Factors Governing the Intelligibility of Speech Sounds

N. R. French, and J. C. Steinberg

Citation: The Journal of the Acoustical Society of America **19**, 90 (1947); doi: 10.1121/1.1916407
View online: https://doi.org/10.1121/1.1916407
View Table of Contents: http://asa.scitation.org/toc/jas/19/1
Published by the Acoustical Society of America

response. The $2^{\frac{1}{3}}$ rule and the 2:3:5 proportions employed in practice satisfy this criterion, but the present result suggests that a relatively broad range of proportions may be equally satisfactory. A quantitative evaluation of response irregularity, involving damping and other factors neglected here, remains for further study.

The author is grateful to P. M. Morse for valuable advice and guidance, to H. Feshbach for helpful discussions and checking of the manuscript and to several colleagues for assistance as noted in the paper.

APPENDIX I

Probability Function for Random (Poisson) Distribution

Consider a random distribution in which every normal frequency μ_N in an interval Δ has an equal probability of occurring at any frequency in that interval. This is an example of Poisson distribution, for which it can be shown that

$$P_m = \frac{N^m e^{-N}}{m!}$$

is the probability that there are m normal frequencies in an interval in which $N = \Delta/\bar{\delta}$ is the expected (average) number.

The probability of finding *no* normal frequencies ($m=0$) in the interval Δ is

$$P_0 = e^{-N} = \exp(-\Delta/\bar{\delta}).$$

But if there are no normal frequencies in Δ then there is a space of width Δ. Therefore, $\exp(-\Delta/\bar{\delta})$ *is the probability of finding a space of width* Δ *where the average expected space width is* $\bar{\delta}$.

The probability of occurrence of an actual space δ is then $\exp(-\delta/\bar{\delta}) = e^{-\rho}$; the probability per unit average space is $(1/\bar{\delta})e^{-\rho}$; and the number per unit $\bar{\delta}$ in the range between δ and $\delta + d\delta$ is $(1/\bar{\delta})e^{-\rho}d\delta$. But $\rho = \delta/\bar{\delta}$ and $d\rho = d\delta/\bar{\delta}$, so $(1/\bar{\delta})e^{-\rho}d\delta = e^{-\rho}d\rho = M_\rho d\rho$, where: $M_\rho = e^{-\rho}$ is the probability of finding spacing ratios between ρ and $\rho + d\rho$, per unit average space $\bar{\delta}$, for Poisson distribution.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 19, NUMBER 1 JANUARY, 1947

Factors Governing the Intelligibility of Speech Sounds

N. R. French and J. C. Steinberg
Bell Telephone Laboratories, New York, New York

(Received November 22, 1946)

The characteristics of speech, hearing, and noise are discussed in relation to the recognition of speech sounds by the ear. It is shown that the intelligibility of these sounds is related to a quantity called articulation index which can be computed from the intensities of speech and unwanted sounds received by the ear, both as a function of frequency. Relationships developed for this purpose are presented. Results calculated from these relations are compared with the results of tests of the subjective effects on intelligibility of varying the intensity of the received speech, altering its normal intensity-frequency relations and adding noise.

1. INTRODUCTION

THIS paper discusses the factors which govern the intelligibility of speech sounds and presents relationships for expressing quantitatively, in terms of the fundamental characteristics of speech and hearing, the capability of the ear in recognizing these sounds. The relationships are based on studies of speech and hearing which have been carried on at Bell Telephone Laboratories over a number of years. The results of these studies have in large measure already been published. The formulation of the results into relationships for expressing speech intelligibility, which has also been in progress for a number of years, has not been previously published. The purpose of this paper is to bring the relationships and basic data together into one report.

Speech consists of a succession of sounds varying rapidly from instant to instant in intensity and frequency. Assuming that the various components are received by the ear in their initial order and spacing in time, the success of the listener in recognizing and interpreting these sounds depends upon their intensity in his ear and the intensity of unwanted sounds that may be present, both as a function of frequency. The relationships presented here deal with intelligibility as a function of these intensities. Relationships having the same objective were formulated about 25 years ago by H. Fletcher. While the present relationships are based largely on data not then available, their development has employed to a considerable extent the concepts of the earlier formulation.

Before proceeding with the subject matter of the paper a word concerning applications of the material may be in order. Material of this type has, of course, been of considerable service for many years in the Bell System. It has, for example, helped to guide the direction of development work on transmission instrumentalities and has aided the preparation of the quantitative transmission data used in engineering the telephone plant.[1] Other factors, however, in addition to those discussed here, often need to be considered in appraising the transmission performance of a speech communication system. For example, echoes, phase distortion, and reverberation may affect intelligibility.[2,3] The naturalness of the received speech may need consideration as a separate item. This is also true of loudness because speech may be too loud for comfort or so faint that the effort of concentrating on the sounds is excessively annoying, even though the sounds are intelligible.

In addition, there is usually the question whether some of the data used in applying the computational methods or, for that matter, in testing the transmission performance of speech communication systems in the laboratory, are truly representative of the conditions of actual use. In either case the value of the results depends upon the degree to which these conditions and the reactions of the users to them can be specified. This information is often difficult to obtain. It is desirable, therefore, in applying the results of computational methods or laboratory tests, to check any modifications of speech communication systems by testing them under actual service conditions and determining their effect on overall performance as judged by the users. The reasons for such a procedure are indicated briefly below and in more detail in a paper by W. H. Martin.[4]

The intensity of the speech received by the ear at each frequency depends on the intensity of the original speech sounds, the position of the mouth of the talker with respect to the microphone, the efficiency at each frequency of the latter in converting to electrical form the speech sounds which reach it, the transmission characteristics of the circuit intervening between the microphone and receiver, the efficiency of the latter in reconverting the speech waves to acoustical form and finally the coupling between the receiver and the ear. It is important to note that those items which are under the control of the user are subject to large variations. For example, there are large natural differences between the intensities of the same sounds spoken by different people or by the same people at different times. In addition, a person tends to adjust the output of his voice in part by the loudness with which he hears his own speech and the incoming speech, both being functions of the response characteristics of the communication system employed. Speech intensities also depend on the intensity of unwanted sounds, such as ambient noise in which the speaker may be immersed. These same factors also partly control the speaker's position with respect to the microphone and the way in which the listener holds the receiver to his ear.

Unwanted sounds in the ear have a masking effect on speech and constitute another major variable. They may arise from electrical disturbances originating within or without the communication system or from ambient noise. The latter may reach the ear by several paths: (1) by leakage between the receiver cap and the ear, or directly when loud speakers are used; (2) by being picked up by the microphone at the listening location and transmitted to the local receiver by sidetone; and (3) by transmission from the distant microphone.

Summarizing, it can be seen that the speech and noise received by a listener are the net result of a large number of factors of which several different types can be discerned: (1) the basic characteristics of speech and hearing, (2) the

[1] F. W. McKown and J. W. Emling, Bell Sys. Tech. J. 12, 331 (1933).

[2] V. O. Knudsen, J. Acous. Soc. Am. 1, 56 (1929).

[3] J. C. Steinberg, J. Acous. Soc. Am. 1, 121 (1929).

[4] W. H. Martin, Bell Sys. Tech. J. 10, 116 (1931).

electrical and acoustical characteristics of the instruments and circuits intervening between talker and listener, (3) the conditions under which communication takes place, and (4) the behavior of the talker and listener as modified by the characteristics of the communication system and by the conditions under which it is used.

By expressing the intelligibility relationships in terms of the intensities of speech and noise in the ear of the listener, the complicating factors discussed in the previous paragraphs do not appear explicitly in the relations. They appear only when the speech and noise intensities in the listener's ear are required in order to apply the relationships to the solution of a particular problem. There is also the question of the effect of variations in the acuity of hearing of the listeners. The relationships presented here apply specifically to young men and women who have good hearing but in general, as discussed later, their field of application is broader than this.

There are a set of consistent and well-defined concepts which underlie the intelligibility relationships. As these may be lost sight of in the details of formulation given in the succeeding pages, they are summarized briefly in the next section.

2. BASIC CONCEPTS

The intelligibility of the received speech sounds is related to a quantity which has been called the articulation index and designated A. It is a quantity such that increments ΔA carried by increments Δf of the speech frequency range may be added together to obtain the total A. The maximum possible value of A is assigned a value of unity; the minimum value is zero.

Any increment Δf of the speech frequency range may at best carry a maximum value of ΔA designated as ΔA_m. When conditions are not optimum for hearing speech in the increment Δf, this increment contributes only a fractional amount W of its maximum, or $\Delta A = W \cdot \Delta A_m$. For convenience in making computations, the frequency range may be divided into twenty bands whose frequency limits are so chosen that the ΔA_m of each band is 0.05, i.e., one-twentieth of the articulation index of the full band under optimum conditions. The general procedure for computing articulation index involves the determination of a value of W for each of the twenty bands, the addition of these twenty values of W and the division of this sum by twenty.

The particular value of W for any one band of speech depends upon a quantity E called the effective sensation level of the band in the ear of the listener, which is simply the sensation level of the band minus the total masking. The sensation level of a speech band is the attenuation needed to reduce the band to the threshold of hearing in the absence of noise and is determined from the intensities of the speech components within the band at the ear of the listener and the acuity of hearing. The total masking is the shift in threshold due to the presence of noise and is the resultant of three kinds of masking: (a) residual masking due to components of preceding speech sounds within the band, (b) interband masking due to speech components in adjacent bands, and (c) masking from extraneous noise components. The factor W is equal to the fraction of $\frac{1}{8}$th second time intervals in which the speech intensity in the particular band is of sufficient intensity to be heard. Stated differently, it is the fraction of these intervals in which the speech intensity in a band exceeds the intensity which corresponds to an effective sensation level of 0 db.

In this paper the relationship between ΔA_m and Δf is obtained empirically from the results of articulation tests on appropriate high pass and low pass filter systems. However, Mr. R. H. Galt has shown, in an unpublished memorandum, that this relationship can be derived from data on the differential pitch sensitivity of the ear. This suggests that the articulation index has a more fundamental significance than might be indicated by its empirical derivation.[5]

Although the response characteristics of a telephone system and its component parts do not enter explicitly into the articulation relationships, they are required in applications of the latter to particular problems. To serve the desired purpose the basic speech, noise, and hearing data and the over-all response of the telephone system must be so specified that they can be combined to obtain intensities received in the listener's ear. The type of response needed is obviously not one based

[5] W. A. Munson, J. Acous. Soc. Am. 17, 103A (1945).

alone on physical measurements of microphone, circuit, and receiver apart from voice and ear. It should include the effects of using real voices and ears. The methods of expressing response characteristics and the characteristics of speech and hearing which underlie the articulation index relationships, are essentially interdependent. Consequently, these subjects are discussed in the following two sections prior to the derivation and detailed discussion of the articulation index relations.

The following are the principal symbols used in this paper. A number of the symbols represent intensity levels; these are in db above 10^{-16} watt/cm^2.

A articulation index,

ΔA increment of articulation index carried by an increment Δf of the speech frequency range,

ΔA_m maximum possible value of ΔA,

W fractional part of ΔA_m obtained when listening conditions are not optimum,

S syllable articulation,

R over-all orthotelephonic response,

β intensity level of a single frequency tone,

β_0 threshold intensity level of a single frequency tone,

K $10 \log_{10}\Delta f_c$,

Δf_c width of critical bands of the ear in cycles,

X $(\beta_0 - K)$,

B the long average intensity per cycle level of the noise received from all sources,

B_f component of B produced in a particular frequency region by speech in the same region,

B_n component of B produced in a particular frequency region by speech in other frequency regions,

B_E component of B from all sources other than speech,

Z level above threshold of a critical band of noise, i.e., effective level,

M masking, i.e., shift of threshold caused by noise,

m $(M-Z)$ for values of Z greater than 50 db,

B_s' the long average intensity per cycle level of an idealized spectrum of speech at one meter from the lips (Fig. 2),

B_s the long average intensity per cycle level of the speech received over a communication system,

V the actual speech level, for any talker, at two inches from the lips, as measured with a sound level meter with 40-db weighting,

H level of a critical band of speech above its threshold level in the absence of noise, i.e., band sensation level,

E the effective sensation level of a band of speech, and

p difference in db between the intensity in a critical band exceeded by 1 percent of $\frac{1}{8}$th second intervals of received speech and the long average intensity in the same band.

3. CHARACTERISTICS OF SPEECH AND HEARING

3.1 The Spectrum of Speech

Figure 1 shows the results of several sets of measurements of the intensity of speech as a function of frequency. Curve A represents the average spectrum of four men and four women members of the testing crew used in carrying out the last extensive program of fundamental articulation tests. The spectrum is at a point two inches directly in front of the lips and is expressed in terms of the long time average intensity per cycle, in db relative to 10^{-16} watt/cm^2. Curves B and C are the spectra given for six men and five women in Fig. 10 of a paper by Dunn and White.[6] In the present paper the latter spectra have been shifted to change from 30 cm, the point of measurement, to the 2-inch position at which curve A applies. In order to provide a better basis for comparing shapes, the curve for the women has been shifted upward an additional 3 db because their total power was that much less than the men's.

It will be observed that there is an appreciable difference between the shapes of the Dunn and White spectra and the spectrum of the articulation testing crew. Because of the long intervai (several years) between the two sets of measurements, it has been impracticable to determine whether the differences are real or result from one or more of the numerous differences in the testing arrangements and procedures. In view of this and the substantial differences which may exist between the spectra of individual voices, the smoothed and somewhat arbitrary compromise

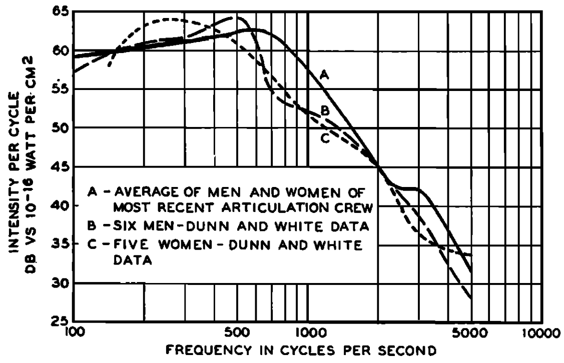

FIG. 1. Comparison of speech spectra at two inches from lips.

