Derivation of primary parameters and procedures for use in speech intelligibility predictions

Chaslav V. Pavlovic
Department of Speech Pathology and Audiology, University of Iowa, Iowa City, Iowa 52242

(Received 5 December 1986; accepted for publication 6 April 1987)

The literature on various parameters that appear in the articulation index-type calculations of speech intelligibility is reexamined. Based on the reported data, the best estimates of these parameters and the most appropriate procedures for their use are suggested. These included: (1) the analysis and specification of the importance of various frequency bands to speech intelligibility; (2) the procedures used for measuring threshold and the calculation of threshold-based parameters used for predicting intelligibility of low-level speech; and (3) the calculation and measurement of relevant speech parameters. All results are given so that the calculations can be performed either in critical bands, 1/3 octaves, or octaves.

PACS numbers: 43.71.Gv

INTRODUCTION

Over the last decade, there has been a renewed interest in the procedures to predict speech intelligibility under various conditions of distortion. The methods that have been used for these predictions are all, to a large degree, based on the articulation index (AI) theory advanced by French and Steinberg (1947).

Although the basic procedures and parameters that appear in the AI calculations were standardized according to ANSI (1969), almost every researcher who has recently used the method as a tool for predicting speech intelligibility has changed it to a greater or lesser degree. The underlying reason for the changes is the availability of the large new body of data related to various parameters that appear in the AI calculations. The fact that different researchers use different modifications (often insufficiently documented) renders comparison of results obtained in various studies virtually impossible.

In this report, the literature on various parameters that appear in the AI calculations is reexamined. Based on the reported data, the best estimates of these parameters and the most appropriate procedures for their use are suggested. These include: (1) the analysis and specification of the importance of various frequency bands to speech intelligibility; (2) the procedures used for measuring threshold and the calculation of threshold-based parameters used for predicting intelligibility of low-level speech; and (3) the calculation and measurement of relevant speech parameters.

These parameters and procedures are the basic ones that typically appear in the AI calculations. There are, however, many others that are important in specific conditions of distortion. Among them are those related to reverberation distortion, those related to the effects of high speech presentation level, and those related to the sharp filtering of speech or of the interfering noise. These "secondary" parameters/procedures are not the topic of this study.

The articulation index method is best described by the following two equations:

$$A = P \sum_{i=1}^{n} I_i W_i, \tag{1}$$

$$s = T(A). \tag{2}$$

Here, A, the articulation index, is an intervening variable that relates speech intelligibility to physical parameters. It is related to speech intelligibility (s) through an empirical transfer function represented by Eq. (2). The fundamental characteristic of A, as seen from Eq. (1), is that it is the algebraic sum of contributions I_i, W_i associated with each band i. The importance function, I_i, characterizes the importance of a speech frequency band i to speech intelligibility. The weighting function, W_i, is equal to the proportion of the speech dynamic range within the band i that contributes to speech intelligibility under conditions that are less than optimal. The factor P, termed proficiency factor, is a measure of how precise the talker's enunciation of the speech material is, and how experienced the listener is in listening to the talker. Under ideal circumstances, its value is 1. The number of bands n that has traditionally been used in AI calculations (Kryter, 1962a) is, in the order of accuracy of the procedure, 20 (bands chosen to be of equal importance), 15 (1/3 octave bands), or 5 (octave bands). In this study, the method of equally contributing bands is abandoned in favor of a method that employs critical bands. Critical bands, as reported by Zwicker (1961), are used.

The material that is discussed in each of the following sections is, to a large degree, independent. Therefore, this report is organized in such a way that relevant discussions and conclusions are contained within the appropriate sections, rather than at the end of the article.

I. IMPORTANCE FUNCTION

Speech recognition is the end product (output) of a complex communication channel whose input is the message conceived by the talker. In order to communicate, the message is transformed into a physical signal that is transmitted over the communication channel and decoded by the listener. In this process a distortion is introduced so that the prob-

 0001-4966/87/080413-10$00.80

ability that the reconstructed message will match the original is less than 1. This probability increases if the decoder mechanism is cognizant of the various statistical properties of the message. That is, reconstruction of the message is easier if the listener can make use of the sequential or contextual constraints existing in the message, or if he or she is aware of the limitation on the size of the vocabulary pertaining to the message. In this study, all of these constraints together will be referred to as the redundancy of the message.[1]

Traditionally, various AI methods for computing speech recognition from the physical parameters of the signal and distortion have assumed that the distribution over frequency of the usable information content of the signal is not a function of the redundancy. In other words, the AI method assumes that importance function does not depend on the message redundancy. The fact that speech recognition improves with an increase in redundancy is accounted for by using different transfer functions for different speech materials. This assumption has been questioned by various authors (e.g., Boothroyd, 1978). In order to analyze whether this traditional approach is justifiable, the importance functions of various speech materials are compared to each other. In this comparison, an assumption is made that the differences, if any, in the phonemic composition of the various speech materials analyzed were not sufficiently large to have an effect on the importance function. (Boothroyd, in 1978, and Duggirala *et al.*, in 1986, have shown that severe changes in the phonemic composition of speech have the capacity to alter the importance function.)

The importance function most widely used for all redundancy levels is one based on a series of studies done principally at Bell Telephone Laboratories in the 1920s and 1930s and reported by French and Steinberg (1947). Modified versions, applicable to the average speech of male talkers only, were reported by Beranek (1947), Kryter (1962a), and the ANSI (1969) standard. The importance function of French and Steinberg was obtained by using nonsense syllables of the CVC type, as shown in Fig. 1 (solid line). Both in this figure and in the figures that follow, the original results were recalculated to correspond to critical bands. The importance function has a peak at critical band 15, or 2500 Hz.

The importance function obtained recently by Studebaker *et al.* (1987) for running speech is shown in Fig. 1 by the dashed line. Their speech sample consisted of exceptionally easy reading passages (seventh grade reading level) and, therefore, this importance function is, in all probability, at the opposite end of the continuum from the one for nonsense syllables. The peak has now changed to critical band 5, or from 2500 to 450 Hz. The direction of this shift was anticipated by Miller and Nicely (1955) in their classical study on the effects of filtering on consonant confusion. Their data show that errors made under low-pass filtering are much less random (i.e., more predictable) than errors made under high-pass filtering. Therefore, when there is redundancy in the message, the listener can better detect and correct, on the basis of context, low-frequency perceptual errors than high-frequency errors. The higher the redundancy, the more pronounced this effect is, and the more information is transmitted via the low frequencies relative to the high frequencies. In addition to the explanation of Miller and Nicely, there also may be other alternatives or additional mechanisms that could be responsible for the shift in the importance function. That is, the perceptual mechanisms for processing on-going contextual information may be considerably different from those for isolated stimuli. Items such as syllables, words, phrases, etc., may have a perceptual unity. In such an event, efforts both to explain perception in terms of sequential phonetic information and to assign to all these units the same importance function would not be successful.

