TECHNICAL REPORT

ISO TR 4870

First edition
1991-12-15

Acoustics — The construction and calibration of speech intelligibility tests

Acoustique — Élaboration et étalonnage des tests d'intelligibilité de parole

Reference number
ISO/TR 4870:1991(E)

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The main task of technical committees is to prepare International Standards, but in exceptional circumstances a technical committee may propose the publication of a Technical Report of one of the following types:

— type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

— type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

— type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

ISO/TR 4870, which is a Technical Report of type 3, was prepared by Technical Committee ISO/TC 43, *Acoustics*.

It contains data which are valuable in speech intelligibility testing, but it is not expected to become an International Standard.

Annexes A and B of this Technical Report are for information only.

International Organization for Standardization
Case Postale 56 • CH-1211 Genève 20 • Switzerland
Printed in Switzerland

Introduction

A variety of perceptual tests have been developed in the past for the assessment of the intelligibility of speech communications as affected by spectral, amplitude, and temporal distortions of the speech signal and by noise that arises from or in the acoustical, electrical (if any), and ear receptor path used for transmitting and transducing speech from a talker to a listener. The principal tests developed for this purpose have been called speech intelligibility tests and will be later defined in detail.

Beyond factors related to the talkers, listeners, and a given communication system, there are two factors common to all speech intelligibility tests that have a significant influence on the scores obtained from a test evaluation. These two common factors are: (1) the speech material employed in the tests, and (2) for a given type of speech material, the total number of alternative members of that material the listeners expect to be presented during a test. Without some knowledge about the contribution of these two factors to the scores obtained on a given speech intelligibility test meaningful comparisons and interrelations cannot be made with respect to the test scores obtained in different investigations of speech intelligibility.

It is the purpose of this document to standardize fundamental methods for the construction and calibration of speech intelligibility tests in ways that reveal the contributions to the test scores of the two common factors mentioned above. Also, illustrative examples of recommended types of speech test materials possibly suited for such purposes as speech audiometry, the evaluation of room acoustics, or an electro-acoustic transmission system are given.

The communication of thoughts and concepts through spoken languages is a broad and complex operation that is influenced by many other factors than the intelligibility based on the perception of acoustical features of the speech signal. However, the basic feature of speech communication is an acoustical signal, and the greater the understanding of the speech derivable from perceptions of certain acoustical relations within the signals, the more effective and general can be the communication process. Intelligibility tests are aimed at the evaluation of the ability of a communication system or component, including the auditory mechanism of a listener, to effectively transmit basic acoustical information that is instrumental to the correct perception of speech.

This page intentionally left blank

Acoustics — The construction and calibration of speech intelligibility tests

1. SCOPE

1.1 The present document is concerned with the description of:

(a) methods for the construction of speech tests for the measurement of the intelligibility of speech transmitted by an analog or combination analog and digital communication system;

(b) a reference communication system and test conditions necessary to the development, calibration, and interpretation of the results of standardized intelligibility tests.

1.2 The description of specific speech tests and related test procedures and measurements that are most appropriate for a given test application are beyond the scope of this document.

2. DEFINITIONS FOR PRESENT PURPOSES

2.1 Speech Sound

A speech sound is the smallest identified unit of speech. These units of speech can be categorized into two general classes known as vowels (V) and consonants (C). Consonants that are, on occasion, used as vowels in forming syllables of speech will, for purposes of this document on those occasions, be classified and included under the label V.

2.2 Syllable

A syllable is a pronounceable unit of speech consisting of a vowel or a combination of a vowel with one or more consonants.

2.3 Polysyllable

A polysyllable is a series of more than one syllable pronounced with liaison between syllables.

2.4 Word

A word is a monosyllabic or polysyllabic unit of speech that has an accepted meaning to the listeners.

2.5 Logatom

A logatom is a monosyllabic or polysyllabic unit that has no meaning to the listeners.

2.6 Test Sound

A particular speech sound to be used in accordance to defined rules to form test items.

2.7 Test Item

A particular monosyllabic or polysyllabic logatom, or word, to be used in accordance with defined rules for intelligibility measurements.