[6] H. K. Dunn and S. D. White, J. Acous. Soc. Am. 11, 278 (1940).

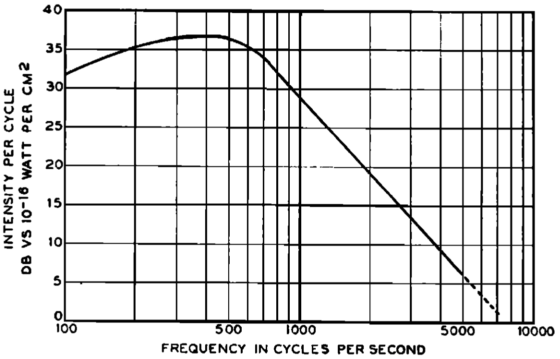

FIG. 2. Idealized long average speech spectrum at one meter from lips in a sound field free from reflections.

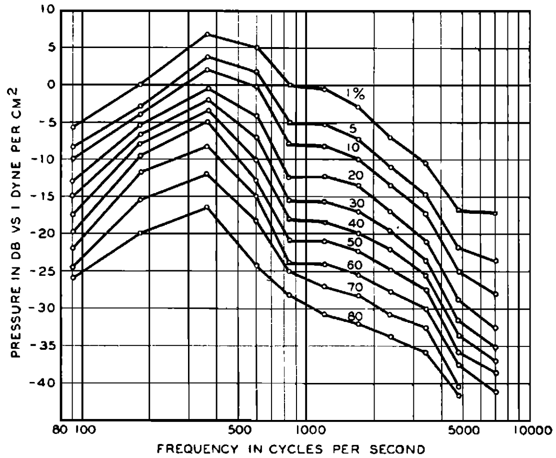

FIG. 3. R.m.s. pressure, during one-eighth second intervals, of speech at 30 cm from lips. Dunn and White composite data for six men (reference 6). Each curve shows the pressure exceeded in the indicated percentage of intervals.

spectrum of Fig. 2 has been adopted for use in this paper. For reasons which will appear later, this spectrum is given at a distance of one meter from the lips. The intensity of this spectrum, integrated over the entire frequency range, amounts to 65 db relative to 10^{-16} watt/cm^2. The corresponding figure at 2 inches from the lips, which is a more accurate point of measurement, is 90 db. If the speech level of a speaker having this idealized spectrum were measured by a sound level meter,[7] using flat weighting, with the microphone at 2 inches from the lips, the observed level would be about 3 db higher than the integrated value or around 93 db. This difference would occur because readings of rapidly varying material tend to be taken on the frequent peaks. With 40 db weighting the observed level should be close to the integrated level or 90 db.

3.2 Level Distribution of Speech

The spectra of speech which have just been discussed represent the average intensity over an appreciable period of time. From moment to moment the intensity of speech fluctuates rapidly above and below this average curve giving rise, at any frequency, to a level distribution of speech as a function of time. This distribution is one of the factors affecting the intelligibility of speech and consequently enters into the relationships presented later. In Fig. 3, taken largely from Fig. 3 of the previously mentioned paper[6] by Dunn and White, are shown the results of level distribution measurements made on a number of male voices. The same paper shows a similar set of data for women's voices. These charts show the distribution of $\frac{1}{8}$ second intervals (roughly the duration of a syllable) with respect to the r.m.s. pressure measured during these intervals in the frequency bands indicated along the abscissa. The differences between levels which are exceeded by 1 percent and 50 percent of the intervals in the bands are shown in Table I for both the men and women talkers. It can be inferred from this table that the range over which the speech intensity fluctuates and the relative occurrence of intervals of different intensities are roughly the same for all bands and for both men and women. Taking all the bands to be alike in these respects results in certain simplifications of the relationships which are presented.

To determine the actual form of the speech level distribution, the data taken with male voices and the 1000–1400 cycle band have been used.

TABLE I. Difference in db between r.m.s. pressures of speech exceeded in 1 percent and in 50 percent of one-eighth second intervals.

Frequency band	Men's voices	Women's voices
250– 500	12 db	15 db
500– 700	18	18
700–1000	21	21
1000–1400	20	21
1400–2000	19	21
2000–2800	18	20
2800–4000	18	20

[7] "ASA—American Standard—Sound Level Meters for Measurement of Noise and Other Sounds" (Z24.3—1944) July 28, 1944.

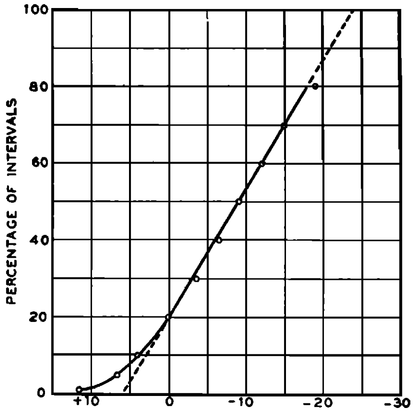

FIG. 4. Cumulative level distribution of average intensity of speech in one-eighth second intervals in db *versus* long average intensity. 1000–1400 cycle band of men's voices.

The first step was to compute the long average intensity by integrating over all of the $\frac{1}{8}$ second intervals in this band. Then the level difference between this long average intensity and the average intensity which was exceeded in 1 percent of the intervals was determined. The value of 1 percent was then plotted against this level difference to determine the point at the lower left corner of Fig. 4. The other points in the figure were obtained by the same process, using the levels exceeded in 5 percent, 10 percent, etc., of the intervals. It will be seen from the resulting curve that 1 percent of the intervals have average intensities 12 db or more above the long average intensity. It will be noted further that over the range between the 20 percent and the 80 percent points of Fig. 4, the distribution can be closely represented by a straight line. Although no accurate data are available to show the shape of the curve above the 80 percent point, it will be advantageous for reasons discussed later to assume that the distribution continues as a straight line up to 100 percent as shown by the dotted line.

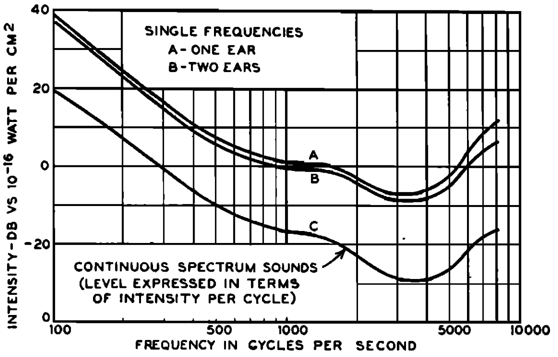

FIG. 5. Zero loudness contours for open air borne sounds.

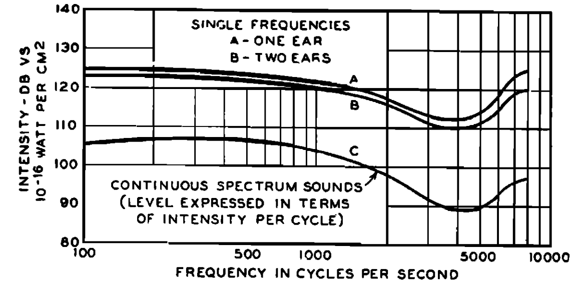

FIG. 6. 120-db loudness contours for open air borne sounds.

If the same procedure is followed for the other bands and for women's voices it will be found that the resulting curves are similar to the curve of Fig. 4, although they tend to be somewhat steeper in slope. On the other hand it would be desirable for the purpose of this paper to measure the level distribution with bands approximating the critical band widths of the ear (Section 3.4), which are narrower bands than those used in the above measurements. This would cause some reduction in the slope of the curves. Figure 4 appears to be a reasonable compromise between these two offsetting factors. In the development of simple relationships it will be convenient and reasonably accurate to use the single curve of this figure as applying to all frequency regions.

3.3 Zero and 120 db Loudness Contours

Curves *A* and *B* of Fig. 5 show the thresholds of audibility for single frequency tones when listening with one and two ears. These curves apply to the most acute ears and indicate about the absolute minimum of sound that can be heard. The two ear curve is identical with the zero loudness contour of the "American Standard for Noise Measurement."[8] The one ear curve is the two ear curve increased by the curve of Fig. 9. In communicating by speech many of the sounds, both wanted and unwanted, tend to approach the continuous spectrum type instead of being discrete frequencies. Under these conditions the application of the single frequency threshold

[8] "ASA—American Standard for Noise Measurement" (Z24.2—1942) J. Acous. Soc. Am. **13**, 102 (1942).

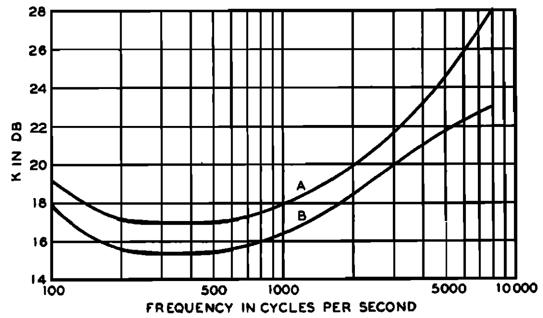

Fig. 7. Critical band widths (K) of ear. In db, $K=10\ \log_{10}W_c$, where W_c is the width of a critical band in cycles. Curves A and B are, respectively, for one- and two-ear listening.

curves requires the specification of a band width over which the intensity of the continuous spectrum sound is integrated. This is discussed later. For the present it is sufficient to note that curve C of Fig. 5 may be considered as a threshold curve for sound of the continuous spectrum type, when its level is expressed in terms of the intensity per cycle.

Curve B of Fig. 6 shows the two ear 120-db loudness contour for single frequency tones taken from the same source as curve B of Fig. 5. Curve A of Fig. 6 for one ear listening was obtained by adding to curve B of the same figure the curve of Fig. 9. The significance of the 120-db loudness contours lies in the fact that more intense sounds lying in the region above these curves are apt to annoy the listener, produce a sensation of feeling or, if of sufficiently high level, produce an actual sensation of pain. Curve C applies to sounds of the continuous spectrum type. This figure is of interest primarily in situations where there is extremely intense noise at the receiving position and higher than normal levels of received speech are required for the attainment of adequate intelligibility.

When there are no unwanted sounds in the ear the practical limits within which the wanted sounds should lie are bounded by the region just above the 120-db loudness contour and the threshold of audibility. These curves apply to the case where the sound waves arrive from a source at some distance from the observer who faces the source in a place free of reverberation. The intensities are measured with the observer out of the sound field, but at the position he takes in listening. Thus they do not represent the intensities which actually exist in the ear except over the lower part of the frequency range. To use these curves in applications in which the listening is done with head receivers, it is necessary to express the output of the receiver in terms of the intensity of open airborne sounds which produce the same sensation as the sounds from the receiver.

3.4 Masking

In most problems involving speech reception, unwanted sounds are present in the ear of the listener and reduce the sensitivity of the ear to other sounds. This reduction in sensitivity is known as masking and at any frequency the amount of the masking M is equal to the difference between the levels β and β_0 of a single frequency tone which are just audible in the presence of the noise and in the total absence of noise, or

$$M=\beta-\beta_0. \qquad (1)$$

The plot of M as a function of frequency is known as the masking spectrum of the noise. In general, interfering noises in the ear of a listener are of the continuous spectrum type, such as room noise. The masking relations provide a means for computing the masking caused by noise of this type when its spectrum affecting the ear is known. The amount of masking which is given by these relations is the threshold shift which would be observed by a highly idealized group of individuals whose thresholds β_0 in the absence of noise are given by the curves of Fig. 5. Actually, the threshold varies greatly among individuals depending upon such factors as fatigue, health, and age, the chosen curves representing about the absolute minimum of sound that can be heard by the most acute ears. The formula will thus, in general, compute a masking figure which is somewhat larger than would be observed by a random crew of observers. This, however, will usually be of no practical importance because computed levels of wanted sounds, above the same threshold, will be too large by the same amount. The margin of the wanted sounds above the unwanted ones is largely independent of the absolute threshold of the observer provided the noise is above the actual threshold. Observed tone levels which can just be heard in the presence of appreciable amounts of noise should be in good agreement with computed tone levels obtained by adding computed maskings to the idealized threshold curve, regardless, within fairly large limits, of the absolute thresholds of the observers.

Tests have shown that the masking effect on single frequency tones of noises having continuous spectra, which do not change in intensity too rapidly with frequency, is dependent only upon the level difference in db between (1) the intensity of the noise integrated over a narrow frequency band whose frequency limits are somewhat below and above the frequency of the masked tone, and (2) the single frequency threshold intensity in the absence of noise. These narrow bands are known as critical bands[9] of the ear and the above level difference at any frequency is referred to as the effective level of the noise at that frequency. The width of the critical bands is a function of frequency, varying from about 30 cycles at low frequencies to several hundred cycles at high frequencies.

The level difference in db between the noise intensity integrated over a critical band and the single frequency threshold (β_0) is given by

$$Z=(B+K)-\beta_0=B-(\beta_0-K), \qquad (2)$$

where

Z = level above threshold of a critical band of noise, i.e., effective level,

B = the long average intensity per cycle level of the noise received from all sources, expressed in db above 10^{-16} watt/cm^2,

$K=10\log_{10}\Delta f_c$, where Δf_c is the critical band width in cycles.

The values of K for one and two ear listening, as derived from masking tests, are shown by Fig. 7. The above expression for effective level is equivalent to referring the noise B to a new threshold which is K db lower than the single frequency threshold. Thus, instead of always being obliged to add a quantity K to the noise spectrum, it will be more convenient, where the noise spectrum is expressed in terms of the intensity per cycle, to subtract from B a new threshold X where

$$X=\beta_0-K \qquad (3)$$

and then

$$Z=B-X. \qquad (4)$$

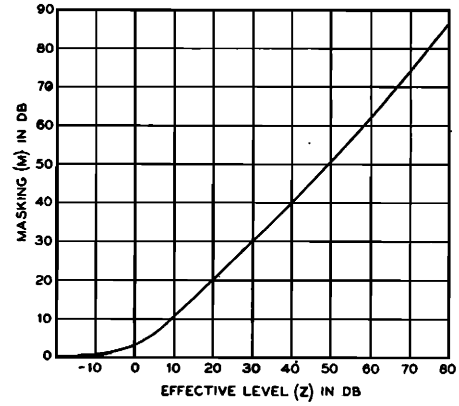

FIG. 8. Relation between the effective level of noise in any frequency region and the resulting masking in the same region.