Figure 2 shows the importance function obtained from

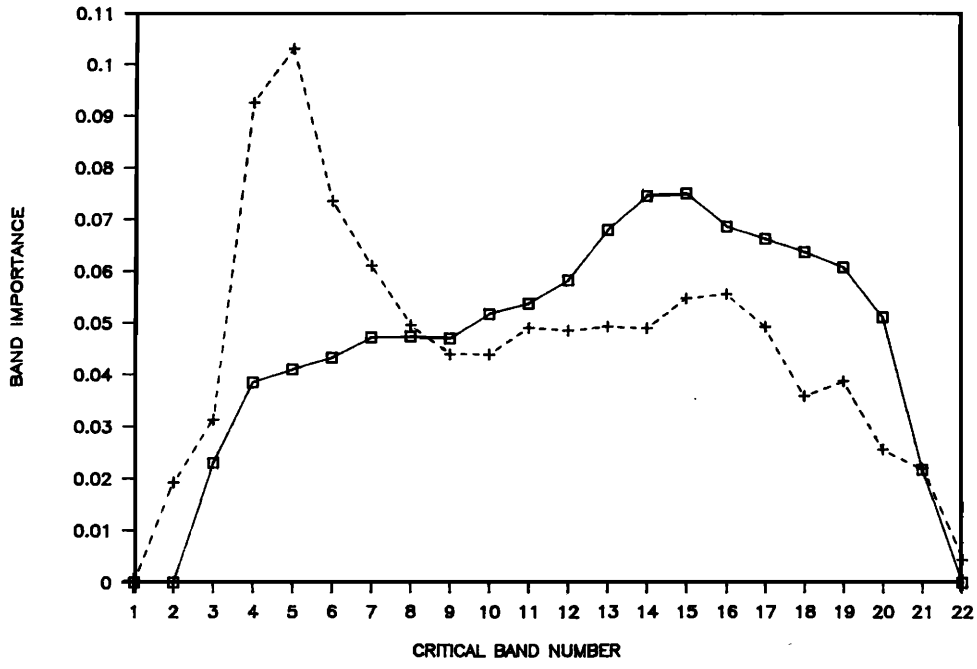

FIG. 1. Importance functions for nonsense syllables (solid line) obtained by French and Steinberg (1947) and for easy running speech (dashed line) obtained by Studebaker *et al.* (1986).

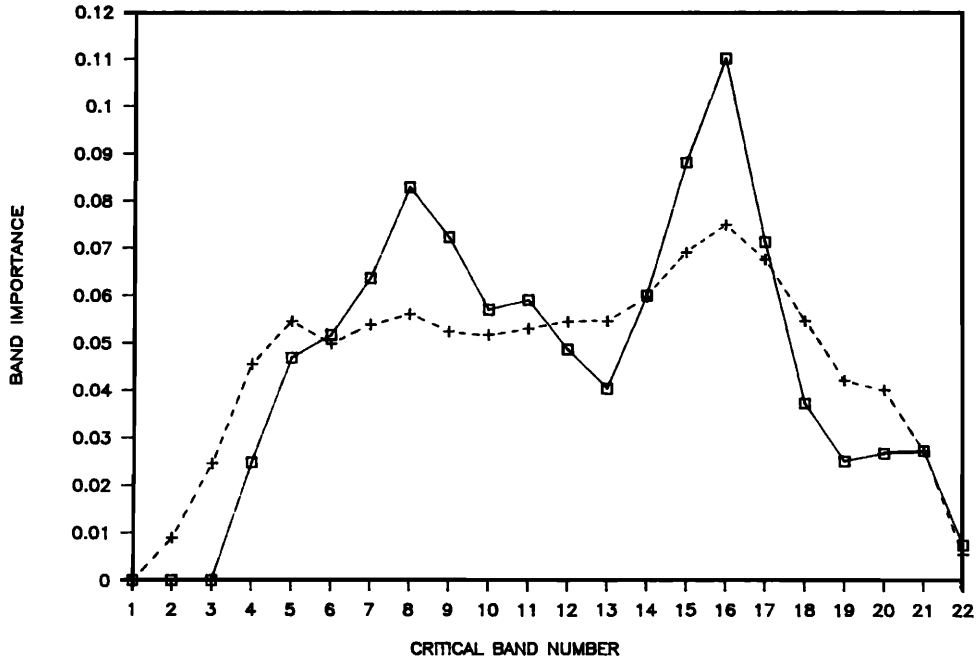

FIG. 2. Importance functions obtained by Black (1959) for phonetically balanced meaningful words (solid line) and for a four-alternative multiple choice words material (dashed line). Because Black's original results are for male speech only, both curves were shifted up in frequency by 16% to arrive at values that better approximate average male and female talkers.

the results reported by Black (1959) for phonetically balanced meaningful words (solid line). It is interesting to note that this medium-redundancy material now shows two peaks. The dashed line gives the importance function for a four-alternative multiple choice words material and is also calculated from the data reported by Black (1959). Because Black's results are for male speech only, the original importance functions were shifted up in frequency by 16% to arrive at values (shown in Fig. 2) that better approximate average male and female talkers. The 16% value was calculated from the study of Peterson and Barney (1952) as the average difference between the male and female formants ($F1$ to $F3$). It is also identical to the value used by Beranek (1947) to obtain the importance function of male speech from the values for average speech.

The solid line in Fig. 3 was obtained by averaging the importance functions discussed thus far. Aside from the edges of the curve, there appears to be a tendency to obtain equal importance per critical band. It seems probable that a proper sample of various speech materials, weighted appropriately in accordance with their representation in daily communication activities, should indeed result in a flat importance function. The flat important function on the bark scale for this, which may be termed "average speech," would mean that the speech code was designed to optimally match the receiver. At the same time, because the matching is done