2.8 Set of Test Sounds

The particular test sounds that have been taken from the total set of possible (in a certain language) or given (according to major frequency in a certain language or according to certain rules) sounds, to form test items. The set of test sounds often is subdivided, dependent on the position of the proper test sound in the test item, into sets of initial, central, and final test sounds.

2.9 Set of Test Items

The particular logatoms or words that have been selected according to defined rules from the total amount of possible or given logatoms or words to be used for intelligibility measurements.

2.10 Subset Item

A speech sound, logatom, or word, to be used in accordance with defined rules for intelligibility measurements.

2.11 Phonemic Structure of Speech

Phonemic structure refers to the relative frequency of occurrence of different speech sounds and their positions relative to other speech sounds, in the syllables and words of a certain language.

2.12 "Phonetically Balanced" Lists

So called "phonetically balanced" (correct definition: phonemically balanced) lists are achieved when each list contains about the same proportion of the various classes or types of speech sounds as are

found to be, or presumed to be, present in specified speech communication with a given language.

2.13 Test List

A number of specially selected test items presented and scored as a single test. Typically, for open or pseudo-open test lists, a relatively large set of items is divided into a number of lists each containing an equal number of test items. Typically for the closed-set lists a number of subset ensembles are grouped together on one list.

2.14 Open Test List

Open lists of test items are made of items drawn randomly from the total set each time a list of test items is to be presented to listeners. Typically a listener writes on an answer sheet each test item the listener believes was presented.

Note: In order to insure that the phonemic structure in the total set appears properly in the test lists, it is necessary not to replace the items drawn randomly for one test list back into the total set of items prior to the random selection of items for succeeding lists.

2.15 Pseudo-open Test List

Pseudo-open lists of test items are made of items drawn on the basis of some specified set of phonemic rules, from the total set of items. The groupings of items within each test list, but not their sequential order, thus drawn is maintained for successive uses of the lists. Typically, a listener writes on an answer sheet each test item the listener believes was presented.

2.16 Repeat Test Items

Items within each open and pseudo-open test lists that are presented more than once within a list.

2.17 Pseudo-open List Scrambling

The items assigned to each pseudo-open test list are reordered on a random basis within each scrambling of each test list, to provide a number of sequences of items for each test list which are novel, or seem to be novel, to the listeners.

2.18 Closed-Set List

Closed ensembles of the order of 2 to 10 items per subset, are displayed visually to listeners during the test. One item of each subset is presented acoustically to the listener during a test, at which

time the listener indicates, typically by a check mark on an answer sheet, which item of the visually displayed subsets involved was most probably presented acoustically.

Note: The subset ensemble is characterized by one speech sound that is the nucleus of every test item in it. All the test items in a given ensemble are initiated (or terminated) by the same speech sound and are terminated (or initiated) by different speech sounds.

2.19 Apparent Message Set Size

Apparent message set size refers to the number of alternative answers (to the presented items) presumed by a listener to be available as possibly correct answers to each item presented during an intelligibility test on the basis of the listener's knowledge of the total number of test items available to the speaker for presentation.

2.20 Real Message Set Size

Real message set size refers to the number of possible alternative answers by a listener to each item presented during an intelligibility test on the basis of the total number of items having audible phonemic similarities with each test items and which are within the set of test items available to the speaker for presentation.

2.21 Intelligible Speech Sound, Logatom, or Word

A speech sound logatom or word is defined as being intelligible when it is correctly perceived by a listener.

2.22 Percent Speech Intelligibility

Percent speech intelligibility is the percentage of items on a list correctly identified by a listener or group of listeners corrected for chance identifications dictated by the number of alternative answers per item available to the listener. This number is taken for the open or pseudo-open test lists to be the number of items in the total message set from which the test lists are drawn; for the small closed set lists, this number is taken to be the number of subset items or alternatives in a sub-set (note, not list size). In formula this can be expressed as follows:

$$I \text{ in } \% = \frac{100}{T} \left(R - \frac{W}{N-1}\right)$$

where T is the number of items in test and N is number of alternatives to each item. R is number of items right, W the number wrong. The last term is the correction for chance in item identification.