The value of X is shown by the bottom curve of Fig. 5. It may be noted that the differences between the one- and two-ear single frequency threshold curves in Fig. 5 are identical with the differences between the one- and two-ear K's of Fig. 7. As a result, a single value of X applies to both one- and two-ear listening.

When the value of the effective level Z is known, the amount of masking M that is produced can be read from the curve* of Fig. 8. As a matter of interest, it will be noted that the masking and the effective levels are equal over the range of 20 to 50 db masking. Within this range a tone can just be heard through a steady noise when the intensity of the former is equal to the intensity of the noise integrated over the critical band in the region of the tone. However, as the effective level in a band increases above 50 db the resulting masking increases at a somewhat faster rate. The tests which gave this result used noises covering a broad frequency range as they generally do in communication problems. This upturn in the masking curve under such conditions has a bearing on some practical problems. For example, consider a case where the only important noise affecting a listener is transmitted along with a signal and the absolute level of reproduction can be varied. Under these conditions, where the signal-to-noise ratio remains constant, the signal may not be heard as well at an intense level of reproduction as at some lower

[9] H. Fletcher and W. A. Munson, J. Acous. Soc. Am. 9, 1 (1937).

* The values of K and M *vs.* Z of the present report differ slightly from those given in reference 9 as a result of additional experimental data.

level. The effect of noise at low levels is also worthy of note. Figure 8 shows that some masking is produced by noise even though it is below the threshold of audibility (Z less than zero db). This is exactly the effect which would be obtained if the threshold in the absence of noise were itself determined by a residual noise, which combines on a power basis with other noises which may be present. The form of the masking curve over its entire range is given by

$$M=(B(+)X)-X+m, \qquad (5)$$

where (+) represents power addition of the quantities B and X, and m is the amount, in db, that the masking exceeds the effective level of the noise. Values of m for effective levels greater than 50 db are given in Table II; for values of Z less than 50 db, m is zero.

At this point it may be of interest to indicate the reasons why the differences between the one- and two-ear thresholds and the one- and two-ear K's are taken to be alike. Figure 9(A) shows observed differences in the acuity of hearing of the best ear and the average of both ears, taken from Fig. 20 of a paper[10] by Fletcher and Munson. The effect of one- *vs.* two-ear listening on K was determined by adjusting the levels of single frequency tones until they could just be heard in the presence of a noise of the continuous spectrum type. This was done alternately with one- and two-ear listening, while maintaining the same noise level for both conditions. These tests showed that higher tone levels relative to the noise levels were required when listening with one ear as compared to two. These differences, which will be shown to represent the differences in K for the two conditions are indicated by Fig. 9(B). It will be seen that a single curve represents these masking data and also the audibility data of Fig. 9(A). That the differences in tone levels for one- and two-ear listening represent the differences in their K's can be shown by noting that the level of a tone which can just be heard in the presence of noise is, from Eq. (1), given by

$$\beta=M+\beta_0.$$

From Eq. (2) the effective level of the noise is

$$Z=B+K-\beta_0.$$

Also, for the levels used in the above tests, the masking M is numerically equal to the effective level of the noise; thus Z in the second equation can be substituted for M of the first equation, which can then be written

$$\beta=B+K. \qquad (6)$$

In this equation B is the intensity level per cycle of the noise and β the intensity level of the tone which can just be heard. It follows that, if the tone level β is greater with one ear listening than with two, the value of K must increase by the same amount since B was constant for the two conditions.

TABLE II. Values of m to the nearest db.

Z in db	m in db	Z in db	m in db
54–60	1	78–80	6
61–65	2	81–83	7
66–70	3	84–86	8
71–74	4	87–89	9
75–77	5	90–91	10

4. RESPONSE CHARACTERISTICS

An over-all response which has been called "orthotelephonic[11] response" is used for applying the information of the preceding section to the derivation of the intensity of speech received over a communication system. This response may be thought of as a usage response, in that it includes the effects on the received speech of distance and coupling between the microphone and the speaker's mouth and the coupling of the receiver to the ear. By definition a telephone system has an orthotelephonic response of zero db at all frequencies when it can be replaced by a one-meter air path, between talker and listener, without changing the loudness of the received speech at any frequency. The speaker and listener face each other in an otherwise unobstructed sound field. Listening to the sound over the air path is done with either one or two ears, depending upon whether one or both ears are used with the communication system.

A telephone system having the above characteristics is designated as an orthotelephonic system. It is convenient to specify the output of such

[10] H. Fletcher and W. A. Munson, J. Acous. Soc. Am. **5**, 82 (1933).

[11] A. H. Inglis, Bell Sys. Tech. J. **17**, 358 (1938).

a system, at any frequency, in terms of the intensity, at the same frequency, of the speech received over the air system, but measured before insertion of the listener's head into the sound field. As a result, the speech received over an orthotelephonic system is identical to that received over the air system in loudness and intensity when the talker speaks at the same level in both cases. Specifying the intensity of the received speech in this manner is in conformity with the manner of expressing the zero and 120-db loudness contours discussed in the previous section.

If a telephone system, which is not an orthotelephonic system, has an orthotelephonic response of R db at any frequency, this means that the speech received over an orthotelephonic system, at the same frequency, must be raised R db in intensity to be as loud as that heard over the telephone system in question. Thus the intensity level of speech received over a telephone system at any frequency is the sum of the intensity level of speech at one meter from the lips and the orthotelephonic response of the system.

In general, a person will talk at a different level than that corresponding to the idealized spectrum of Fig. 2. Correction for this can be made by raising the spectrum by an amount $V-90$, where

V = the actual speech level, for any talker, at two inches from the lips in db *vs.* 10^{-16} watt/cm^2, as measured with a sound level meter using 40 db weighting,

90 = the corresponding level for the idealized speech spectrum of Fig. 2 when shifted to the two inch point.

The above information can be combined into the following equation for computing the intensity levels of received speech:

$$B_s = B_s' + (V-90) + R, \qquad (7)$$

where

B_s = the long average intensity per cycle level of the speech received over a communication system, expressed in db *vs.* 10^{-16} watt/cm^2,

B_s' = the long average intensity per cycle level of an idealized spectrum of speech (Fig. 2) at one meter from the lips in a place free of reverberation, expressed in db *vs.* 10^{-16} watt/cm^2.

The intensity level of the received speech B_s is, of course, in terms of the free field intensity which produces, in the uncovered ear of an observer placed in the sound field, the same sensation obtained with speech delivered by a telephone receiver. It is equally important to note that the intensity level of received noise B, discussed previously, is also expressed in the same terms.

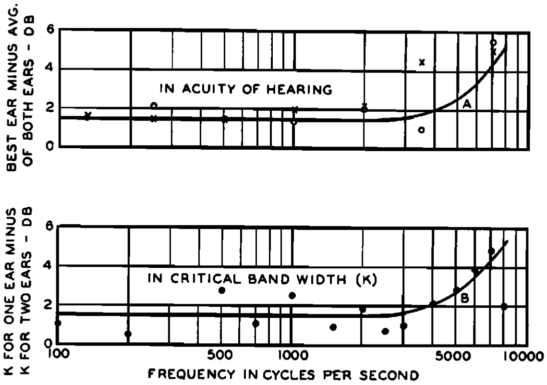

FIG. 9. Differences between one-ear and two-ear listening.

In concluding this section it may be in order to bring out some of the practical aspects of the problem of obtaining the orthotelephonic response of a telephone system. The over-all response is not usually measured as a whole in accordance with the above description but is derived from separate measurements of the response of microphone, electrical circuit, and receiver. The circuit responses are derived from purely physical measurements using single frequency tones. The real ear responses of receivers are also determined with single frequency tones by balancing the tone heard in the receiver against a comparison tone of the same frequency transmitted over a one-meter air path. The intensity of the output of the receiver is specified, exactly as described above, in terms of the intensity of the comparison tone, measured in the free sound field. The input to the receiver is measured in any suitable terms which will combine properly with the measurements of circuit response up to the receiver. Receiver measurements made in this way are usually accompanied by purely objective measurements, using mechanical couplers for example, from which conversion factors are obtained which do away with the need for further real ear measurements on other receivers of the same type.

The real voice response of a linear microphone can be obtained from two sets of measurements. In one, a person speaks into the microphone, taking whatever position with respect to it that is regarded as typical, while measurements are made of the output of the microphone in narrow frequency bands throughout the entire frequency range. In the other, a similar analysis is made of the speech intensity near the lips of the speaker, usually at two inches, the microphone having been removed from the sound field. These latter speech intensities, or these intensities reduced to one meter from the lips, are taken as the input to the microphone. Supplementing these measurements by objective response measurements using, for example, single frequency tones and an artificial voice, provides conversion factors which enable real voice responses of other microphones of the same type to be derived from purely objective measurements. These conversion factors allow for the interaction effects and distance losses between the artificial source and microphone relative to these effects between a real voice and the microphone. The application of this method without modification, to non-linear microphones, can give results which may be somewhat in error due to modulation products, generated by the microphone when complex waves of speech are impressed upon it. This may be avoided by a more complicated procedure beyond the scope of this paper to describe. It is also beyond the scope of the paper to go into details concerning the responses needed for determining the levels of noise in the ear. It should be sufficient to point out that the basic noise data and the response of each separate path by which noise can enter the ear should be so coordinated and expressed that they can be combined to give the noise intensity in the ear in the same terms as the received speech.

5. ARTICULATION INDEX

5.1 General

A distinguishing characteristic of speech is movement. Conversation at the rate of 200 words per minute, corresponding to about four syllables and ten speech sounds per second, is not unusual. During the brief period that a sound lasts, the intensity builds up rapidly, remains comparatively constant for a while, then decays rapidly. The various sounds differ from each other in their build-up and decay characteristics, in length, in total intensity, and in the distribution of the intensity with frequency. With the vowel sounds the intensity is carried largely by the harmonics of the fundamental frequency of the voice and tends to be concentrated in one or more distinct frequency regions, each sound having its own characteristic regions of prominence. The consonant sounds, as a group, have components of higher frequency and lower intensity than the vowel sounds. In addition, the intensity tends to be scattered continuously over the frequency region characteristic of each sound. Thus when the elementary sounds are combined in sequence to form syllables, words, and phrases, there is a continuous succession of rapid variations in intensity, not only in particular frequency regions but also along the frequency scale. The interpretation of speech received by the ear depends upon the perception and recognition of these constantly shifting patterns.

The importance of the different regions of intensity and frequency to the recognition process was determined, in the investigation described here, by articulation tests, using a test circuit into which electrical networks and different amounts of attenuation were introduced to alter the intensity-frequency distribution and level of the called material prior to its reception by the listeners. The material consisted of meaningless monosyllables of the consonant-vowel-consonant type. The results were expressed as the percentage of syllables of which all three component sounds were perceived correctly. This percentage is designated as the syllable articulation, or simply the articulation, of the condition tested. The sounds used in these syllables include those commonly used in conversation.[12] A detailed description of this method, including the reasons for its choice, is given in other papers.[13,14]

Syllable articulation, in common with all other known subjective measures of intelligibility such as word or sentence intelligibility, has certain limitations which impair its usefulness as a basic index. First, the value obtained from tests is not independent of the skill and experience of the testers. This difficulty can be partially overcome by calibrating a crew and correcting the results

[12] N. R. French, C. W. Carter, Jr., and W. Koenig, Jr., Bell Sys. Tech. J. **9**, 290 (1930).

[13] H. Fletcher and J. C. Steinberg, Bell Sys. Tech. J. **8**, 806 (1929).

[14] T. G. Castner and C. W. Carter, Jr., Bell Sys. Tech. J. **12**, 347 (1933).

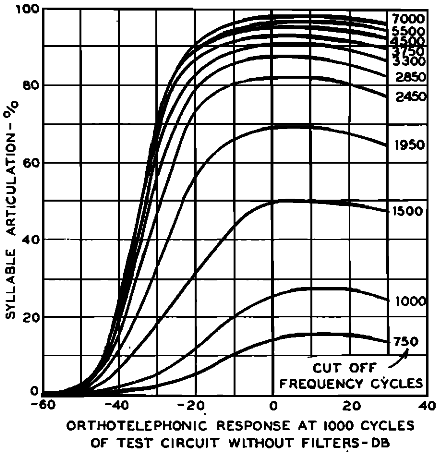

FIG. 10. Smoothed results of 1928–1929 articulation tests on low pass filters having the indicated cut-off frequencies.

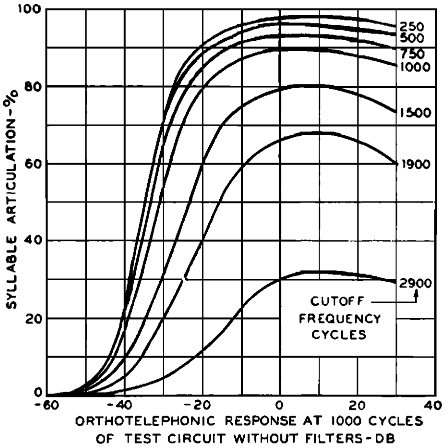

FIG. 11. Smoothed results of 1928–1929 articulation tests on high pass filters having the indicated cut-off frequencies.

by methods described elsewhere.[13] Of more importance is the fact that syllable articulation, in common with other subjective measures, is not an additive measure of the importance of the contributions made by the speech components in the different frequency regions. Stated differently, the articulation observed with a given frequency band of speech is not equal to the sum of the articulations observed when the given band is subdivided into narrower bands which are then individually tested. For the purpose of establishing relations between the intelligence carrying capacity of the components of speech and their frequency and intensity, a more fundamental index free of the above defects is needed. Such an index, called "articulation index," can be derived from the results of articulation tests. The magnitude of this index is taken to vary between zero and unity, the former applying when the received speech is completely unintelligible, the latter to the condition of best intelligibility.