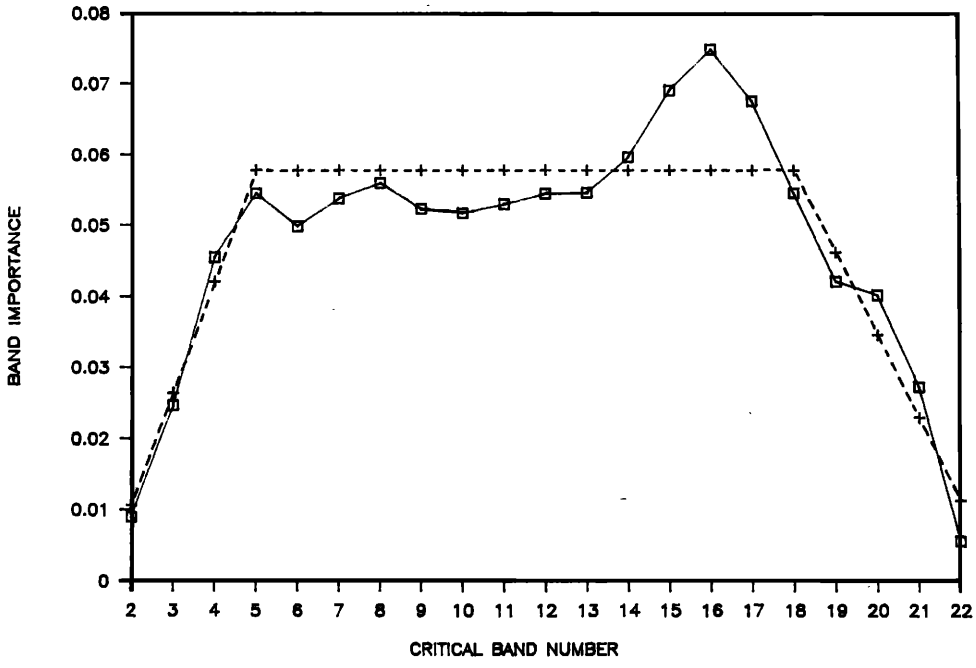

FIG. 3. The average importance function. The actual data are given by the solid line. The best estimate is shown by the dashed line.

in respect to the critical band, it would also mean that the weakest link in the speech recognition apparatus is the auditory filter. The optimization discussed above refers to maximizing the reception rate of information. If each channel of the receiver, i.e., each critical band, receives the same amount of independent information, this rate is maximized (Shannon, 1948).

The dashed line in Fig. 3 represents the best estimate of the importance function of average speech. It was obtained by assuming that the importance function over critical bands 5–18 is constant and equal to the mean value of the actual data over these bands. It was further assumed that the importance function indeed tapers off at the edges of the curve, as indicated by the actual data. In these areas, it was calculated as the best-fit straight line to the data on the condition that it does not change the cumulative importance values of the areas. Tables I, II, and III give the estimated importance function of average speech per critical band, per 1/3 octave band, and per octave band, respectively. In addition, the equivalent data for nonsense syllables, calculated from the results of French and Steinberg (1947), and for easy running speech calculated from the results of Studebaker *et al.* (1987), are shown.

The importance function for average speech does not relate to any specific speech material. Therefore, it is suggested that it be used to obtain an AI that is a more general measure of speech intelligibility than the speech intelligibility performance with any specific material. This, however, does not preclude the use of the importance function for predicting the intelligibility of a specific speech material. Because it was developed for material of average redundancy, it is likely to produce, across various speech materials, a more accurate prediction than the importance functions for materials of very high or very low redundancy.

TABLE I. The critical band importance functions for nonsense syllables, easy running speech, and average speech.

Crit. band No.	Crit. band C.F. (Hz)	Band importance		
		Nonsense syllables	Easy speech	Average speech
2	150	0.0000	0.0192	0.0103
3	250	0.0230	0.0312	0.0261
4	350	0.0385	0.0926	0.0419
5	450	0.0410	0.1031	0.0577
6	570	0.0433	0.0735	0.0577
7	700	0.0472	0.0611	0.0577
8	840	0.0473	0.0495	0.0577
9	1000	0.0470	0.0440	0.0577
10	1170	0.0517	0.0440	0.0577
11	1370	0.0537	0.0490	0.0577
12	1600	0.0582	0.0486	0.0577
13	1850	0.0679	0.0493	0.0577
14	2150	0.0745	0.0490	0.0577
15	2500	0.0750	0.0547	0.0577
16	2900	0.0685	0.0555	0.0577
17	3400	0.0662	0.0493	0.0577
18	4000	0.0636	0.0359	0.0577
19	4800	0.0607	0.0387	0.0460
20	5800	0.0511	0.0256	0.0343
21	7000	0.0216	0.0219	0.0226
22	8500	0.0000	0.0043	0.0110

TABLE II. The 1/3-oct-band importance functions for nonsense syllables, easy running speech, and average speech.

1/3-oct C.F. (Hz)	Band importance		
	Nonsense syllables	Easy speech	Average speech
160	0.0000	0.0114	0.0083
200	0.0000	0.0153	0.0095
250	0.0153	0.0179	0.0150
315	0.0284	0.0558	0.0289
400	0.0363	0.0898	0.0440
500	0.0422	0.0944	0.0578
630	0.0509	0.0709	0.0653
800	0.0584	0.0660	0.0711
1000	0.0667	0.0628	0.0818
1250	0.0774	0.0672	0.0844
1600	0.0893	0.0747	0.0882
2000	0.1104	0.0755	0.0898
2500	0.1120	0.0820	0.0868
3150	0.0981	0.0808	0.0844
4000	0.0867	0.0483	0.0771
5000	0.0728	0.0453	0.0527
6300	0.0551	0.0274	0.0364
8000	0.0000	0.0145	0.0185

II. PARAMETERS THAT DETERMINE W_i

In Eq. (1), W_i was defined as the propagation of the speech dynamic range that contributes to speech intelligibility under conditions that are less than optimal. Therefore, in the case of external noise distortion or low-level speech, the predictions of speech intelligibility critically depend on the accurate specification of both the dynamic range of speech and its relationship to the external noise and the threshold of hearing.

Speech represents a time-varying signal. Therefore, its dynamic range will depend on the time constant used in measuring the distribution of its level over time. However, accurate speech intelligibility predictions are obtained when a 125-ms integration time is used (French and Steinberg, 1947; Kryter, 1962b; Pavlovic and Studebaker, 1984). In noise, a 125-ms speech sample (hereafter referred to as "speech sample") will, in a given critical band, contribute to speech intelligibility if its power in the band is larger that the power of the masking noise in the same band. This represents

TABLE III. The octave band importance functions for nonsense syllables, easy running speech, and average speech. (The 150-Hz octave is not included. Its relatively small contribution has been added to that of the 250-Hz octave.)

Oct C.F. (Hz)	Band importance		
	Nonsense syllables	Easy speech	Average speech
250	0.0437	0.1004	0.0617
500	0.1294	0.2551	0.1671
1000	0.2025	0.1960	0.2373
2000	0.3117	0.2322	0.2648
4000	0.2576	0.1744	0.2142
8000	0.0551	0.0419	0.0549

one of the basic principles of the AI method and has been found to be reasonably accurate (Kryter, 1962b; Pavlovic and Studebaker, 1984).

In regard to the discussion that follows, it is important to emphasize that we are concerned here with speech intelligibility rather than with speech detectability. It is conceivable that a sample of speech may be detectable at levels lower than are sufficient for any contribution to speech intelligibility. In other words, at the threshold of speech detectability in noise, the speech power in a critical band may be lower than the power of the masking noise in the same band. At the threshold of intelligibility, these two levels are equal. In the case of a pure tone signal, its power at threshold of detectability in noise is indeed about 4 dB lower than the power of the masking noise (Scharf, 1970).