Note 1: By chance is meant that the listener is able to correctly guess a certain number of test items inasmuch as the listener knows, because of training or test format, the identity of all the possible alternative answers for each test item presented. For example, if the message-set consists of but 5 words, the listener would, on the average, score one out of five correct, or 20%, merely by guessing the identity of each test item.

Note 2: Under good listening conditions and high intelligibility scores, the size of the real, as opposed to the apparent set size of open or pseudo-open test list format, is of minor concern, because, as reflected in the last term of the formula for calculating percent speech intelligibility, the correction for chance is negligible when most items are correctly perceived. As the listening conditions and, accordingly, the intelligibility scores are degraded, the real message set size approaches the apparent size; i.e., the number of alternative responses is perceived as being much larger in number than is the case under good listening conditions. For tests in which the number of apparent alternative answers to each test item presented to the listener is greater than about 50, the correction for chance becomes negligible and percent speech intelligibility can be taken as the percent items correct on a test.

Example 1. If there are 50 test items on an open or pseudo open test list with each item having 1000 alternatives and 26 of the 50 test items were answered correctly and 24 of the items were answered incorrectly, the percent of speech intelligibility would be 52%. (100/50 (26 - 24/1000)= 51.952% , or, to round off, 52%.)

Example 2. If there are 50 small-closed set test items each consisting of 5 alternative subset items, and 26 of the 50 test items were answered correctly and 24 of the items were answered incorrectly, percent speech intelligibility would be 40%.(100/50 (26 - 24/4) = 40%.)

2.23 Carrier Sentence or Phrase

A sentence or phrase of at least 4 words and that contains a test item but such that the correct understanding of the test item is not dependent upon the context or meaning of the sentence in which it it embedded.

Note 1: The purpose of the carrier sentence is to provide: (1) the talker with means of enunciating the words in a natural manner and a controlled and measurable level of effort; (2) a regular temporal separation of test items of sufficient duration to permit listeners to decide and record

their answers to each perceived test item; and (3) to provide a "steady" stream of speech sounds that would be natural and necessary to provide operation of certain electronic devices, such as automatic gain controls, and/or the acoustic reverberations that would be present in a room.

Note 2: An example of an English carrier sentence used in some speech intelligibility tests is "You will mark (or write) (test item) now," It is important that the speech sound immediately preceding the test items be pronounceable without liaison with the test items, otherwise a variable interaction between that sound and different test items will occur and influence the perception of the test item.

2.24 Vocal Effort of Talker in Terms of Measured Sound Level of Speech

The vocal effort used by the talker in a speech intelligibility test is measured in terms of the arithmetic average of the maximum sound level reached during each of the test items, or the words of the carrier phrase, respectively (see 3.7 below). The sound pressure level will be A-weighted and measured with a sound level meter complying with IEC 651 type 1, set on S characteristic and observed at, or referred to, a point 1 meter in front of, and level with, the talker's lips when speaking in a free-field, or effective (in terms of there being no adverse reverberation effects on the understandability of the speech) free-field being present at that position.

2.25. Rate of Talking

The carrier sentence or phrase and the test items will be uttered by the talker in a normal fashion. Normally continuous speech is uttered at a rate of approximately 5 syllables per second.

2.26 Idealized Speech Spectrum

Figure 1 shows the idealized spectrum level of male voices at the level typical for everyday talking and listening conditions.

Note: The average of the maxima SPL, A-weighted, slow meter, per word of conversation typically equals 65 dB at one meter in front of the talker in a business office environment , and 55 dB for conversations in the home.

2.26.1 Table of Idealized Speech Spectrum

The spectrum level relative to 400 Hz of the idealized speech shown in Fig. 1 is as follows at the frequencies specified.