The articulation index is based on the concept that any narrow band of speech frequencies of a given intensity carries a contribution to the total index which is independent* of the other bands with which it is associated and that the total contribution of all bands is the sum of the contributions of the separate bands. Letting ΔA represent the articulation index of any narrow band of speech frequencies and n the number of narrow bands into which the total band is subdivided for computational purposes, the articulation index A of the total band reaching the listener is

$$A = \textstyle\sum_1^n \Delta A. \tag{8}$$

The value of ΔA, which is carried by any narrow frequency band, varies all the way from zero to a maximum value ΔA_m as the absolute levels of speech and noise in the ear are independently varied over wide ranges. Letting W represent the fractional part of ΔA_m which is contributed by a band with a particular combination of speech and noise, the value of articulation index for that band is given by

$$\Delta A = W \cdot \Delta A_m. \tag{9}$$

Hence,

$$A = \textstyle\sum_1^n W \cdot \Delta A_m. \tag{10}$$

The establishment of relations for computing A thus involves two main steps: (1) the determination of the increments of frequency which give equal values of ΔA_m throughout the frequency range and (2) the determination of relationships between W and the levels of speech and noise in the ear.

The desired relations are derived below from the results of articulation tests on a broad-band transmission system into which high pass and low pass filters were inserted. The system included distortionless attenuators and amplifiers for varying the absolute level of the received speech.

* Not absolutely true; the contribution of a band may be modified somewhat by masking produced by intense speech in neighboring bands.

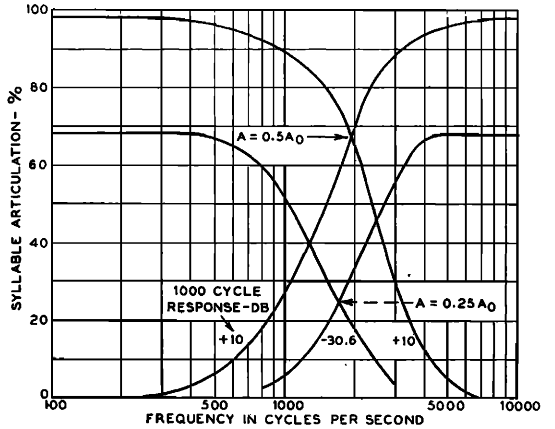

FIG. 12. Syllable articulation *versus* cut-off frequency of high pass and low pass filters at two different settings of test circuit. A_0 is articulation index of test circuit at its optimum setting.

The orthotelephonic response of the system with particular settings of attenuators and amplifiers is shown by curve A of Fig. 28. The departures of this response from flatness largely reflect usage factors and the method adopted for specifying the receiver output, which were discussed earlier in the paper. For example an imperfect seal between a receiver and the ear provides a shunting leakage path to the outside air and causes the drop in response noted at low frequencies.

The results of a few articulation tests with the frequency band limited by filters appear in a previous paper[3] by one of the writers. The smoothed results of more comprehensive tests, which provide the basis for the following relations, are given by Figs. 10 and 11. The former applies to low pass and the latter to high pass filters. The results are composite data taken with men's and women's voices. The ordinate of the curves represents the percentage of syllables which were recorded correctly. The abscissa is the orthotelephonic response at 1000 cycles of the test circuit before insertion of the filters. The filters introduced a negligible loss within their passed bands and also caused practically complete suppression of the speech components beyond their cut-off frequencies. Thus the abscissa of Figs. 10 and 11, in combination with the response Curve A of Fig. 28 and the cut-off frequency of the filters, permits the determination of the response, at all frequencies, of the test condition corresponding to any value of articulation shown on these figures.

During each articulation test electrical measurements were made of the total speech output of the microphone. Computations were also made to determine what the output of the microphone would be with a talker having the speech spectrum of Fig. 2. By comparing these results it is estimated that this particular articulation testing crew talked at an acoustic level 4 db higher than that to which Fig. 2 applies.

5.2 Relation between ΔA_m and Frequency

Referring now to the curves of Figs. 10 and 11, it will be noted that articulation rises rapidly as the circuit response is varied to raise the level of the received speech and reaches a maximum value at about the same setting of the system with each of the filters. The 1000-cycle orthotelephonic response of the system at this generally optimum setting is +10 db. The articulation values indicated for the different filters at this setting, plotted against the filter cut-off frequencies, are shown by the top pair of curves of Fig. 12. Now letting S_1 represent the indicated value of syllable articulation when the frequency range below a certain cut-off frequency is transmitted and S_2 the syllable articulation when the range above the same cut-off frequency is transmitted, it will be noted that the sum of S_1 and S_2 is generally greater than the articulation S_3 observed when both bands are transmitted together. In other words, the articulation of 27 percent for a 1000-cycle low pass filter, when added to 89 percent for the complementary high pass filter, does not yield the value of 98 percent which was observed for the full band. It follows, therefore, that syllable articulation is not an additive index. This is also true for observed values of letter articulation, word articulation and sentence intelligibility. However, the curves of Figs. 10 and 11 offer a means of deriving an additive index from the articulation test data, as described below.

Since the full-band system which was used may not be an optimum system for articulation, the articulation index of the speech received over it, which is presumably close to but not necessarily equal to unity at the optimum level, is here designated A_0. For this value of articulation index the syllable articulation is that observed for the full-band at a setting of +10 db, or $S=98$ percent. It will be noted also from Fig. 12 that the high pass and low pass filter curves for this +10-db condition intersect at about 1900 cycles; this means that for this particular system half of the articulation index carried by the full-band of received speech is below and half above this frequency. At the point of intersection the observed value of S was 68 percent and consequently a syllable articulation of 68 percent for this particular testing group corresponds to an articulation index of $0.5A_0$.

If the top curve of Fig. 10 is now referred to, this curve applying to a 7000-cycle low pass filter and also to the unrestricted band, it will be noted that, by increasing the attenuation of the system, the S of the full band can be reduced to 68 percent which, as previously noted, corresponds to an articulation index of $0.5A_0$. The 1000-cycle response of the system at which this occurs is -30.6 db. If the syllable articulation obtained with the different filters at this setting of the system is now plotted against the cut-off frequency of the filters, another pair of intersecting curves will be obtained as shown in the lower part of Fig. 12. The articulation index of each of the two complementary bands, below and above the frequency of intersection (1700 cycles), consequently has by definition an articulation index of $0.25A_0$ and the corresponding value of S is 25 percent. This procedure may be followed further to find that a syllable articulation of 8 percent corresponds to an articulation index of $0.125A_0$.

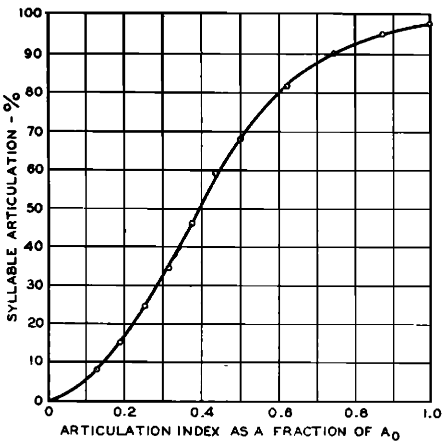

FIG. 13. Relation between syllable articulation and articulation index. The latter is expressed as a fraction of the articulation index (A_0) of the test circuit at its optimum setting.

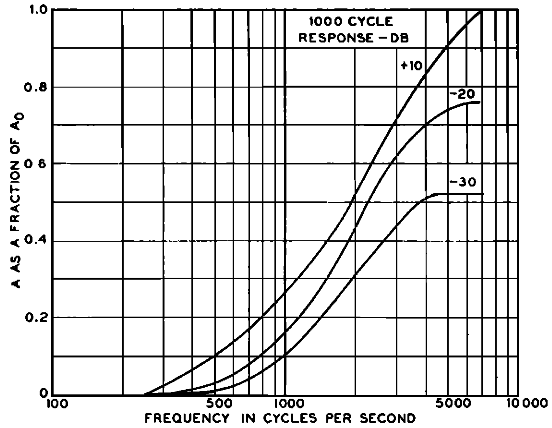

FIG. 14. Relation between articulation index and cut-off frequency at three different settings of the test circuit. Articulation index is expressed as a fraction of the articulation index (A_0) of test circuit at its optimum setting.

Knowing from the above that a syllable articulation of 25 percent corresponds to an articulation index of $0.25A_0$ reference is again made to the +10 db curves of Fig. 12. It will be seen that a low pass filter (about 950 cycles) yielding an articulation index of $0.25A_0$ has as its complement a high pass filter having a syllable articulation of 90 percent. Since the contributions of these two complementary filters must add to A_0 it follows that $S=90$ percent corresponds to an articulation index of $0.75A_0$. By following these procedures a sufficient number of points may be found to determine satisfactorily the curve shown in Fig. 13. This curve shows the relationship between syllable articulation and articulation index expressed as a fraction of the articulation index A_0 of the full-band of the speech received at its optimum level over the system which was tested.

Having obtained the relationship shown in Fig. 13, it is now possible to construct a set of curves showing, for each of several levels of the full-band of speech, the cumulative total of articulation index, expressed as a fraction of A_0, as the upper end of the passed band is increased in frequency. This is accomplished by reading from Fig. 10 the syllable articulation values obtained with all the filters at each of several fixed settings of the full band system, converting these values of S into fractional values of A_0 by means of the curve of Fig. 13 and plotting the

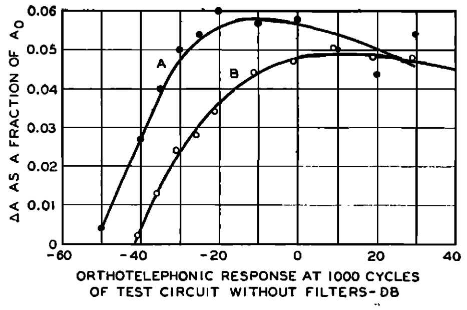

FIG. 15. Fractional values of A_0, the articulation index of the test circuit at its optimum setting, carried by individual bands. Curve *A*—band from 1300 to 1520 cycles. Curve *B*—band from 490 to 620 cycles.

results against the cut-off frequency of the filters. The results of this operation are shown in Fig. 14.

The next step is to separate the frequency range into a large number of bands (20 were used) having equal fractional values of A_0. The +10-db curve of Fig. 14 is used for this purpose since this is the optimum setting of this system with the full band. Having established by this method the frequency limits of bands of equal importance (.05A_0) *in the system tested*, the contribution of these bands at other levels can be read from a complete family of curves like those of Fig. 14. The resulting values are then plotted against the orthotelephonic response of the system at 1000 cycles* to obtain, for each of the twenty bands, curves of the type illustrated by Fig. 15. These curves show that the increment ΔA, carried by a band, first increases as the gain of the system is increased, then reaches a maximum value after which it drops off slowly as the gain is further increased. If the system tested had been an optimum system, the maximum contribution of each of the twenty bands should be .05, since the frequency limits of the bands were selected on the basis of a 5 percent contribution by each band at the optimum setting of the system. Also the maximum contribution of each band should occur at the same setting of the system. Inspection of the curves of Fig. 15 shows that neither of these expectations is precisely fulfilled, thus indicating that the testing system fell somewhat short of being an optimum system. Actually, a summation of the maximum values of ΔA_0 of the twenty curves gives a value about 3 percent above unity. If a value of unity is assigned to the articulation index of the speech received over an optimum system, this means that the speech received over the system tested had an articulation index of 0.97, or $A_0=0.97$. With this information the curve of Fig. 13, showing the relation of syllable articulation to articulation index as a fraction of A_0, can be converted to a relation between syllable articulation and absolute values of articulation index by multiplying the abscissa by 0.97. The resulting curve is shown on Fig. 23. Although this curve may be lacking in general interest the detailed description of how it was obtained is of general interest since the same method could be used by others who might start with a system having different response characteristics from that used in the tests which have been described.

We are now in a position to draw up a cumulative curve of the absolute value of articulation index *versus* frequency when all bands are simultaneously at their optimum settings. The maximum value of articulation index which can be contributed by each of the twenty bands discussed above is obtained by multiplying the maximum value of each of the twenty curves, like those of Fig. 15, by 0.97. The resulting value for the band of lowest frequency, plotted against the upper frequency limit for this band, provides one point on the desired curve. By adding successive bands, one at a time, the final relation, shown by the curve of Fig. 16 is obtained. It differs only slightly from the top curve of Fig. 14 which

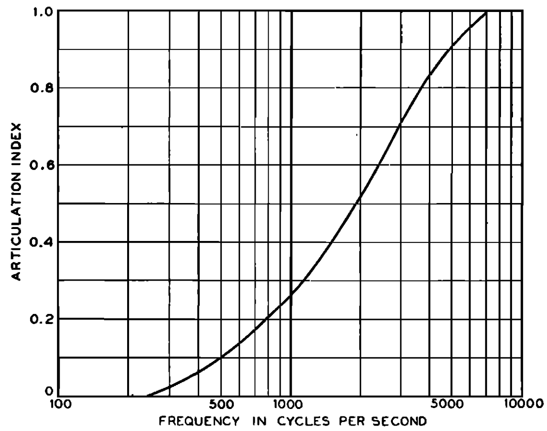

FIG. 16. Articulation index *versus* cut-off frequency. All bands are at their optimum levels. Curve is based on about equal numbers of men's and women's voices.

* Any other parameter which reflects changes in received level, such as the response of the system within each of the 20 bands, could be used equally well.

applied when the system used was tested at its optimum setting. In fact, the differences between the two curves are so small that it is open to question whether the data are sufficiently precise to justify the above operation in this case. It is believed, however, that the operation will be of interest in the event of additional basic studies of this nature which, caused by the particular characteristics of the circuits which may be employed, may require greater corrections.