A. Threshold problem

In quiet, the AI method of French and Steinberg (1947) compares the speech power to the power of a fictitious internal noise. This noise is calculated so that, if it were an external masker, it would give rise to the observed pure-tone threshold in quiet Q. Therefore, the power spectrum density X of the internal noise is

$$X = Q - R, \tag{3}$$

where R is the critical ratio in dB. The variable X will be referred to as "threshold spectrum density."

Let us denote the power spectrum density of a speech sample as Y. In the method of French and Steinberg, analogous to the external noise situation, the difference between Y and X determines whether the sample contributes to speech intelligibility (the difference is positive) or not (the difference is less than or equal to zero). This difference is, therefore,

$$Z = Y - Q + R. \tag{4}$$

In the ANSI (1969) standard, the quantity Z' (which determines whether a speech sample is intelligible) is specified as the difference between Y and the spectrum density of a just detectable noise. Relatively recent research (Berger, 1981; Cox and McDaniel, 1986) suggests that the spectrum density of a just detectable noises is equal to the difference between the pure-tone threshold Q and the critical band in decibels C. Therefore, Z' can be expressed as

$$Z' = Y - Q + C. \tag{5}$$

Combining Eqs. (4) and (5), we obtain

$$Z' = Z + K, \tag{6}$$

where K is the difference between C and R. When R is calculated as the mean of the data of Fletcher (1953) and Hawkins and Stevens (1950), and C is calculated from Zwicker (1961), the average value of K for the frequency region used in the AI is 3.9 dB. Therefore, it appears that the procedure of French and Steinberg and the procedure of the ANSI (1969) would disagree by this amount. The data in Table IV support this conclusion. In column B, the threshold values for "the sounds having continuous spectra" from the ANSI standard are given. These are the values that are subtracted from the Y values to obtain the quantities Z' discussed above. Column F gives the threshold spectrum density X calculated by subtracting the critical ratio R (column C) from the monaural pure-tone threshold (column D plus column E). These are the values that are subtracted from the Y values to obtain the quantities Z. Therefore, the difference between column B and column F represents the difference between Z' and Z. The mean difference between these two columns is 3.8 dB. This is almost exactly equal to the expected difference (3.9 dB) based on Eq. (6).

The use of Z rather than Z' in the AI procedures appears to be justifiable. Analogous to the external noise procedure, the issue here is speech intelligibility rather than speech detectability. Here, Z is based on the threshold of intelligibility ($Z = 0$ when the speech sample and the masker have equal energy), while Z' is based on the threshold of detection ($Z' = 0$ when the speech sample is at the threshold of detection). The possible overestimates of the true signal-to-threshold ratio in the ANSI (1969) procedure ($Z' > Z$) may have been the cause of the observed speech intelligibility scores being greater in noise than in quiet for the same AI (AI validation study, Kryter, 1962b). In recent studies of Pavlovic and Studebaker (1984), Dirks *et al.* (1986), and Pavlovic *et al.* (1986), where Z was used, there were no discrepancies between the predictions in noise and in quiet.

In regard to Pavlovic and Studebaker, Dirks *et al.*, and Pavlovic *et al.*, it should be pointed out that the authors did not use the ISO (1961) values for deriving X, as was done in Table IV; rather, they measured individual thresholds. However, an analysis of the techniques they employed suggests that their thresholds converge to the ISO values corrected for monaural listening.[2] This justifies the use of the ISO/R 226-1961 in Table IV. Moreover, it follows that, in the case of a procedure that does not converge to the ISO (1961) values, appropriate correction factors should be used.

An important consequence of the discussion above relates to the treatment of 1/3-oct noise band thresholds (Q_n) in the AI calculations. Since, according to Berger (1981) and Cox and McDaniel (1986), the 1/3-oct noise threshold is virtually identical to the pure-tone threshold Q, the quantity Z should be determined in the same manner as in Eq. (4):

$$Z = Y - Q_n + R. \tag{7}$$

It would not be correct to subtract Q_n from the 1/3-oct speech level ($Y + C$).

In summary, the following is recommended in regard to the threshold.

(1) The approach of French and Steinberg (1947) for determining the intelligibility of low-level speech is preferred to that of ANSI (1969). For the average normal-hearing listener, the values X that should be compared to the speech signal are given in Table V (column D).[3] These values are derived as explained in regard to Table IV. For the octave band AI procedure, it would be more accurate to use the X values obtained by averaging on the power basis the corresponding 1/3-oct X values, rather than using the X values at octave center frequencies.[4] However, this results in changes of less than 1.5 dB. Given the large margin of error

TABLE IV. Comparison of different threshold values used in AI procedures. The values in column B are from ANSI (1969). Critical ratio values in column C are obtained by averaging the data of Fletcher (1953) and Hawkins and Stevens (1950). The values in column D, which represent the differences between MAF values for binaural and monaural listening, are from French and Steinberg (1947). The binaural MAF values (column E) are from ISO (1961). The values for X in column F have been obtained by subtracting column C from the sum of columns D and E.

(A) 1/3-oct C.F. (Hz)	(B) ANSI threshold (dB SPL)	(C) Crit. ratio R (dB)	(D) Diff. 2 − 1 ear MAF (dB)	(E) ISO MAF binaural (dB SPL)	(F) Threshold spectrum density X = (E) + (D) − (C) (dB SPL)
200	0.5	17.0	1.5	13.8	−1.7
250	−5.3	16.6	1.5	11.2	−3.9
315	−10.1	16.6	1.5	9.0	−6.1
400	−12.5	16.9	1.5	7.2	−8.2
500	−14.2	17.2	1.5	6.0	−9.7
630	−16.0	17.3	1.5	5.0	−10.8
800	−16.0	17.8	1.5	4.4	−11.9
1000	−16.0	18.2	1.5	4.2	−12.5
1250	−17.7	18.8	1.5	3.8	−13.5
1600	−20.4	19.5	1.5	2.6	−15.4
2000	−24.3	20.2	1.5	1.0	−17.7
2500	−28.2	21.5	1.5	−1.2	−21.2
3150	−30.0	22.6	1.6	−3.2	−24.2
4000	−28.6	24.0	2.0	−3.9	−25.9
5000	−22.8	25.0	2.5	−1.1	−23.6
6300	···	26.2	3.8	6.6	−15.8

inherent in the octave AI procedure, it is suggested not to incorporate these corrections in the octave procedure and, thereby, not to set it apart from the other two.

(2) In applications where individual thresholds need to be measured, the variable X should be determined as the sum of its value in Table V and the hearing loss. For any given threshold estimation procedure, the hearing loss component is determined as the difference between the measured threshold and the average threshold of young normal-hearing individuals.

(3) In applications where it is more convenient to use a reference point other than the free field at the listener's position, the threshold spectrum density X should be determined using Eq. (3). The variable Q should be calculated so as to correspond to the value that would have been obtained using psychoacoustical procedures compatible with those used for obtaining the ISO (1961) thresholds corrected for monaural listening (see footnote 1). Either pure tones or subcritical bands of noise could be used as the signal for determining Q. The critical ratio values that may be needed for these calculations are given in Table V (column C).