125 Hz	-6.0 dB
250 Hz	-1.0 dB
400 Hz	0 dB
500 Hz	+0.5 dB
1000 Hz	-10.0 dB
2000 Hz	-22.0 dB
4000 Hz	-34.0 dB
6300 Hz	-43.0 dB

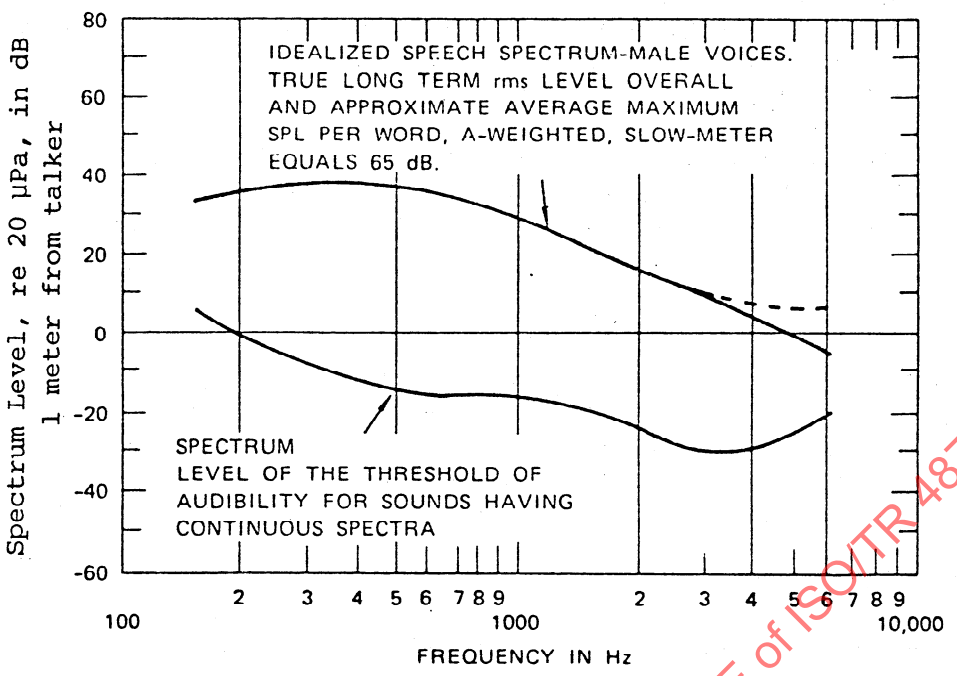

FIGURE 1 IDEALIZED SPEECH SPECTRUM AND SPECTRUM LEVEL OF AUDIBILITY FOR CONTINUOUS SPECTRA SOUNDS. SPEECH LEVEL SHOWN IS FOR TYPICAL EVERYDAY LISTENING AND TALKING CONDITIONS AND IS ABOUT 10 dB HIGHER THAN THE LEVEL FOUND UNDER QUIET CONVERSATIONAL CONDITIONS.

The speech spectrum shown by the solid, to 2500 Hz and dashed line above 2500 Hz has been incorporated into a standard for the calculation of the Articulation Index (Ref.1). The solid curve from 125 to 6300 Hz is deemed to be more proper on the basis of recent studies (Ref. 2) and is the idealized speech spectrum recommended for present purposes.

2.27 Speech Level at the Listeners' Ears

The speech level with no noise present and with earphone listening is to be expressed as the arithmetic average of the maximum sound levels (frequency weighted A, time weighted S) reached during each test time. This level is to be estimated from coupler calibrations as specified in IEC Publication 318.

2.28 Speech-shaped Masking Noise

Speech-shaped masking noise is defined as random white noise filtered such that its spectrum level falls within ± 1 dB over the frequency range of 125 to 6300 Hz of the idealized speech spectrum for male voices shown by the upper solid curve in Fig. 1 and the values shown in table of 2.26.1, except that this will fall off at the rate of at least 6 dB below 125 Hz and above 6300 Hz.

Note 1: This spectrum shape can be approximated with a third order filter.

Note 2: One purpose of this definition is to provide a noise spectrum that is somewhat representative of the everyday real-life noises, including the babble of many voices, that often interfere with speech communications.

Note 3: A second purpose of using a speech-shaped masking noise is that it equally interferes, on a temporal-average basis, with speech in all regions of the acoustical spectrum.