The derivative or slope of the curve of Fig. 16 at any frequency shows the importance of that frequency with respect to its maximum possible contribution to articulation index. At any frequency the product of the slope of this curve and the factor *W*, discussed in the next section, represents the contribution of this frequency to the total articulation index. In general, the levels of speech and noise in the ear, and hence *W*, will vary sufficiently slowly with frequency to permit the use of a single value of *W* over a considerable frequency range. For the general run of computations twenty values of *W* at suitably selected frequencies should be adequate. For this purpose, it is convenient to divide the frequency range into twenty parts or computation bands such that the maximum possible contribution of each band is equal to that of the others and to determine *W* at the mid-frequency of each band. The limits of the twenty bands chosen in this way are obtained by reading from the continuous curve of Fig. 16 the frequencies corresponding to all the articulation indices which are multiples of .05. These band limits are given in Table III.

The importance curve of Fig. 16 is based on composite data taken with about equal numbers of men and women talkers. Men's voices are about an octave lower in pitch than women's and the latter tend to be somewhat richer in high frequency sounds. As a result it is probable that separate importance curves for men's and women's voices would approximate the curve of Fig. 16 in shape but be shifted somewhat toward lower and higher frequencies, respectively.

TABLE III. Frequency bands making equal (5 percent) contributions to articulation index when all bands are at their optimum levels. Composite data for men's and women's voices.

Band	Frequency limits cycles	Band	Frequency limits cycles
1	250–375	11	1930–2140
2	375–505	12	2140–2355
3	505–645	13	2355–2600
4	645–795	14	2600–2900
5	795–955	15	2900–3255
6	955–1130	16	3255–3680
7	1130–1315	17	3680–4200
8	1315–1515	18	4200–4860
9	1515–1720	19	4860–5720
10	1720–1930	20	5720–7000

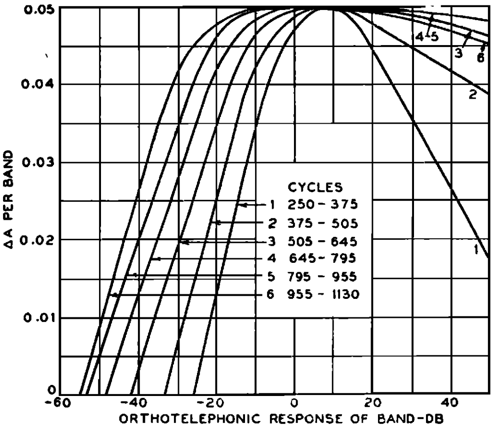

FIG. 17. Effect of level variations on articulation index carried by narrow bands. Band limits are so chosen that the articulation index of each band is 0.05 at its optimum level.

5.3 Variation of ΔA with Level

Having obtained the frequency limits of the twenty bands which individually contribute 0.05 to articulation index when each band is making its maximum possible contribution, the next step is the determination of the contribution of each band under other than optimum conditions. This includes the specification of the conditions in usable terms. The starting point is the twenty curves illustrated by the two curves of Fig. 15. The ordinates of these curves are first multiplied by 0.97 to convert to absolute values of articulation index, as discussed previously. They are then used to draw up additional curves of cumulative articulation index *vs.* frequency, similar to the curve of Fig. 16 but for levels 10, 20, 30, etc., db below the optimum level of each band, or above the reasonably well-defined settings at which the contribution of the individual bands drops to zero. After smoothing these curves, one for each relative level, they are divided into bands having the frequency limits of Table III. The contribution of each of these bands at each level is then obtained from the new set of curves (not shown)

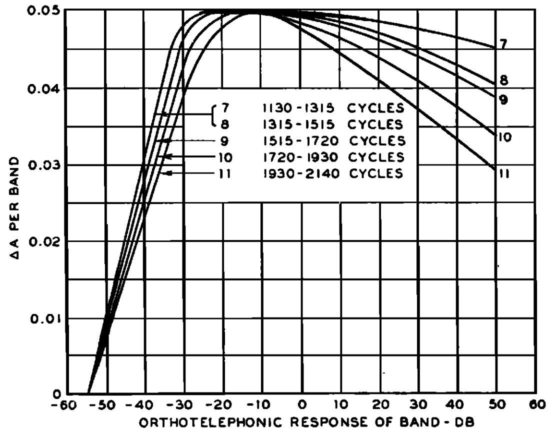

FIG. 18. Effect of level variations on articulation index carried by narrow bands. Band limits are so chosen that the articulation index of each band is 0.05 at its optimum level.

and plotted to obtain the twenty curves of Figs. 17–20, inclusive. These show the effect of level changes on the articulation index carried by each of the twenty equally important bands. The abscissa of each curve is the orthotelephonic response of the transmission system at the mid-frequency of the particular band. The specification of the response of each band in this way was accomplished by shifting the abscissa of the curves illustrated by Fig. 15 by the amount that the orthotelephonic response at the band frequencies exceeds the orthotelephonic response at 1000 cycles, Curve *A* of Fig. 28 providing the necessary data.

The absolute placement of the curves applies only to the particular acoustic talking level used in the basic articulation tests. However, the curves can obviously be specified on an absolute basis, if desired, in terms of the absolute intensity of the received speech, by adding to the abscissa the intensity, in each band, of the crew's speech at one meter. Such a group of curves could be used for computational purposes. A different procedure is followed, however, to obtain a solution which will not only more readily handle problems involving noise but also more clearly bring out the nature of the relationships.

The fraction of the maximum possible contribution which a band makes when it is not at an optimum level is designated by *W*. Curves of *W* against level would consequently be identical in shape to the curves of Figs. 17–20. It will be noted that these curves are essentially straight lines except in the region where the articulation index is approaching a maximum and that the slopes of the straight line portions are approximately alike and equal to about 3 db for a change of 10 percent in *W*. This is the same slope that was derived earlier for the level distribution of speech in narrow bands (Fig. 4).

When speech, which is constantly fluctuating in intensity, is reproduced at a sufficiently low level only the occasional portions of highest intensity will be heard, but if the level of reproduction is raised sufficiently even the portions of lowest intensity will become audible. Thus the similarity in slope of the straight line portions of the *W* curves and the speech distribution curve suggests that *W is equal to the fraction of the intervals of speech in a band that can be heard.* It will be noted, of course, that the shapes of the *W* and speech curves are different in the region where *W* is approaching zero. Actually the *W* curves in this region cannot be determined accurately and probably do taper off in much the same manner as the speech level distribution (low portion of curve of Fig. 4).

As regards the upper part of the curves of Figs. 17–20 it will be seen that their shapes in this region do not agree with the speech level distribution of Fig. 4. As pointed out previously the latter was extrapolated in the 80–100 percent region and consequently may be in error. For reasons which will be pointed out later, it appears advantageous to assume that the straight line of

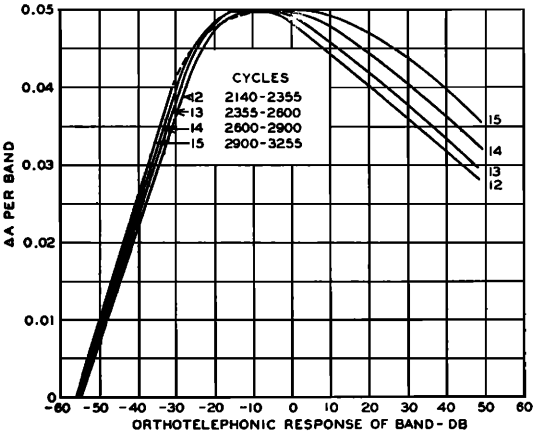

FIG. 19. Effect of level variations on articulation index carried by narrow bands. Band limits are so chosen that the articulation index of each band is 0.05 at its optimum level.

Fig. 4 represents the true speech level distribution up to 100 percent (i.e., down to the lowest level intervals) and to offer a different explanation for the bending over of the ΔA curves, and hence of W, as the level of maximum contribution is approached. This is taken up later as the explanation to be given involves a consideration of the effects of noise.

If W is equal to the fraction of the time intervals that speech in a critical band can be heard, it should be possible to derive W from the characteristics of speech and hearing and to use Figs. 17–20 for testing the method. The first step in this process is the definition of a new term H, where

H = the level of a critical band of speech above its threshold level in the absence of noise. This is termed the band sensation level.

The band sensation level of speech is given by

$$H=B_S+p+K-\beta_0=B_S+p-X. \quad (11)$$

The terms B_S, K, β_0, and X have been defined previously. The term p is the difference in db between the intensity in a critical band exceeded by 1 percent of $\frac{1}{8}$th second intervals of received speech and the long average intensity in the same band.

Tests have shown that speech does not become inaudible until its long average intensity per cycle is reduced to about 30 db below the single frequency threshold β_0. This results from two causes which bring about the introduction of the p and K terms in the above equation. Since speech is far from constant in intensity its threshold level in any frequency region is determined by the most intense sounds in that region. As pointed out previously, the intensity of these sounds integrated over $\frac{1}{8}$th second time intervals and over frequency bands which approximate the critical bands in width, is about 12 db above the long average intensity within the same bands; hence $p=12$ db. Actually this difference varies somewhat from band to band and in the direction of smaller values at low frequencies. In the interests of simplicity it is here considered to be independent of frequency.

The need for the K term, which is of the order of 20 db, has already been pointed out in connection with the discussion of masking of continuous spectra sounds. While speech is not rigorously of this type, the spacing of its single frequency components, which are constantly varying up and down the frequency scale, corresponds roughly to the width of the critical bands over which the intensity has to be integrated to obtain a true measure of the sensation which is produced. Therefore, without much loss of accuracy, the same values of K and hence X, which have been determined for sounds having continuous spectra can be applied to speech.

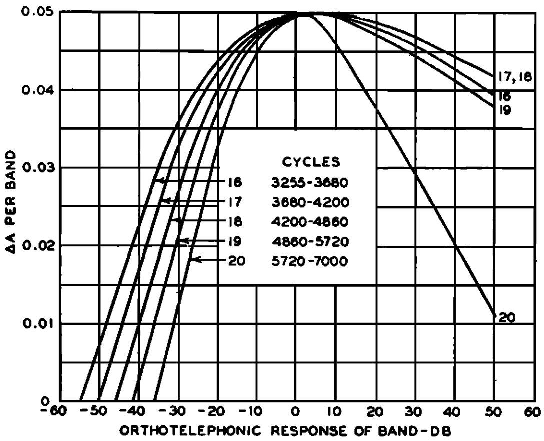

FIG. 20. Effect of level variations on articulation index carried by narrow bands. Band limits are so chosen that the articulation index of each band is 0.05 at its optimum level.

Now referring back to Figs. 4 and 17–20 it will be appreciated that there are certain consequences that can be tested if the hypothesis is correct that W is equal to the proportion of the intervals of speech in a band which can be heard. These are

(1) The computed sensation levels of the speech received in the 20 bands should be substantially alike when these bands all have the same value of W.

(2) The computed sensation level in each band for the zero point of the twenty ΔA curves, which are drawn down to the zero point as straight lines, should be 6 db. This results from the shape of the speech level distribution (Fig. 4) and the choice of the 1 percent highest intervals for expressing the sensation level of the speech in a band.

The sensation level corresponding to $W=0$ is desired for each of the twenty frequency bands of Figs. 17–20. Although these bands are wider than the critical bands their sensation levels are nevertheless given correctly by Eq. (11). This equation involves B_S which in turn is given by

TABLE IV. Computation of the sensation level (H) of the received speech at which $W=0$ in the 1928–1929 articulation tests. For the particular crew, $H=B_S'+16+R-X$.

Band	B_S' (db)	R (db)	X (db)	H (db)
1	36.5	−26	−1.5	28
2	36.6	−33.5	−8.0	27
3	35.7	−42	−11.6	21
4	33.4	−48.5	−14.1	15
5	30.7	−53.5	−15.7	9
6	28.3	−55	−16.7	6
7	26.0	−55	−17.5	5
8	24.0	−55	−18.3	3
9	22.1	−55	−19.4	2
10	20.4	−55	−21.0	2
11	18.9	−55	−23.3	3
12	17.5	−55.5	−25.2	3
13	16.1	−56	−26.6	3
14	14.6	−56	−27.8	2
15	13.0	−56	−28.5	2
16	11.3	−55.5	−28.9	1
17	9.5	−50.5	−28.8	4
18	7.5	−46	−27.8	5
19	5.1	−41.5	−25.1	5
20	2.5	−36	−19.7	2

TABLE V. Values of W for values of E between 0 and +12 db.

E in db	W	E in db	W
1.0–2.2	.01	8.4–8.7	.11
2.3–3.1	.02	8.8–9.1	.12
3.2–3.9	.03	9.2–9.5	.13
4.0–4.6	.04	9.6–9.9	.14
4.7–5.3	.05	10.0–10.3	.15
5.4–6.0	.06	10.4–10.7	.16
6.1–6.6	.07	10.8–11.1	.17
6.7–7.2	.08	11.2–11.5	.18
7.3–7.8	.09	11.6–11.8	.19
7.9–8.3	.10	11.9–12.1	.20

Eq. (7). Combining Eqs. (7) and (11) we obtain

$$H=B_S'+(V-90)+p+R-X.$$

The term $(V-90)$ represents the acoustic talking level of the particular articulation test crew relative to the talking level corresponding to the idealized spectrum of Fig. 2; hence $(V-90)$ is +4 db as mentioned previously. Also p is +12 db as discussed above. Combining these numerical values, the values of H for this particular crew are given by:

$$H=B_S'+16+R-X.$$

This equation has been applied to the computation of H for each of the twenty bands whose frequency limits are given on Figs. 17–20. The values of B_S' and X, at the mid-frequencies of the bands, were taken from Fig. 2 and curve C of Fig. 5. The values of the orthotelephonic response R of the circuit were read from the abscissa of Figs. 17–20 at the points of zero contribution of the twenty bands. The results of the computations are given in the last column of Table IV. For bands 5 to 20, inclusive, the computed levels are all within a range of 8 db and the average level for these bands is within $2\frac{1}{2}$ db of the required value of 6 db. In view of the many sources of error, involving the measurement of the acoustic level of the talkers, the real voice and ear calibrations of microphone and receivers and possible differences in the manner in which the latter were talked into and held to the ear in the calibrating and articulation tests, the results for bands 5–20 are considered to be in reasonable agreement with the requirements which are being tested.