B. Long-term rms speech spectrum level

The long-term rms speech spectrum level refers to the speech sound-presssure level averaged on the power basis over time and contained within a band 1 Hz wide. In the remainder of the manuscript, it will also be referred to as the speech spectrum density, or simply speech spectrum, and denoted as either S or $S(f)$. We can write

$$S(f) = 20 \log\left[\left(\lim_{T\to\infty}\sqrt{\frac{1}{T}\int_0^T p_f^2(t)dt}\right)\Big/0.00002\right], \tag{8}$$

where $p_f(t)$ is the sound pressure in pascals at the output of a 1-Hz-wide ideal bandpass filter.

The term overall speech level (S_{tot}) will be used to denote the unfiltered speech sound-pressure level averaged on the power basis over time:

$$S_{\text{tot}} = 20 \log\left[\left(\lim_{T\to\infty}\sqrt{\frac{1}{T}\int_0^T p^2(t)dt}\right)\Big/0.00002\right], \tag{9}$$

where $p(t)$ is the speech sound pressure in pascals.

In actual applications, speech spectrum is measured at the output of a bandpass filter wider than 1 Hz. If the long-term rms sound-pressure level at the output of the filter (band level) is S_{band}, then $S(f)$ can be approximated as

$$S(f) = S_{\text{band}} - 10 \log B. \tag{10}$$

The accuracy of the approximation is inversely proportional to the magnitude of B. For AI applications, however, B as small as one critical band (i.e., close to 1/3 octave band) is quite satisfactory (Kryter, 1962b).

Table V (column E) gives the values of $S(f)$ for normal vocal effort calculated in this study.[5] They were obtained by summarizing the data reported earlier by various investigators, as discussed below. All values have been recalculated to correspond to the speech levels in the free field 1 m from the talker's lips. The values represent the arithmetic averages between the male and female spectra. The overall level of this speech is 63.0 dB SPL.

Either in the original studies, or in this study, Eq. (10) was used to derive $S(f)$. Therefore, the $S(f)$ values in Table V should be understood to represent speech spectra at all frequencies for the specified band. They should not be interpreted as representing only the values at the center frequencies of the bands, nor should the values in between these

TABLE V. Recommended values of variables needed for AI calculations. They are given at the center frequencies of 1/3 octave bands [ANSI (1984)] and at the center frequencies of critical bands (Zwicker, 1961). The values at the center frequencies of the oct bands are identical to those listed under the 1/3-oct with the same center frequencies. The values of P are given for purposes of completeness. It is suggested that 12 dB be used at all frequencies.

(A) 1/3-oct C.F. (Hz)	(B) Crit. band C.F. (Hz)	(C) Crit. ratio R (dB)	(D) Thresh. spec. dens. X (dB SPL)	(E) Aver. speech spec. dens. S (dB SPL)	(F) Peaks minus speech spec. P (dB)	(G) Hearing level of speech (dB HL)
	150	17.8	1.5	31.8	8.6	30.3
160		17.7	0.6	33.0	8.5	32.4
200		17.0	−1.7	35.2	8.5	36.9
250	250	16.6	−3.9	35.3	9.2	39.2
315		16.6	−6.1	34.6	9.7	40.7
	350	16.8	−7.2	34.9	9.7	42.1
400		16.9	−8.2	35.2	9.7	43.4
	450	17.0	−8.9	35.0	9.7	43.9
500		17.2	−9.7	34.8	9.8	44.5
	570	17.2	−10.3	33.7	9.9	44.0
630		17.3	−10.8	32.7	9.9	43.5
	700	17.6	−11.4	31.4	10.0	42.8
800		17.8	−11.9	29.0	11.6	40.9
	840	17.9	−12.0	28.5	13.0	40.5
1000	1000	18.2	−12.5	25.8	12.3	38.3
	1170	18.6	−13.2	24.0	11.0	37.2
1250		18.8	−13.5	23.4	11.0	36.9
	1370	19.0	−14.0	22.6	11.0	36.6
1600	1600	19.5	−15.4	20.5	12.5	35.9
	1850	20.0	−16.9	18.4	13.0	35.3
2000		20.2	−17.7	17.6	12.4	35.3
	2150	20.6	−18.8	17.0	9.8	35.8
2500	2500	21.5	−21.2	14.7	9.8	35.9
	2900	22.2	−23.2	13.2	10.0	36.4
3150		22.6	−24.2	12.5	10.8	36.7
	3400	23.0	−24.9	12.0	11.8	36.9
4000	4000	24.0	−25.9	10.2	11.2	36.1
	4800	24.8	−24.2	6.3	10.1	30.5
5000		25.0	−23.6	5.8	10.1	29.4
	5800	25.8	−19.0	4.1	10.8	23.1
6300		26.2	−15.8	3.1	11.9	18.9
	7000	26.9	−11.7	2.2	12.8	13.9
8000		27.7	−7.1	1.1	12.7	8.2
	8500	28.0	−6.0	−0.6	12.6	5.4

frequencies be obtained by approximation. This misinterpretation results, under some circumstances, in a serious overestimation of a speech band level.

To obtain $S(f)$, the results from various studies were weighted and averaged. The weighting coefficients were chosen in the following manner. The importance of the spectrum obtained in a given study was doubled if it referred to American English rather than to British or Australian English. The importance of the spectrum was also doubled if conversational speech had been used rather than the sentence "Joe took father's shoebench out, she was waiting at my lawn." This sentence has traditionally been assumed to have the spectrum equal to that of conversational speech. Benson and Hirsh (1953) found this assumption to be reasonably correct, although, at some frequencies, the agreement was less satisfactory. If, in a given study, both conversational speech and the sentence "Joe ... lawn" had been used, the importance of the spectrum was multiplied by 1.5. No weighting according to the number of subjects that participated in the study was performed because it would have drastically reduced the importance of studies performed on American conversational speech.

