of the standard articulation testing lists, as used by Dr. Fletcher in this laboratory. The results for energy distribution are shown in Fig. 9, Curve *A* being

the mean energy distribution for the four male speakers, using the syllables, while Curve *B* is the mean energy distribution of the two female speakers. Curves 9*A* and 9*B* may be compared with Curves 6*A* and 6*B* which represent the sentence of continuous speech. The two sets of curves are essentially the same as shown in Fig. 8, *C* and *B*

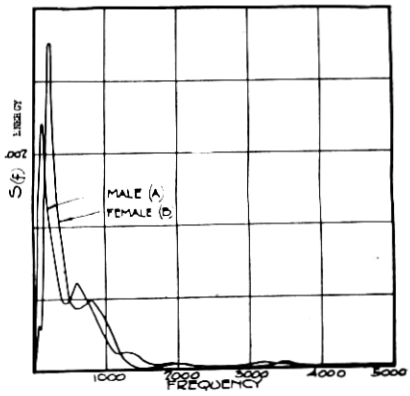

Fig. 9—Energy Distribution in Disconnected Speech

being respectively the composite curves for all speakers, using connected and disconnected speech.

Such small differences as exist between Curves *C* and *B* of Fig. 8 may probably be due to differences in the distribution of the vowel sounds in the connected and disconnected test material. This distribution is given in the following table:

Vowel Sounds	a	ā	á	e	ē	ér	i	ī	o	ō	ó	u	ū	ou	Total
In Sentence.....	6	6	3	7	3	3	7	2	2	5	3	0	2	1	50
In Syllabic List (No. 174).....	4	4	3	4	3	4	3	4	3	4	3	4	4	3	50

Key to Vowel Sounds: a, as in father (or o as in top); ā, as in tape; á, as in tap; e, as in ten; ē, as in team; ér, as in term; i, as in tip; ī, as in time; o, as in ton; ō, as in tone; ó, as in for; u, as in pull; ū, as in rule; ou, as in house

The similarity between Curves *C* and *B* of Fig. 8 is evidence of the general reliability of the method, and leads to two rather important conclusions.

In the first place, characteristic results have been obtained for a given set of speakers, using two different types of test materials. This

seems to show that the choice of test material does not require especial consideration, provided it is of sufficient length. It seems to be a matter of rather greater importance to increase the number of observers.

In the second place, it seems that for the actual energy distribution, the results previously obtained from the vowel analyses are definitely in error, in that they show relatively little energy associated with the lower voice frequencies.

Criticism of the Results

The foregoing treatment provides a curve showing the frequency distribution of the square of the excess pressure on the diaphragm.

In an undisturbed field of sound energy we have for the intensity

$$I = \frac{P^2}{2\rho a}$$

in which ρ is the mean density of the medium, a the velocity of sound in the medium and P the maximum excess pressure.

It remains for us to consider in how far the results obtained represent the frequency distribution of sound energy in speech.

Due to the fact that at frequencies where the sound wave-length is short and comparable with the diameter of the transmitter, considerable reflection takes place, and the pressure on the diaphragm is proportionately greater for these frequencies than for those which are not accompanied by strong reflection. In this respect again the higher frequencies provoke the greater response in the system.

The following experiment was tried to investigate this variation. A wall six feet square, with a central hole to fit over the condenser transmitter, was brought up to make the transmitter a part of a plane wall. The clearance around the periphery of the transmitter was tightly closed, and reflection was to be expected at all frequencies. Where total reflection takes place, a given quantity of sound energy results in twice the alternating pressure on the diaphragm as when no reflection occurs. That is, the resulting electrical energy observed should be four times as great for total reflection as for no reflection. The wall was expected to cause reflection at all frequencies, and the experiment consisted in reading the electrical response, with and without the wall, the condenser transmitter being exposed to tones of frequencies from 200 to 10,000 cycles per second under definite adjustments of the supply circuit of a receiver producing this tone. When the frequency is low, little reflection takes place from the transmitter standing alone, and bringing up the wall should cause a great increase in the response

of the system. At high frequencies the transmitter should reflect nearly as much alone as when part of a large wall, and the readings with and without the wall should be nearly equal. Plotting ratio of response without, to response with the wall was expected to yield a curve which could be used to make the final reduction of electrical output to incident sound energy, and so permit a more accurate determination of the spectrum of sound energy of the voice.

No consistent results were obtained after several trials and the experiment was abandoned. The failure is doubtless to be ascribed to standing waves, the character of which is very sensitive to the location in the room of the transmitter and the wall. This experiment is to be repeated under more favorable conditions when standing waves can be eliminated.

Thus, the curves finally obtained show no more than the frequency distribution of energy in speech in terms of the mechanical energy of a more or less ideal transmitter diaphragm. However, this information has its value because in any given configuration of transmitter, speaker, and room, there is a definite correspondence between the sound energy of the voice and the force acting on the diaphragm on which it falls, and in telephony at any rate it is this action on the diaphragm with which we are immediately concerned.

In conclusion we may give a determination of the total energy rate of speech, obtained as a by-product of the preceding investigation. Knowing the calibration of the system in absolute units, it is possible to determine the alternating pressure on the condenser transmitter diaphragm exposed to continuous speech from the normally modulated voice under the conditions of the experiment. Using the mean of the values obtained with 9 observers we find for the alternating pressure 11.3 dynes per sq. cm. (r.m.s.) for a distance of 2.5 cm. from mouth to diaphragm. This corresponds to an energy flow of 3.2 ergs per sq. cm. per second. Assuming that this energy flow is distributed uniformly over a hemisphere of 2.5 cm. radius, we may take 125 ergs per second as the total sound energy flow from the lips with the normally modulated voice.