The levels computed for bands 1 to 5, inclusive, are too high. However, they are qualitatively in agreement with what would be expected if there had been a low level of room noise in the listening booth during the tests, resulting in masking of the speech. Since room noise usually falls off rapidly with increasing frequency and the shielding effect of receivers held against the ear increases with frequency, extraneous low level noise would have its greatest masking effect in the lowest bands, and negligible effects above 1000 cycles or so. One of several possible sources of noise is the movements of the four observers who were in the booth at the same time. Another uncertainty at the lower frequencies lies in the manner in which the receiver is held to the ear. The above computation of absolute levels in the ear involves the real ear response of the receiver, and consequently the tacit assumption that the coupling between receivers and ears in the articulation tests was the same as in the subjective determinations of the receiver response. Here again any differences which may exist between the responses in the two cases are likely to be greatest at low frequencies. In view of these various effects it is believed that the computed absolute levels are sufficiently close to those required by the above hypothesis of the significance of the W factor, to justify it as a working basis in the formulation of a method for computing the articulation index of received speech.

5.4 Derivation of W—Noisy Conditions

It is apparent that values of W over the range from 0 to about 0.7 can be determined closely, for speech reproduced over linear systems and listened to under quiet conditions, by computing the fractional part of the speech distribution of Fig. 4 which is above threshold. When more than about 70 percent of the speech distribution is above threshold in the absence of noise, an additional factor is included to account for the rounded portion of the ΔA curves of Figs. 17–20, covering values of W in the range from about 0.7 to unity. This part of the curves can be arrived at on the basis of a fatigue effect which may be considered as self-masking. On this basis the hearing of the relatively infrequent low level sounds in a band is considered to be impaired through a temporary loss of sensitivity owing to the preceding sounds of higher level in the same band. This loss of sensitivity will be treated as equivalent to the effect of noise. It is necessary, therefore, to develop relations for noisy conditions before the development for quiet condition can be completed.

If there were no such loss of sensitivity and no other source of masking, and if the speech level distribution is taken to be a straight line, the value of W for any speech band would be given by the fraction of the speech intervals which have sensation levels above 6 db, or

$$W=(H-6)/30 \qquad (12)$$

for sensation levels between 6 and 36 db. To provide a basis for accounting for the gradual tapering off of the twenty ΔA curves as $W=1$ is approached, and also for evaluating the effects of noise generally, this equation will be rewritten as follows:

$$W=(E-6)/30, \qquad (13)$$

where E is a new term called the effective sensation level of a band of speech, given by the following equation:

$$E=H-M, \qquad (14)$$

where M is the masking resulting from all sources of interference, including the masking of speech on itself. By application of Eq. (11) this becomes

$$E=(B_S+p-X)-M. \qquad (15)$$

This can be written in the following more convenient form for computations by replacing M by its equivalent from Eq. (5), or

$$E=B_S+p-m-(B(+)X). \qquad (16)$$

To obtain W, this expression is substituted in Eq. (13), and

$$W=1/30[B_S+p-6-m-(B(+)X)]. \qquad (17)$$

This is the equation ordinarily used for computing W. Actually it is an approximation for values of W less than 0.2 (effective sensation levels less than 12 db). In cases where reception is poor and the effective sensation levels of the

TABLE VI. Values of β_0, X, K, and B_S' at selected frequencies. Values of K are in db; other quantities are in db *vs.* 10^{-16} watt/cm².

Bands for which ΔA_{max} =0.05	Band center cycles	One ear β_0	One ear K	Two ears β_0	Two ears K	X	$B_{S'}$
1	310	15.5	17.0	14.0	15.5	− 1.5	36.5
2	440	9.0	17.0	7.5	15.5	− 8.0	36.6
3	575	5.5	17.1	4.0	15.6	−11.6	35.7
4	720	3.3	17.4	1.8	15.9	−14.1	33.4
5	875	2.0	17.7	0.5	16.2	−15.7	30.7
6	1040	1.4	18.1	−0.1	16.6	−16.7	28.3
7	1225	.9	18.4	−0.6	16.9	−17.5	26.0
8	1415	.5	18.8	−1.0	17.3	−18.3	24.0
9	1615	− .2	19.2	−1.7	17.7	−19.4	22.1
10	1825	−1.4	19.6	−2.9	18.1	−21.0	20.4
11	2035	−3.3	20.0	−4.8	18.5	−23.3	18.9
12	2250	−4.8	20.4	−6.3	18.9	−25.2	17.5
13	2475	−5.9	20.7	−7.4	19.2	−26.6	16.1
14	2750	−6.6	21.2	−8.2	19.6	−27.8	14.6
15	3080	−6.9	21.6	−8.5	20.0	−28.5	13.0
16	3470	−6.7	22.2	−8.6	20.3	−28.9	11.3
17	3940	−5.9	22.9	−8.0	20.8	−28.8	9.5
18	4530	−4.1	23.7	−6.5	21.3	−27.8	7.5
19	5300	−0.3	24.8	−3.3	21.8	−25.1	5.1
20	6350	+6.5	26.2	+2.6	22.3	−19.7	2.5

TABLE VII. Values of $(B(+)X)-X$ as a function of $B-X$.

$B-X$ (db)	$(B(+)X)-X$ (db)
−9	0.5
8	0.6
7	0.8
6	1.0
5	1.2
4	1.5
3	1.8
2	2.1
−1	2.5
0	3.0
+1	3.5
2	4.1
3	4.8
4	5.5
5	6.2
6	7.0
7	7.8
8	8.6
+9	9.5

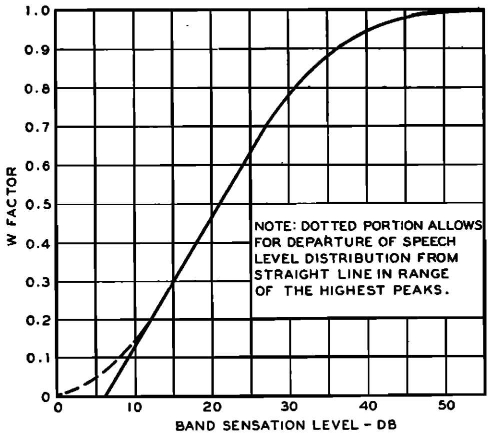

Fig. 21. The W factor for quiet conditions.

speech in a number of computation bands are less than 12 db, computations can be improved in accuracy by using Eq. (16) and Table V for these particular bands to allow for the departure of the most intense part of the speech level distribution from a straight line (Fig. 4). Values of m are given in Table II for values of Z above 50 db; at lower levels m is zero. Values of B_S', from which B_S is derived, and of X are given by Table VI. Values of $(B(+)X)$ relative to X are given by Table VII as a function of $(B-X)$. Outside the range for which values are given, $(B(+)X)$ equals either B or X, depending upon which is the larger.

In the above equations B represents the level above 10^{-16} watt/cm^2 of the combined intensity per cycle of all the various noises reaching the ear at any particular frequency. In addition to the usual sources of noise, B includes the noise equivalent in its effect to the self-masking of a band of speech on itself and also the noise equivalent in its effect to the masking of one speech band on another. These are all combined on a power basis and the sum then expressed in db. Self-masking and interband masking are further considered in the following sections.

5.5 Derivation of W—Quiet Conditions

It is now possible to consider self-masking and its effects on the form of the ΔA curves. Referring back to Eq. (13) it will be seen that if W is not to exceed unity the effective sensation level must not exceed 36 db. This is accomplished by taking the equivalent noise of self-masking as 24 db below B_S, where B_S is the long average intensity per cycle level of speech. This equivalent noise, designated by B_f, is

$$B_f = B_S - 24. \qquad (18)$$

Substituting this value of B_f for B in Eq. (17) it follows that, for quiet conditions,

$$W = 1/30[B_S + p - 6 - m - ((B_S - 24)(+)X)]. \qquad (19)$$

The relationship between W, as computed by this equation, and the sensation level of a speech band, as computed by Eq. (11), is shown by the continuous curve of Fig. 21. This curve applies to the case where there is no noise and no non-linear elements are between the voice and the ear to change the form of the time variation of speech received in a band from that of the original speech. If this curve is compared with the twenty curves of Figs. 17–20, it will be seen that it is a reasonable representation of their shapes over the entire range below their maximum values. It may be worth noting here that the self-masking factor, which produces the tapering effect as W approaches unity under quiet conditions, will also produce the same sort of an effect when other noises are present, and when the speech is raised to a level considerably above the noise level.

Figure 21 indicates that the maximum contribution of a band of speech under quiet conditions, except for the equivalent noise of self-masking B_f, is reached at a band sensation level of 50 db. At this speech level the effective level Z of the noise having an intensity per cycle level of B_f is only 14 db since B_f is 24 db below B_S and the sensation level of a speech band is determined by the levels of speech $p = 12$ db above B_S. The value of m in Eq. (19) is consequently zero over the range of the curve of Fig. 21.

Referring now to Fig. 8, it will be noted that masking does not start to increase faster than the effective level of noise until the latter exceeds 50 db. Consequently, when self-masking is the only source of masking, the value of m in Eq. (19) does not change from zero until the sensation level of speech in a band rises above 86 db. It follows that W for quiet conditions, as given by the above equations, has a value of unity for band sensation levels of speech ranging between 50 and 86 db. Thus the relations do not account for the reduction of A at high speech levels as shown on Figs. 17–20. Overloading in the ear, resulting in the generation of intermodulation products, which could act as noise, is a possible explanation of this reduction. If the noise equivalents of these products could be determined their effects could be allowed for, presumably, in the same manner as other noises. It is possible that the downward droop of the curves of Figs. 17–20 above the optimum levels is excessive. Because of the small variation of the measured values of articulation with level above the optimum level with the various filters, the derivation of the variation of the contributions of the individual bands with level in this region is not at all precise.

The unimportance of m, as discussed above, applies specifically to quiet conditions. In problems involving high levels of extraneous noise the inclusion of m in the above equations may have a considerable effect on W.

5.6 Interband Masking of Speech

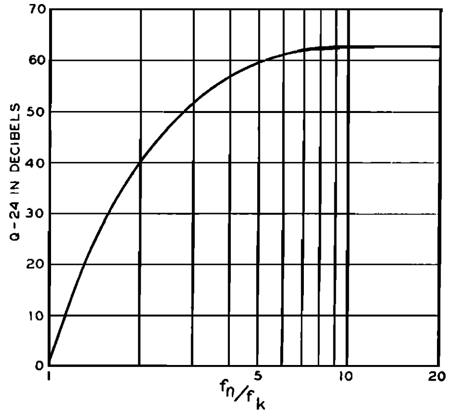

FIG. 22. Function used in determining the masking of speech by speech in lower bands.

Articulation tests have shown that at high received levels the articulation tends to decrease as the level is increased. The effect is most evident in systems that contain pronounced peaks or frequency regions that are partially suppressed. The effect is believed to be caused in part by speech in one frequency region masking the speech sounds in other frequency regions. One rather elaborate method for allowing for this effect has been developed but is too lengthy to describe here and also too laborious in applications involving many computations. This method involves, for example, the determination of the effect of each band of speech on each of the other bands. These effects are functions of the levels in each of the bands. In the computational method described here a simpler but presumably less accurate procedure has been followed. One simplification is to consider the effect of a speech band on only those bands which are of higher frequency. As in the case of self-masking, it will be convenient to consider the interband masking as equivalent to the masking produced by a noise B_n in the speech band being masked. The estimated intensity level of this equivalent noise B_{nk}

TABLE VIII. Q, i.e., the number of db that the noise produced in any band, by speech in any lower band, is below the long average intensity of the speech in the lower band.

Producing band	Band in which the noise is produced																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	49	62	70	75	78	81	83	85	85	85	86	86	87	87	87	87	87	87	87
2		44	56	64	70	74	78	80	82	83	84	85	85	86	86	87	87	87	87
3			42	52	61	67	71	75	78	79	81	83	84	85	85	86	86	87	87
4				40	50	57	64	68	71	75	78	79	81	82	83	84	85	86	87
5					38	49	56	62	66	70	72	75	78	79	81	83	85	85	86
6						36	46	54	59	64	67	70	72	75	77	80	82	84	85
7							35	45	51	56	61	65	69	72	75	77	79	81	84
8								35	43	50	55	59	64	67	71	74	77	80	83
9									35	42	47	52	57	63	67	71	74	78	81
10										34	41	47	52	57	63	67	71	76	79
11											33	40	46	52	57	63	68	72	77
12												33	40	47	54	60	65	71	75
13													34	41	50	55	61	67	72
14														34	42	49	56	64	70
15															35	43	51	59	65
16																35	45	54	61
17																	35	46	56
18																		36	49
19																			39

produced in band n by speech in band k is given by:

$$B_{nk}=B_{sk}-Q, \qquad (20)$$

where

B_{sk} = the intensity level of speech in band k which is doing the masking,
f_k = the mid-frequency of band k, and
f_n = the mid-frequency of the band n in which speech is being masked.

The quantity Q, derived empirically, is given on Fig. 22 as a function of (f_n/f_k). Values of Q for the particular frequency bands of Table III are given in Table VIII. To simplify the computations Q is here taken to be independent of the absolute level of B_{sk}.

Assuming the equivalent noises from the various bands to combine on a power basis, the total equivalent noise in band n produced by speech from all lower bands is given by

$$B_n=B_{n1}(+)B_{n2}(+)\cdots(+)B_{n,\,n-1}. \qquad (21)$$

In cases where very high levels of speech are necessary to ride over excessive levels of noise, and the response of the communication system contains sharp peaks or dips, interband masking may be appreciably larger than these formulas indicate.

5.7 Summary of Relationships—Linear Systems

If the speech frequency range is subdivided, for computational purposes, into twenty bands having the frequency limits of Table III, the value of ΔA_m for each band is 0.05 and the articulation index of the received speech by Eq. (10) is

$$A=0.05(W_1+W_2+\cdots W_{20}). \qquad (10a)$$

The subscripts refer to the individual bands of Table III. The value of W in any particular computation band is determined by the following relation in which the quantities that vary with frequency are usually specified at the mid-frequencies of the bands

$$W=1/30[B_S+p-6-m-(B(+)X)]. \qquad (17)$$

The symbol $(+)$, between two terms expressed in db, indicates that they are to be combined on a power basis and then reconverted to db. This is the basic equation for determining W except for non-linear systems, discussed in Section 5.8, or in cases where reception is poor and the effective sensation level (Eq. (16)) of the speech in a number of the computation bands is less than 12 db. In this event Eq. (16) should be used for these particular bands and the values of W read from Table V.