The following results were used in the averaging: (a) the American conversational speech spectrum reported by Dunn and White (1940); (b) the American conversational speech spectrum of talkers used in one of the series of articulation studies done at Bell Laboratories ("Bell Laboratory spectrum") reported by French and Steinberg (1947); (c) the spectrum of American talkers (conversational speech and "Joe ... lawn") reported by Benson and Hirsh (1953); (d) the spectrum of "Joe ... lawn" of American talkers obtained by Pearson *et al.* (1976); (e) the spectrum of Australian conversational speech used by National Acoustics Laboratories (Byrne, 1977); and (f) British conversational speech spectrum (Loye and Morgan, 1939). Results from various other studies[6] were not included in either the table or in further calculations, for one or more of the following reasons: (1) Less than three subjects were used; (2) no adequate data were reported to determine the speech spectrum, as defined in Eq. (8); (3) averaging of spectra for different subjects was done on the power basis rather than by calculating the arithmetic average; (4) the speech levels were specified in bandwidths larger than 1/2 oct; (5) talking level was used that was not normal conversational level; and (6)

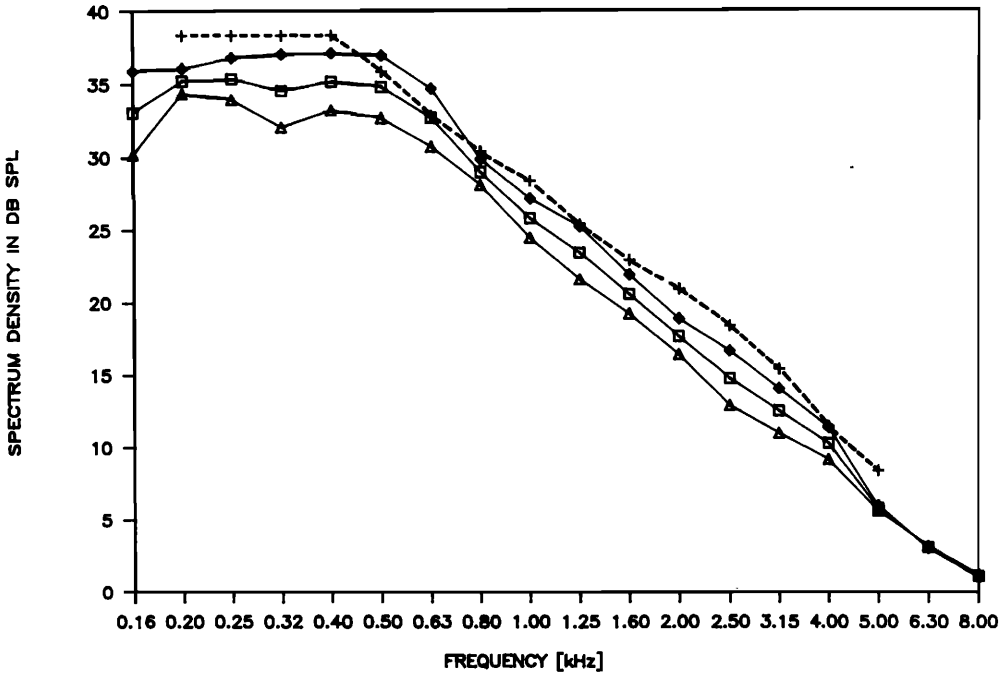

FIG. 4. The speech spectra of male speech (upper solid line), female speech (lower solid line), and average speech (middle solid line). The ANSI (1969) spectrum is given by the dashed line. The lines that connect the indicated data points are there only to facilitate visualizing the data. For calculation purposes, it should be assumed that the spectrum is flat within each band.

speech spectrum was recorded in extreme proximity to the talker's mouth (less than 10 cm).

Based on the studies in which both male and female spectra are reported (all but the Bell Laboratory study), an average difference $D(f)$ was found between the two spectra. No weighting was used for this purpose. In order to obtain the male spectrum and female spectrum, the average spectrum $S(f)$ was increased and decreased by $D(f)/2$, respectively.

Figure 4 shows the speech spectra of male speech (upper solid line), female speech (lower solid line), and average speech (middle solid line). The ANSI (1969) spectrum is given by the dashed line. The latter is mainly based on the Dunn and White spectra (1940) and on the spectra that are referred to here as the Bell Laboratories spectra (French and Steinberg, 1947). In addition, it was adjusted to correspond to male talkers only.

In applications where speech spectrum is actually measured, it should be used in place of the values in Table V. If $S'(f)$ is the measured speech spectrum [obtained, if necessary, through the use of Eq. (10)] that varies over a band of interest B', its mean value is calculated as[7]

$$S(f) = 10 \log \int_{B'} 10^{0.1S'(f)}\, df - 10 \log B'. \qquad (11)$$

In many applications, most notably with hearing-impaired listeners, it is convenient to have the speech spectrum represented on an audiogram, and compare it to individual thresholds on the same chart. It is necessary, therefore, to express the speech levels in reference to the normal thresholds. These levels are termed "hearing levels of speech" and are calculated as the difference between $S(f)$ and the threshold spectrum densities[8] X, which are given in Table V (column G).

C. Speech dynamic range considerations

Articulation index predictions, as discussed in Sec. I, are based on comparing the intensities of the signal and noise. In this comparison, the time interval over which the signal and noise are integrated is 125 ms. With this integration time, the distribution of the corresponding speech rms values is approximately linear over a 30-dB-wide range in any given frequency band (Dunn and White, 1940). The upper limit of this dynamic range is determined by the level of speech "peaks." This level is defined as the sound-pressure level exceeded only 1% of the time by speech energy integrated over 125-ms intervals. In some AI studies, values other than the 30-dB dynamic range were used (French and Steinberg, 1947), while, in other studies, a different than uniform distribution of speech samples was assumed (French and Steinberg, 1947; Pavlovic, 1984). However, the data of Pavlovic and Studebaker (1984) indicate that the 30-dB dynamic range and the uniform distribution provide for more accurate AI predictions.

The difference $P(f)$ between the speech peaks and the $S(f)$ are calculated from the study of Dunn and White (1940).[9] These values ar reported in Table V (column F). French and Steinberg (1947) suggested that for greater accuracy of results the peaks calculated in this way should be used. However, the increase in accuracy is minimal (Pavlovic and Studebaker, 1986). It is suggested that 12 dB across all frequencies be used because it results in simpler calculation procedures. The resultant dynamic range [30-dB width, $P(f) = 12$, uniform distribution] is appropriately termed "perceptual dynamic range" (Boothroyd, 1986).

III. SUMMARY

Tables I–III and V contain all the primary parameters needed in AI calculations. The calculations can be per-

formed, in the order of accuracy, either in critical bands, 1/3 octave bands, or octave bands. For even higher accuracy, the speech spectrum and the individual thresholds could be measured and used as detailed in Sec. II. In Table V, the speech spectrum density, rather than band levels, is used regardless of which procedures is desired. This is in contrast to ANSI (1969), where the spectrum density is used only for the 20-band procedure. The approach suggested here greatly simplifies the calculation of some secondary AI parameters (e.g., spread of masking).

Three importance functions are given in the tables.[10] Two are for very specific types of speech material (nonsense syllables and easy running speech). The other is for average speech and, therefore, does not relate to any specific speech material. It is suggested that, unless it is of interest to predict the speech performance with either nonsense syllables or easy running speech, the latter be used. The AI obtained using this importance function should be a more general measure of speech intelligibility than the speech intelligibility performance on any specific test material.

The AI is determined using Eq. (1). The weighting factor W_i is calculated as the ratio between the optimal dynamic range (30 dB) and the difference between the speech peaks $(S+12)$ and the threshold spectrum density X. This difference is restricted to the 0- to 30-dB range. In noise, X is substituted by the spectrum density of the noise, providing the noise exceeds X. This is preferable to the power summation of the noise spectrum density and X (Pavlovic and Studebaker, 1984). When individual thresholds are measured, X should be increased by the dB HL value of the threshold.