2.29 Noise Level at the Listeners' Ears

The noise level, with no speech present and with earphone listening is to be expressed as the arithmetic average of sound levels (A-weighted, slow meter action) reached during the moments when each test item would be present. This level is to be estimated from standard coupler calibrations as specified in IEC Publication 318.

2.30 Speech-to-Noise Ratio

Speech-to-noise ratio is the numerical difference between the sound level of the speech signal compared to the sound level of the noise when each are measured separately and at or referred to the same acoustical or electrical point in a communication system.

2.31 Non-distorting Reference Speech Transmission System

For present and practical purposes a transmission system will be deemed free of any significant distortion effects upon the transmitted speech signal when: (1) its frequency response characteristics are uniform (± 2 dB) over the frequency range of 125 to 6300 Hz; (2) the noise floor, on a spectrum level basis, is at least 40 dB below the spectrum level of undistorted speech at all frequencies from 125 to 6300 Hz for the speech at its average A-weighted sound pressure level, slow meter action (see 2.27 and 2.28 above); and (3) harmonic distortion does not exceed 1% with pure-tone input signals at frequencies from 125 to 6300 Hz at input levels, A-weighted, that exceed by up to 18 dB the input speech signal. This reference transmission system is to be capable of 50 dB linear dynamic range with signal and noise each measured on a spectrum level basis and providing speech to the listener's ears at a sound level of 65 dB, as defined in 2.27 above.

3. RECOMMENDED TYPES OF TEST MATERIALS AND PROCEDURES FOR THEIR PREPARATION AND CALIBRATION

3.1 Large Set Tests

One type of intelligibility test recommended for evaluating the effectiveness of systems used for normal speech communication requires the use of a set of at least 1000 meaningful words or at least 650 logatoms presented for conducting the test under several test conditions. It is recommended that the chosen large set of items be separated into pseudo-open lists of at least 50 items each, each list should contain the same number of each type of phonemes in similar positions in the words or logatoms and in about the same overall proportions as can be best estimated, and achieved with the

chosen test items, for the everyday speech of the language. Each list should contain at least one item that is presented more than once within the list.

Note 1: Because of differences among languages with respect to the relative number of monosyllabic and polysyllabic words occurring in the language, it is not feasible to recommend that only one type of syllabic structure (monosyllabic or polysyllabic) be utilized in the construction of these tests.

Note 2: The choice as to whether open-set or pseudo-open set test lists are to be constructed for a given test situation depends on how carefully the user wishes to maintain balance in the test evaluation, the number of systems or variables to be evaluated, and the amount of testing time or expense that can be devoted to the test evaluation.

3.2 Small Closed-Sets

A second type of test recommended for speech intelligibility utilizes the small closed-set format. In this format test items are organized into lists of small subsets (the subset size can vary for different test materials, typically ranging from 2 to 10 alternatives) in which all but one of the speech sounds, and always at the same position within the syllable, is varied. It is recommended that the number of chosen small closed subsets be sufficiently large to test all, or nearly all, consonantal types, and syllabic positions, with at least several different vowels, and that each subset contain as many alternative response answers as feasible.

Note 1: This test format can be responded to easily by the listener because the subsets can be visually displayed so that the listener merely checks or indicates which item of the visually displayed subsets the listener thinks corresponds to the item presented audibly.

Note 2: As a result of the forced choice format with the displayed answers available to the listeners, the training time is a matter of but a few minutes before performance reaches a level that remains consistent, with a given communication system and test conditions, for an indefinite period of time for each listener. Accordingly, the test-retest reliability is very good.

Note 3: It is to be noted that the small message set format permits one to readily identify the confusions that occur among the phonemes and that this information is sometimes of value in the diagnosis of the ability of a speech communications system to transmit specific phonemic characteristics of speech.

3.3 Calibration Function with Reference System and Masking Noise

For present purposes, the relation between percent speech intelligibility test scores and signal-to-noise ratios at the listeners' ears, is to be called a "speech intelligibility test calibration function". It is recommended that the reference transmission system defined in paragraph 2.31, the noise defined in 2.28 and the speech as defined in 2.26, be used as the standard for the calibration of all speech intelligibility tests.