The quantity B_S is the level, in db *vs.* 10^{-16} watt/cm^2, of the long average intensity per cycle of the received speech, the intensity being expressed as a free field intensity in the manner described in Section 4. B is a similar quantity but applies to the total noise per cycle received from all sources. The value of p is ordinarily taken as 12 db at all frequencies. X is a function of frequency only; its values are given in Table VI. Values of $(B(+)X)$ relative to X are given in Table VII as a function of $B-X$. The term m can be omitted unless extraneous noise of high level is present; values of m as a function of the effective level Z of the noise B are given in Table II, where

$$Z=B-X. \qquad (4)$$

The value of B_S in Eq. (17) is given by

$$B_S=B_S'+(V-90)+R, \qquad (7)$$

where B_S' is the intensity level, at the appropriate frequency, of the idealized speech spectrum of Fig. 2, values of which are tabulated in Table VI. The symbol V represents the actual speech level of any particular talkers, at two inches from the lips, as determined by a sound level measurement using 40-db weighting. R is the orthotelephonic response of the communication system at the appropriate frequency.

The value of B in Eqs. (4) and (17) is given by a new equation

$$B=B_E(+)B_f(+)B_n, \qquad (22)$$

where B represents the intensity per cycle level of the total noise from all sources except that produced by the received speech, and

$$B_f=B_S-24 \qquad (18)$$

and

$$B_n=B_{n1}(+)B_{n2}(+)\cdots B_{n,\,n-1}, \qquad (21)$$

where n is the number of the particular band in which the noise is being determined and the subscripts 1, 2, etc., refer to the bands, one to

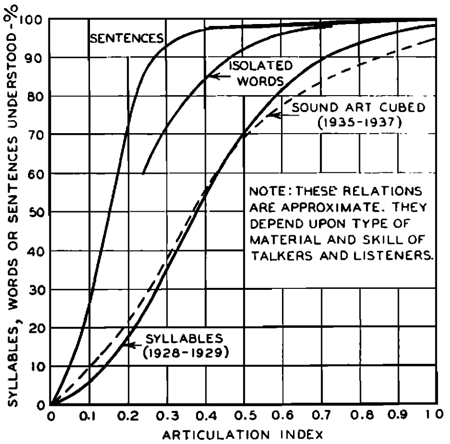

FIG. 23. Approximate relations between articulation index and subjective measures of intelligibility.

$(n-1)$, from which the noise arises because of the speech in these bands. The values of B_{n1}, B_{n2}, etc., relative to the levels B_S of speech in bands $1, 2\cdots(n-1)$, can be read from Table VIII.

In applying these relations it should be noted that the equivalent noises B_f and B_n vary with the level of the received speech B_S. B_n can usually be omitted entirely unless the response of the communication system falls off rapidly with increasing frequency or has sharp peaks and valleys.

5.8 Non-Linear Relation between Original and Received Speech

The above derivation of the W factor applies to cases where the intensity of the received speech in any band is proportional to the initial speech. It is now necessary to consider whether the same relations which specify the W factor in such cases will hold for cases where the speech is transmitted through systems containing a non-linear element, such as a carbon transmitter. Tests have shown that for a given value of received talking volume the articulation obtained with a carbon transmitter may be somewhat less than that obtained with a linear transmitter which has the same shape of frequency response characteristic. Attempts have been made to explain this effect by considering as noise the resulting inter-modulation products of speech. While there probably is such an effect, the reduction in articulation can also be accounted for by self-masking in conjunction with the effect of the non-linear device in altering the level distribution of the speech sounds. In some cases the output of a carbon transmitter changes r db for each db change in input, where r is nearly constant over a considerable range of levels and is usually greater than unity. It follows that the level distribution of the output of such a transmitter in a speech band would cover a range which is broader by the factor r than that of the original speech. Consequently on the basis of the self-masking theory, a greater fraction of the lower level intervals of speech would be masked by speech of higher levels in the band, thus reducing the maximum possible value of W.

The effect of such an expanding action where the output-input characteristic on a db basis is approximately linear, but with a slope different from unity, can consequently be computed from the relations which have already been given, by considering the basic speech level distribution to be r times as broad as that shown in Fig. 4 and then proceeding with the computations exactly as if the instrument were a linear one. The relationship between effective sensation level and the W factor has already been given, as follows:

$$W=(E-6)/30. \qquad (13)$$

In this equation the number 30 represents the range between the maximum and minimum levels of speech in a band, assuming a straight line

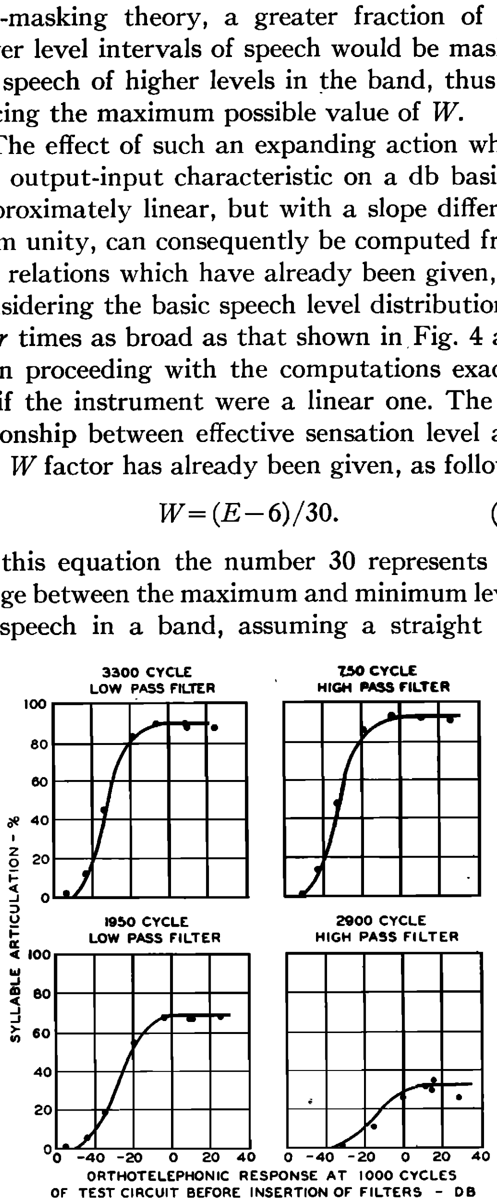

FIG. 24. Comparison of observed and computed results for 1928–1929 articulation tests. Points show observed data.

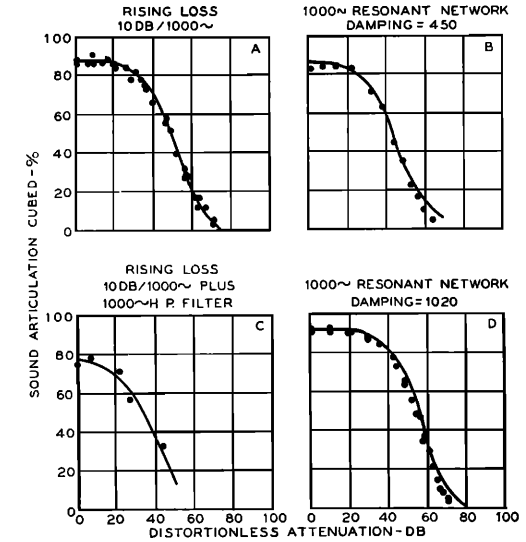

FIG. 25. Comparison of observed and computed results for 1935–1936 articulation tests. Points show observed data.

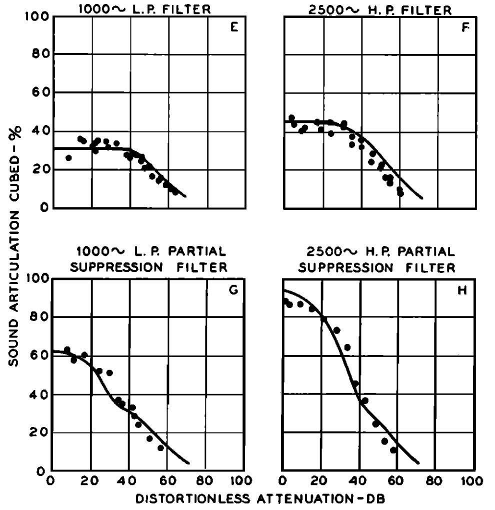

FIG. 26. Comparison of observed and computed results for 1935–1936 articulation tests. Points show observed data.

relationship between percentage of intervals and levels over the entire range. Consequently with an expanding type device the range is $30r$ and Eq. (13) can be rewritten as follows:

$$W=(E-6)/30r. \tag{23}$$

The values of the effective sensation levels are computed by Eq. (16) as before. This equation contains a peak factor p, representing the level difference between the intensity exceeded by 1 percent of $\frac{1}{8}$th second intervals of speech and the long average intensity of all intervals, which may be changed in value by the expanding action from the 12-db figure which applies to the ordinary distribution of speech levels. This can be easily computed for any value of r, but for values of r between 1 and 1.2 the values of p are practically constant at 12 db.

In cases of compression, large ratios of expansion or expansion that varies with level, it is necessary to calculate the level distribution of the received speech sounds from the characteristics of the non-linear device and the level distribution of speech shown in Fig. 4. The lower curved portion of this distribution, rather than the straight line approximation, should be used in cases of compression. Then W can be computed by determining the fraction of sounds in the modified distribution that are audible. It can be seen that, in general, this procedure will be laborious and cannot be expressed in a convenient mathematical form.

At the present time this treatment of non-linearity should be regarded primarily as an hypothesis. It has, however, been successful in explaining qualitatively the results obtained with a few systems containing non-linear elements, but other complicating factors are also involved. For example, in computations involving non-linear elements the conception of response is not as clear as it is in the case of linear elements and the shape of the response characteristic that is obtained may vary widely, depending upon the type of measurement that is made. Considerable caution must, therefore, be used in interpreting the results of computations of the articulation index of speech, received over systems containing non-linear elements.

5.9 The Effect of Hearing Loss

The term β_0 used in the above relations is an idealized threshold for single frequency tones and is close to the minimum of sound that can be heard by people having the most acute ears. The hearing of most people will be some 10–15 db less acute than this.[15,16] In practical problems

[15] J. C. Steinberg and M. B. Gardner, J. Acous. Soc. Am. 11, 270 (1940).

[16] J. C. Steinberg, H. C. Montgomery, and M. B. Gardner, J. Acous. Soc. Am. 12, 291 (1940).

there is usually sufficient noise to cause a threshold shift of more than 10–15 db by masking. Under these conditions calculations should be valid even though they are based on the acute β_0. In general, computations employing β_0 should be valid, except perhaps in the region where W normally is approaching unity, for all individuals having hearing losses somewhat less than the masking caused by noise, up to masking values of 40–50 db. For quiet conditions it is necessary to replace the idealized β_0 by the actual threshold values of the individuals under consideration. This procedure should give reasonably valid results for hearing losses up to 40–50 db from'the idealized threshold. For greater hearing losses, the validity of the methods becomes questionable because of modulation and other effects.

6. RELATION OF ARTICULATION INDEX TO SUBJECTIVE MEASURES

While the computational method which has been described was derived from syllable articulation tests, it is possible to interpret the resulting index in terms of subjective measures which use words or sentences. For this purpose it is only necessary, with a particular testing crew, to make subjective tests of the desired character under a variety of conditions where all the required data on the circuits, the speech spectrum, etc., are sufficiently well known to permit computing the articulation index of the received speech. These computed indices plotted against the measured word or sentence intelligibility will thus provide an empirical relationship for interpreting the results of other computations. Approximate relations of this character are given by Fig. 23. Taking, as a starting point, the curve of this figure which shows the relation, derived previously, between articulation index and syllable articulation, the other curves were obtained by using published relations[13] between syllable articulation and word and sentence intelligibility.

Although these relations apply only to specific testing crews and types of material, several features are worth noting. For instance, if speech is impaired sufficiently to lower its articulation index to one-half, sentence intelligibility may still remain high. For comparing transmission systems which have articulation indices in the range of 0.5 to unity, sentence intelligibility tests, affording only a small range of errors, are consequently impracticable. Articulation tests are more useful for this purpose. Sentence tests are useful, however, for comparing conditions which provide poor reception. For articulation indices in the range of zero to 0.3, sentence intelligibility is a sensitive measure, varying from zero to about 90 percent. Another point—although sentence intelligibility may fall by only a small amount when the articulation index is reduced to only half its maximum value, it is apparent, by referring to the curve for syllables, that a listener fails under this condition to recognize correctly a substantial portion of the sounds which are received. The high sentence intelligibility in this case must be attributed to the listener's ability to utilize context and to guess the unintelligible sounds, owing to the restricted number of sound

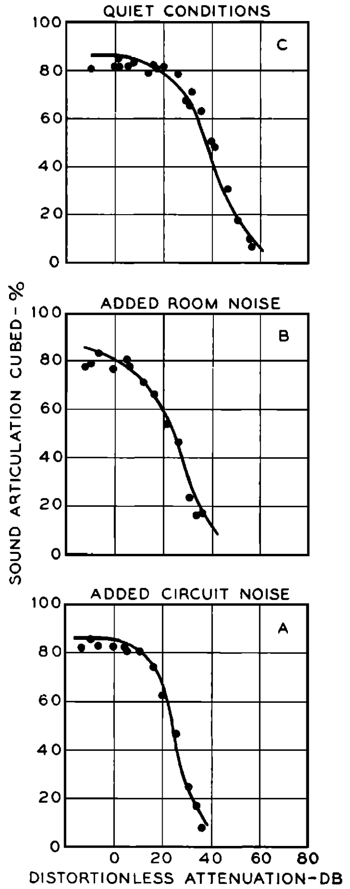

FIG. 27. Comparison of observed and computed results for 1936–1937 articulation tests. Points show observed data.