ACKNOWLEDGMENTS

I wish to thank Donald Dirks for bringing to my attention the disagreements of the various AI procedures with respect to threshold. I also wish to thank him, as well as Arthur Boothroyd and Patricia Tillman, for valuable comments on an earlier version of this manuscript.

[1]Boothroyd (1978) points out that the term redundancy should not be taken to imply unnecessary or useless. Rather, "... it is the means by which the probability of perceptual error can be kept to an acceptably low value despite the physical and physiologic imperfections of the real world."

[2]Pavlovic and Studebaker (1984) and Pavlovic *et al.* (1986) used a Békésy tracking technique. The rate of attenuation change was 2.5 dB/s, while the step size was 0.25 dB. The pulse duration was 250 ms and the duty cycle was 50%. The midpoints of the tracings defined the threshold. Harbert and Young (1966) found that this variation of Békésy procedure (see their Fig. 3) results in a threshold that is 2.8 dB better than that obtained by an audiometric procedure consisting of two ascending and two descending series. [The authors referred to the latter as the "conventional" procedure. For more details on this procedure, see Carhart and Jerger (1959).] Because only positive responses obtained in ascending and descending series are averaged, the step size of 5 dB leads to this threshold being half a step size worse than the thresholds given in ISO (1961). The latter is mainly based on studies by Sivian and White (1933), Churcher and King (1937), and Robinson and Dadson (1956), where the step size was 1 dB and either the method of limits or the method of constant stimuli was used. Therefore, the Békésy procedure used by Pavlovic and Studebaker (1984) and Pavlovic *et al.* (1986) is likely to result in virtually the same values as those specified in ISO (1961). This conclusion is further supported by the study by Burns and Hinchcliffe (1975), in which no differences were found between the threshold obtained by the method of limits and the threshold obtained by a Békésy tracking procedure comparable to the one used by Pavlovic and Studebaker (1984) and Pavlovic *et al.* (1986). The only major difference between the Békésy procedure of Burns and Hinchcliffe and the one used in the above studies is that a continuous signal was used in the former, while an interrupted signal was used in the latter. This, however, does not cause any differences in the thresholds of normal-hearing individuals (Harbert and Young, 1966).

The study of Dirks *et al.* (1986) used a 2 IFC procedure with a 2-dB step size. A 2 IFC procedure with 1.2 step size results in a threshold that is 6.5 dB better (Marshall and Jesteadt, 1986) than the clinical threshold procedure specified in ANSI (1978). Because the latter averages only positive responses and employs a 5-dB step size, it results in values 2.5 dB greater than the ISO (1961) values. Therefore, the procedure of Marshall and Jesteadt (1986) results in values that are 4 dB better than the ISO values. The procedure of Dirks *et al.* (1986) converges to the 79% point on the psychometric function, while the procedure of Marshall and Jesteadt (1986) converges on the 71% point. Based on the data of Marshall and Jesteadt (1986), the 79% point corresponds to an increase in threshold of 2.5 dB as compared to the 71% point (see their Fig. 2). Taking this into account, the procedure of Dirks *et al.* should result in thresholds that are 1.5 dB better than the ISO values. The fact that the step size used by Dirks *et al.* is larger by almost 1 dB than the step size of Marshall and Jesteadt should further decrease this difference and render it insignificant.

[3]As in the original AI studies, the X values in this table were derived using Eq. (3). Therefore, they should be understood to represent the threshold spectrum density at all frequencies in the specified band. They should not be interpreted to represent the value at the center frequencies of the bands only, nor should the values between the center frequencies be obtained by interpolation. Under some circumstances, the latter may result in a very serious overestimation of the X values integrated over the band of interest.

[4]In the octave-band AI procedure, speech and noise spectrum densities based on the total octave power should ideally be used. The X values in column C (Table V) are valid only over one critical band (approximately 1/3 octave band). Thus more accurate X values than those specified in Table V could be obtained by averaging the corresponding 1/3-oct X values on the power basis as follows:

$$X = 10\log\left[\left(\sum_{i=1}^{3} 10^{0.1X_i} B_i\right)\Big/B\right],$$

where X_i is the X value from the Table V for the ith 1/3 octave within a given octave, B_i is the bandwidth of the ith 1/3 octave, and B is the bandwidth of the octave. Applying this equation, we obtain for the octaves centered at 250–8000 Hz the following values of X, respectively: -3.9, -9.7, -12.7, -17.9, -24.4, and -7.1 dB SPL.

[5]At the same center frequencies, speech spectrum levels that correspond to 1/3 octave bands in one instance and critical bands in another were found to be virtually identical. [These two are theoretically different because $S(f)$ represents the average intensity within the 1/3 octave band, in one intance, and within the critical band, in the other.] When a difference was observed, the average value was taken. In doing this, the maximum error made was 0.1 dB. This is a small price to pay in order to be able to treat the critical band AI method and 1/3 octave AI method together, both in the table and in the further analysis. Ideally, in the octave-band AI procedure, speech and noise spectrum densities based on the total octave power should be used. Therefore, the $S(f)$ should optimally be obtained by averaging the corresponding 1/3-octave $S(f)$ values on the power basis, as was explained in the case of the threshold in footnote 4. Again, given the large margin of error inherent in the 1-octave procedure, it is suggested that this not be done. Thus the 1-oct procedure will not be set apart from the other two procedures. The accurate $S(f)$ values that correspond to the octaves centered at 250–8000 Hz are, respectively: 35.0, 34.2, 26.2, 17.8, 9.8, and 1.1 dB SPL. Therefore, the maximum difference between these values and those from Table V is 0.6 dB.

[6]These include: Stevens *et al.* (1947); Rudmose *et al.* (1948); Western Electro-Acoustic Laboratory (1959); Harris and Waite (1965); Niemoller *et al.* (1974); De Gennaro *et al.* (1981); and Cox (1983).

[7]The mean noise spectrum over the band of interest is found in exactly the same way.

[8]If needed, the hearing level of noise is calculated in the same way.

[9]At the same center frequencies, $P(f)$ that correspond to 1/3 octave bands, on one hand, and critical bands, on the other, were found to be virtually identical. When a difference was observed, the average value was taken. In doing this, the maximum error was 0.2 dB.

[10]It is of interest to point out that it is not possible to use one importance function for all speech materials and then compensate for the errors so

made by instituting compensatory changes in the transfer function. To illustrate this point, let the transmission systems X and Y be two ideal bandpass filters that do not overlap. Assume that the input speech signal is amplified so that the entire dynamic range is above threshold. In this example, the calculations will not be affected by the relative relationship of the speech energy in the two bands. Let us further assume that the articulation indexes for the systems X and Y that correspond to the "true" importance function are not equal, while those that correspond to the importance function actually used in calculations turn out to be equal. The question is whether a transfer function could be found so that the speech recognition scores for the two bands are predicted to be different. The answer is clearly no. The transfer function is not a function of frequency, but only of the magnitude of the AI. Thus, if two AIs are equal to each other, the predicted scores will also be equal to each other.