3.4 Noise Mixing

The noise is to be electrically mixed with the speech signal prior to its transduction to an acoustical form for presentation to the listener. The speech will be presented at a level of 65 dB (see 2.27) and the noise for different tests, at levels of 71, 65, 59, 53 and 47 dB, or speech-to-noise ratios of -6, 0, +6, +12, and +18 dB (see 2.30). A schematic block diagram of this reference calibration test system is given in Figure 2.

3.5 Talker(s)

The training and tests will be conducted with at least one male and one female talker, each having speech deemed by test or expert opinion to be typical of a given nationality and language.

3.6 Recording of Tests

The tests will be recorded with the talker in a part of the field in which the reverberant sound is negligible and with a microphone and on a medium that is free of amplitude distortion, has harmonic distortion of less than 1 percent, dynamic range of at least 50 dB and a frequency response ± 2 dB between 125 and 6300 Hz.

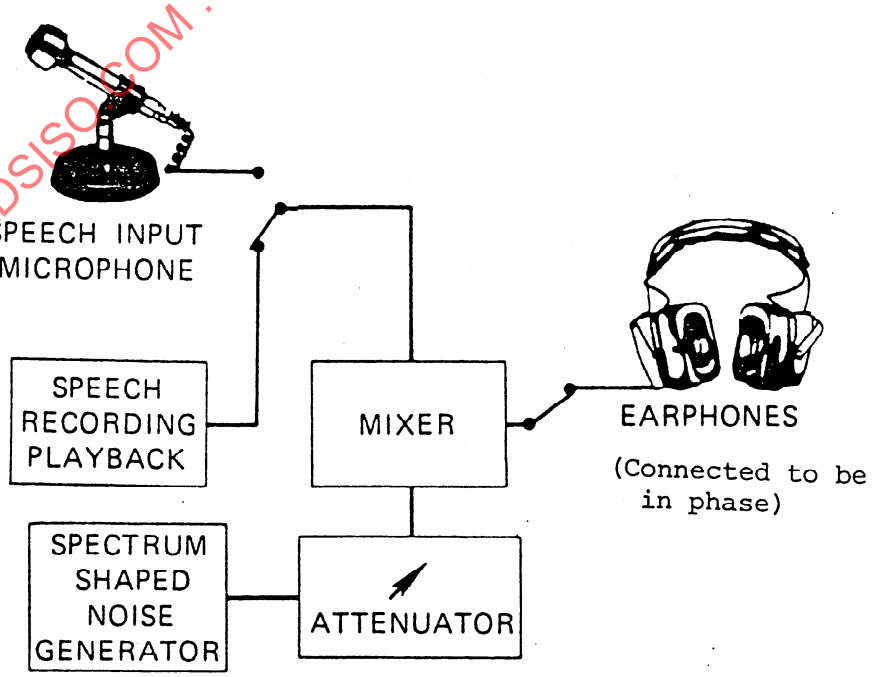

FIGURE 2 SPEECH TRANSMISSION SYSTEM WITH MASKING NOISE SOURCE FOR CALIBRATION OF SPEECH INTELLIGIBILITY TESTS

3.7 Voice Level Monitoring

For purposes of these calibration function tests, the talker will use a level of vocal effort that would provide a level of intensity that measures on the average 65 dB ± 3 for the words of the carrier sentence or phrase, at a distance of 1 meter in front of the talker's lips. The vocal effort of the talkers will be measured or estimated in accordance to definition 2.24.

Note 1: It is recognized that the arithmetic average of the SPL of the individual test items may not be the same as that of the words of the carrier sentence or phrase. The purposes of the carrier sentence or phrase are given in 2.23.

Note 2: A carrier sentence or phrase is recommended for the purpose of calibrating a speech intelligibility test and can be recommended for most intelligibility testing purposes. However, there are test applications where the carrier sentence or phrase may not be deemed necessary or appropriate. If a speech intelligibility test is to be constructed for use without a carrier sentence or phrase, then the calibration of that test will also be conducted without the use of a carrier sentence or phrase.