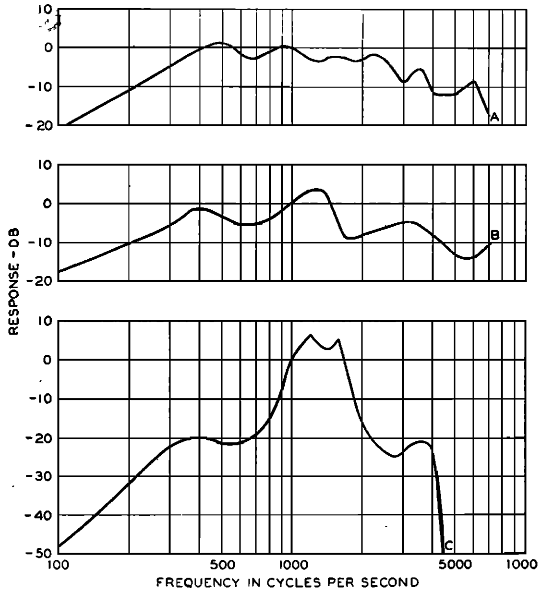

FIG. 28. Orthotelephonic response of test circuits before insertion of networks—expressed in db relative to *OT* response at 1000 cycles. Curve *A* applies to 1928–1929 articulation tests, curve *B* to 1935–1936 tests and curve *C* to 1936–1937 tests.

combinations which form actual words. This requires some effort. It seems probable that the relatively flat portion of the sentence curve is accompanied by an appreciable change of effort.

7. APPLICATIONS—ILLUSTRATIVE EXAMPLES

Although the development of the relationships which have been discussed appears rather complex, they are comparatively simple in application. For certain types of problems, at least, the solutions are practically self-evident once the fundamental data are available, as will appear from an illustrative example.

Let it be assumed that speech is being received in the presence of noise of the continuous spectrum type, and that the response of the communication system is such, that there is a constant difference in db at all frequencies between the average intensity of the speech per cycle and the average intensity of the noise per cycle, both in the ear. Let it be assumed that this difference is 0 db, i.e., that the noise and speech spectra are identical at all frequencies and that there is no limitation on band width. Under these conditions, unless both the speech and noise have large variations in their absolute levels from one frequency band to another, the controlling source of interference will be the noise. In other words, masking of speech on itself, either in the same band or on adjacent bands will be negligible relative to the noise. From Eq. (17)

$$W=1/30(B_S+p-6-m-(B(+)X)).$$

In the example under discussion, $B=B_S$ and, if the level of the noise is well above threshold, B is so large with respect to X that $(B(+)X)$ is equal to B and hence to B_S. Also, unless the noise is very intense, m is zero. Then letting $p=12$ db, the equation reduces to

$$W=6/30=0.2.$$

In other words, each of the twenty bands makes one-fifth of its maximum possible contribution to articulation index, and thus the articulation index of the received speech is 0.2. According to Fig. 23 this corresponds to a sentence intelligibility of about 70 percent.

Now let it be supposed that the band width is restricted to the frequencies below 1900 cycles which, according to Table III, eliminates ten of the twenty bands. The articulation index is consequently cut from 0.2 to 0.1 and sentence intelligibility to about 25 percent.

It is now natural to ask how much the remaining passed band would have to be raised in level, with respect to the noise, to restore the intelligibility to its original amount, namely 70 percent. To do this the articulation index of the limited band has to be restored to its original amount, namely 0.2. Since only half the bands are now contributing, the contribution of each band must consequently be doubled, or W for each of the ten lower bands must be raised from 0.2 to 0.4. Substituting $W=0.4$ in Eq. (17) the corresponding value of B_S-B is found to be +6 db. Thus the speech has to be raised by 6 db as compared to the original condition to restore the intelligibility to 70 percent.

Although a great deal can be learned by analyzing problems in the above manner this is not always possible without some loss of accuracy. With intense noise the term m in Eq. (17) should be evaluated. Also the equivalent noise of self-masking B_f should be included in the total noise B whenever the effective sensation level E of the speech in a band is greater than about 25 db.

8. COMPARISON OF COMPUTED AND OBSERVED ARTICULATIONS

Articulation index computations have been made, covering the test circuits which provided the basic data for the formulation, and also a wide variety of additional circuits for which articulation test data were available. Comparisons of the observed and computed results are given by Figs. 24–27. These figures, representing only a small part of similar comparisons, were selected because they cover a rather wide range of types of distortion and are representative of the kind of agreement between observation and calculation that is generally obtained. These figures cover tests made with three circuits having, before insertion of the networks to be tested, the response characteristics shown in Fig. 28. Curve A applies to Fig. 24, curve B to Figs. 25 and 26, and curve C to Fig. 27.

In the tests of Fig. 24 the observed quantity was syllable articulation, shown by the points. The data were obtained during 1928–1929 and are part of the fundamental data on which the present formulation is based. The curves represent computed values of articulation index translated into syllable articulation by means of the previously derived relationship (Fig. 23) between these quantities for the 1928–1929 crew.

The data of Figs. 25–27 were obtained during 1935–1937 with a different testing crew. The same type of syllables used in the 1928–1929 tests were employed, but the automatic equipment used by the observers for recording and totaling the results gave the results in terms of sound articulation, i.e., the percentage of the called sounds that were correctly understood. Each point on Figs. 25–27 is the cube of the observed sound articulation, which is approximately equivalent to syllable articulation. The curves of these figures represent computed values of articulation index translated into sound articulation cubed by means of the relationship between these quantities shown on Fig. 23. This relationship was established by determining with the 1935–1937 crew the maximum sound articulation (as a function of received level) of each of a number of sharp cut-off filters, and plotting the cube of the maximum sound articulation of each filter against its known value of articulation index at optimum volume, as derived from the 1928–1929 tests.

The computations leading to the curves of Figs. 24–27 were carried out in accordance with the procedures summarized in Section 5.7. The data needed are the over-all response characteristics of the test circuits, the acoustic speech level of the callers and the spectrum, in the observer's ears, of any interfering noise that may have been present. The over-all response of the circuits of the Fig. 24 tests may be obtained, at any value of abscissa, by adding the abscissa value to curve A of Fig. 28 and assuming an infinite loss beyond the filter cut-off frequency. The response corresponding to any value of the abscissas of Figs. 25 and 26 may be obtained by subtracting the abscissa value from curve B of Fig. 28, adding a constant value of 39 db and subtracting the insertion loss of the appropriate network of Fig. 29. For Fig.. 27 the response is obtained by subtracting the abscissa value from curve C of Fig. 28 and adding a constant value of 21.5 db.

Measurements of the speech output of the microphone were made during the tests. The speech output of the microphone was also computed by adding the acoustic speech spectrum of Fig. 2 to the response of the microphone and integrating the resulting spectrum. The amount

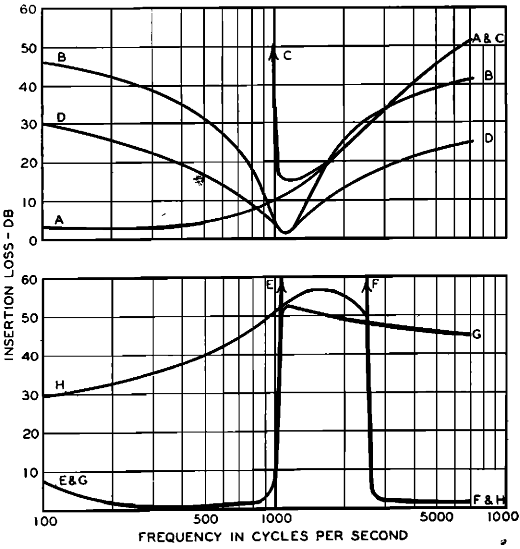

FIG. 29. Insertion loss of networks used in 1935–1936 tests.

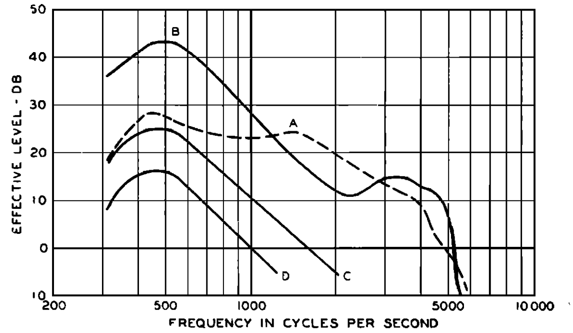

FIG. 30. Estimated effective level of noise in observer's ears during articulation tests. Curve *A*—added circuit noise (several 1936–1937 tests). Curve *B*—added room noise (several 1936–1937 tests). Curve *C*—assumed residual noise during 1936–1937 tests. Curve *D*—assumed residual noise during 1935–1936 and 1928–1929 tests.

that a measured value exceeds a computed value represents the amount that the acoustic speech level (*V*) at two inches from the lips for the tests exceeds 90 db. It was determined in this way that the average value of *V* for the tests of Fig. 24 was 94 db and 92.5 db for the tests of Figs. 25–27. The values of *V* for the individual test conditions were generally within ±1 db of these average values.

Preliminary computations of an articulation index of the three basic circuits *without added networks*, yielded at low levels of received speech, somewhat higher values than the test results, indicating that the effective sensation levels (*E*) of the received speech during the tests were somewhat lower than the computed levels. Such a disagreement was to be expected in view of the discrepancies shown in Table IV, since they indicate that the effective sensation levels of the low frequency bands were considerably lower in the tests than the calculated values would predict. The reasons for the discrepancies are not definitely known. They could arise from one or more of the following factors:

(a) A low level residual noise, such as might be produced by the movements of the several observers in the test booths, might have been present during articulation tests. Owing to the leakage characteristic of a receiver held to the ear, such a noise would be expected to produce its principal masking in the low frequency bands.

(b) The observers might have held their earphones less tightly to their ears during articulation tests than the earphones were held in the real ear calibrating tests of receiver response. The principal effects of such a variation would be to decrease the receiver response in the low frequency bands.

(c) In calculating the effects of interband masking, the masking effects of a given band on bands of lower frequency were neglected. These effects may not be negligible and the lower bands might be masked to some extent by adjacent higher frequency bands.

(d) In calculating the thresholds of the speech bands, the method assumes that, irrespective of the frequency of the band, the threshold is determined by the 1 percent points of the speech intensity distributions. These are taken to be 12 db above the long average intensity for all bands. There is some evidence that the peak factors for the low frequency bands are less than 12 db.

(e) The 1928–29 derived ΔA curves (Figs. 17–20) were drawn to zero articulation contribution ($W=0$) as straight lines. It has been assumed that the effective band sensation level at this zero point is 6 db for all bands. It may be that this factor should be larger for the low frequency bands which would be in the direction to reduce the discrepancies in Table IV.

The effects of all of these factors if known and taken into account, would tend to result in the calculation of lower effective sensation levels for the low frequency bands and hence smaller values of *W*, than is now done with the present method. The procedure that has been used was to take residual noise as the entire cause of the discrepancy, assume this noise to have the same shaped spectrum as room noise, modify it by the response of the leakage path between receiver and ear and then make computations for different absolute levels of the noise until the best agreement was obtained between the computed and observed articulations *at low levels* on the basic circuits *before* insertion of any of the distorting networks. Having derived in this manner estimates of an assumed residual noise, these same estimates were then used in all the res maining computations. Curve *D* of Fig. 30 show-the effective levels (*Z*) of the residual noise used in the computations of Figs. 24–26. Curve *C* of Fig. 30 applies to Fig. 27.

The tests of the two lower curves of Fig. 27 were made with noise added deliberately. Curve *B* of Fig. 30 shows the effective levels (*Z*) of the room noise introduced into the booth for the tests of the center curve of Fig. 27. The values of (*Z*) were obtained by subtracting *X* from the intensity levels (*B*) of this noise in the ear. The intensity levels were obtained by combining the measured spectrum of the noise in the booth with the measured shielding effect of the receiver on the ear. Curve *A* of Fig. 30 shows the effective levels of the added noise used in the tests of the lower curve of Fig. 27. This noise was introduced electrically into the receiver. The intensity levels in the ear, used in computing the effective levels, were obtained by combining the electrical spectrum of the noise with the real ear response of the receiver.

Certain broad conclusions can be reached from the comparisons of computed and observed results afforded by Figs. 24–27: (1) the computational method appears to define reasonably well the steep parts of the articulation *vs.* received speech level curves and (2) the calculated results at high received levels tend to be too large. The latter tendency may result in part from omission of intermodulation products produced in the ear at high speech levels as discussed in Section 5.5. Also, it may result in part from the assumption of equivalent noises of self and interband masking as fixed numbers of decibels below the speech levels, independent of absolute level, and the omission, in the interband masking functions, of masking on speech by speech at higher frequencies.

Although the assumption of a residual noise in the observing booth improves the agreement between computed and observed articulation at low speech levels by reducing the effective sensation levels of the speech in the lower frequency bands, it should be noted that the procedure is arbitrary and does not indicate that residual noise of this magnitude was actually present during the tests. It simply indicates that for the range of transmission conditions for which tests and calculations have been made, the effects of the discrepancies in the low frequency bands may be lumped into an effect produced by the assumed residual noise.

Although the present computational procedure has given reasonably good agreement between calculated and observed results for a wide variety of systems and hence throws light on the factors which govern the intelligibility of speech sounds, it is hoped that future work will improve the approximations and also throw additional light on the empirical procedures that have been used.

9. ACKNOWLEDGMENTS

The work reported here has drawn on the contributions of a number of people. First should be mentioned the names of H. Fletcher, W. H. Martin, and E. C. Wente, who not only were responsible for the broad content of the work but contributed significantly to the fundamental studies and their application to problems in telephone engineering and design. In the formulation of the relationships, the major portion of the work was carried out by A. B. Anderson, P. V. Dimock, J. W. Emling, R. H. Galt, W. Koenig, Jr., and W. A. Munson. The basic speech, hearing, and articulation data, which enter into the relationships, involved these same people and others to whom reference has been made throughout the text.