ANSI (**1969**). ANSI S3.5-1969, "American national standard methods for the calculation of the articulation index" (American National Standards Institute, New York).

ANSI (**1978**). ANSI S3.21-1978, "American national standard methods for manual pure-tone threshold audiometry" (American National Standards Institute, New York).

ANSI (**1984**). ANSI S1.6-1984, "American national standard preferred frequencies, frequency levels, and band numbers for acoustical measurements" (American National Standards Institute, New York).

Benson, R. W., and Hirsh, I. J. (**1953**). "Some variables in audio spectrometry," J. Acoust. Soc. Am. **25**, 499–505.

Beranek, L. L. (**1947**). "The design of speech communication systems," Proc. IRE **35**, 880–890.

Berger, E. H. (**1981**). "Re-examination of the low-frequency (50–1000 Hz) normal threshold of hearing in free and diffuse sound fields," J. Acoust. Soc. Am. **70**, 1635–1645.

Black, J. W. (**1959**). "Equally contributing frequency bands in intelligibility testing," J. Speech Hear. Res. **2**, 81–83.

Boothroyd, A. (**1978**). "Speech perception and sensorineural hearing loss," in *Auditory Management of Hearing-Imparied Children*, edited by M. Ross and T. G. Giolas (University Park, Baltimore, MD), Chap. 4, pp. 117–144.

Boothroyd, A. (**1986**). Personal communication.

Burns, W., and Hinchcliffe, R. (**1975**). "Comparison of the auditory threshold as measured by individual pure tone and by Békésy audiometry," J. Acoust. Soc. Am. **29**, 1274–1277.

Byrne, D. (**1977**). "The speech spectrum—some aspects of its significance for hearing aid selection and evaluation," Br. J. Audiol. **11**, 40–46.

Carhart, R., and Jerger, J. J. (**1959**). "Preferred method for clinical determination of pure-tone thresholds," J. Speech Hear. Disord. **24**, 330–345.

Churcher, B. G., and King, A. J. (**1937**). "The performance of noise meters in terms of the primary standard," J. Inst. Electr. Eng. **81**, 57–90.

Cox, R. M. (**1983**). "Using ULCL measures to find frequency/gain and SSPL90," Hear. Instrum. **34**, 17–21, 39.

Cox, R. M., and McDaniel, D. M. (**1986**). "Reference equivalent threshold levels for pure tones and 1/3-oct noise bands: Insert earphone and TDH-49 earphone," J. Acoust. Soc. Am. **79**, 443–446.

De Gennaro, S., Braida, L. D., and Durlach, N. I. (**1981**). "A statistical analysis of third-octave distributions," J. Acoust. Soc. Am. Suppl. 1 **69**, S16.

Dirks, D. D., Bell, T. S., Rossman, R. N., and Kincaid, G. E. (**1986**). "Articulation index predictions of contextually dependent words," J. Acoust. Soc. Am. **80**, 82–92.

Duggirala, V., Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1986**). "Band importance functions for certain consonant features," J. Acoust. Soc. Am. Suppl. 1 **79**, S23.

Dunn, H. K., and White, S. D. (**1940**). "Statistical measurements on conversational speech," J. Acoust. Soc. Am. **11**, 278–288.

Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, Princeton, NJ), p. 101.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Harbert, F., and Young, I. M. (**1966**). "Amplitude of Békésy tracings with different attenuation rates," J. Acoust. Soc. Am. **39**, 914–919.

Harris, C. M., and Waite, W. M. (**1965**). "Measurements of speech spectra recorded with a close-talking microphone," J. Acoust. Soc. Am. **37**, 926–927.

Hawkins, J. E., and Stevens, S. S. (**1950**). "The masking of pure tones and speech by white noise," J. Acoust. Soc. Am. **22**, 6–13.

ISO (**1961**). ISO/R 226, "Normal equal-loudness contours for pure tones and normal threshold of hearing under free field listening conditions" (International Organization for Standardization, Switzerland).

Kryter, K.D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Kryter, K. D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1702.

Loye, D. P., and Morgan, K. F. (**1939**). "Sound picture recording and reproducing characteristics," J. Soc. Mot. Pict. Eng. **27**, 631–647.

Marshall, L., and Jesteadt, W. (**1986**). "Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures," J. Speech Hear. Res. **29**, 82–91.

Miller, G. A., and Nicely, P. E. (**1955**). "Analysis of perceptural confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Niemoller, A. F., McCormic, L., and Miller, J. D. (**1974**). "On the spectrum of spoken English," J. Acoust. Soc. Am. **55**, 461.

Pavlovic, C. V. (**1984**). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," J. Acoust. Soc. Am. **75**, 1253–1258.

Pavlovic, C. V., and Studebaker, G. A. (**1984**). "An evaluation of some assumptions underlying the articulation index," J. Acoust. Soc. Am. **75**, 1606–1612.

Pavlovic, C. V., Studebaker, G. A., and Sherbecoe, R. L. (**1986**). "An articulation index based procedure for predicting the speech recognition performation of hearing-impaired individuals," J. Acoust. Soc. Am. **80**, 50–57.

Pearson, K. S., Bennett, R. L., and Fidell, S. (**1976**). "Speech levels in various environments," BBN Rep. No. 3281 (Bolt, Beranek and Newman).

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Robinson, D. W., and Dadson, R. S. (**1956**). "A redetermination of the equal-loudness relations for pure tones," Br. J. Appl. Phys. **7**, 166–181.

Rudmose, H. W., Clark, K. C., Carlson, F. D., Eisenstein, J. C., and Walker, R. A. (**1948**). "Voice measurements with an audio spectrometer," J. Acoust. Soc. Am. **20**, 503–512.

Scharf, B. (**1970**). "Critical bands," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), Vol. 1, Chap. 5, pp. 159–202.

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656.

Sivian, L. J., and White, S. D. (**1933**). "On minimum audible sound fields," J. Acoust. Soc. Am. **4**, 288–321.

Stevens, S. S., Egan, J. P., and Miller, G. A. (**1947**). "Methods of measuring speech spectra," J. Acoust. Soc. Am. **19**, 771–780.

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138.

Western Electro-Acoustic Laboratory (**1959**). "Study and investigation of specialized electroacoustic transducers for voice communication," U.S. Air Force Contract No. AF33 616–3710, Task No. 43060, pp. A4–4 (National Technical Information Service, Springfield, VA).

Zwicker, E. (**1961**). "Subdivision of audible frequency range into critical bands," J. Acoust. Soc. Am. **33**, 248.