3.8 Listener Crew Selection for Calibration Testing

The listeners, male or female, shall be selected on the basis of having good hearing, normal experience in the use and spelling of the language to be used in the speech tests, and be capable in other respects of normal performance in test learning and participation. Good hearing for present purposes is said to be present in a person with a pure-tone audiogram that does not exceed a hearing level of 10 dB at any test frequency up to 4000 Hz and 15 dB at any test frequency up to and including 6000 Hz.

3.9 Crew Size for Calibration Testing

It is recommended that a listening crew be sufficiently large so that the variability of test scores is found to not decrease with increase in crew size. It is found that this will usually occur with a minimum of 5 well-motivated, capable and well-trained crew members when small closed message set test format is used and 10 members when the large pseudo-open set test format is used. More crew members may be required if the crew members are less than well motivated to obtain reliable test results. The degree of statistical reliability achieved for a given test series should be calculated (see also 3.14 below).

3.10 Talker and Listener and Testing Time Training for Acoustical and Electrical System Testing

Table 2 summarizes the estimated time required for the training of typical talkers and listeners on the recommended speech tests in order to reduce to a near minimum the changes that may occur in successive presentation to the same test material due to learning of the materials and procedures. The allowed time for the testing of a crew of listeners following training is dictated by the fact that after the periods of time indicated, the same listeners will

begin to know, for the large pseudo-open set tests the order of appearance of items on the different scramblings of the test items as well as the grouping of items within each list (i.e., the lists are only pseudo-open to the entire large set of items).

TABLE 2 Typical Talker and Listener Crew Training and Testing Time

	Large Pseudo-Open Set Tests*		Small Closed Set Tests	
	Logatom	Words	Logatom	Words
Typical talker training time	8 hours	1 hour	2 hours	5 minutes
Typical Listener training time	24 hours	12 hours	2 hours	5 minutes
Typical permissible testing time following training of listener.	200 hours	120 hours	No limit	No limit

Note 1: The first one-half of the training time for the large pseudo-open set tests should be conducted with a speech transmission system and noise condition that permits perfect or near perfect speech intelligibility and the second half of the training time should be conducted with a speech transmission system and/or noise condition that gives speech intelligibility scores that do not, on the average, exceed 50% correct for the large pseudo-open set tests and 70% for the small closed set tests.

Note 2: It is recommended that, in general, speech intelligibility training sessions with the same listening crew not exceed 2 hours per morning and 2 hours per afternoon and that within each 2-hour session, rest periods from the training of about 5 to 10 minutes be provided about every 30 minutes.

* Based on approximately 20 lists of 50 test items, each list internally scrambled.

3.11 Number of Tests

At the completion of the training tests, a minimum of 3 different lists per talker of 50 items each for large set tests, or 3 different lists of 50 sub-tests per lists of small, closed set tests will be administered to the listening crew at each of the signal-to-noise ratios specified in 3.4 . The different conditions of signal-to-noise ratios and talkers will be presented in a random order sequence with no more than 20 test lists presented within a three-hour test session, including rest periods for the listeners for about 10 minutes each between test periods of about 20 minutes each.

3.12 Scoring

Percent intelligibility scores will be found for each test as given in accordance with paragraph 2.22. The percent intelligibility scores for each type of test under calibration will then be averaged for each talker and each signal-to-noise ratio. These results will be plotted on graph paper with curves of best fit interpolated and extrapolated to the data points.

3.13 Estimated Calibration Function for Reference System and Masking Noise

Figure 3 illustrates the general nature of the final, smoothed curves to be expected with the subject reference speech system and masking noise for intelligibility tests utilizing different types of speech materials. The functions shown in Figure 3 are drawn on the basis of general conclusions of speech intelligibility tests conducted with English language test materials. However, it is believed that similar relations will be found with similar tests made from nearly all other languages.

3.14 Statistical Measures of Variability

In addition to the calibration function representing the average results for the entire crew of listeners and talkers, statistical measures, including the means and standard deviations, of the results for given test conditions, listeners and talkers should be calculated and reported.

Note: Statistical methods to take into account, to some extent, variability in test scores generally introduced by different states of training of the test subjects and to reduce the errors of asymmetrical variances of the test results near the extremes of 0% and 100% correct are described in Ref. 3.

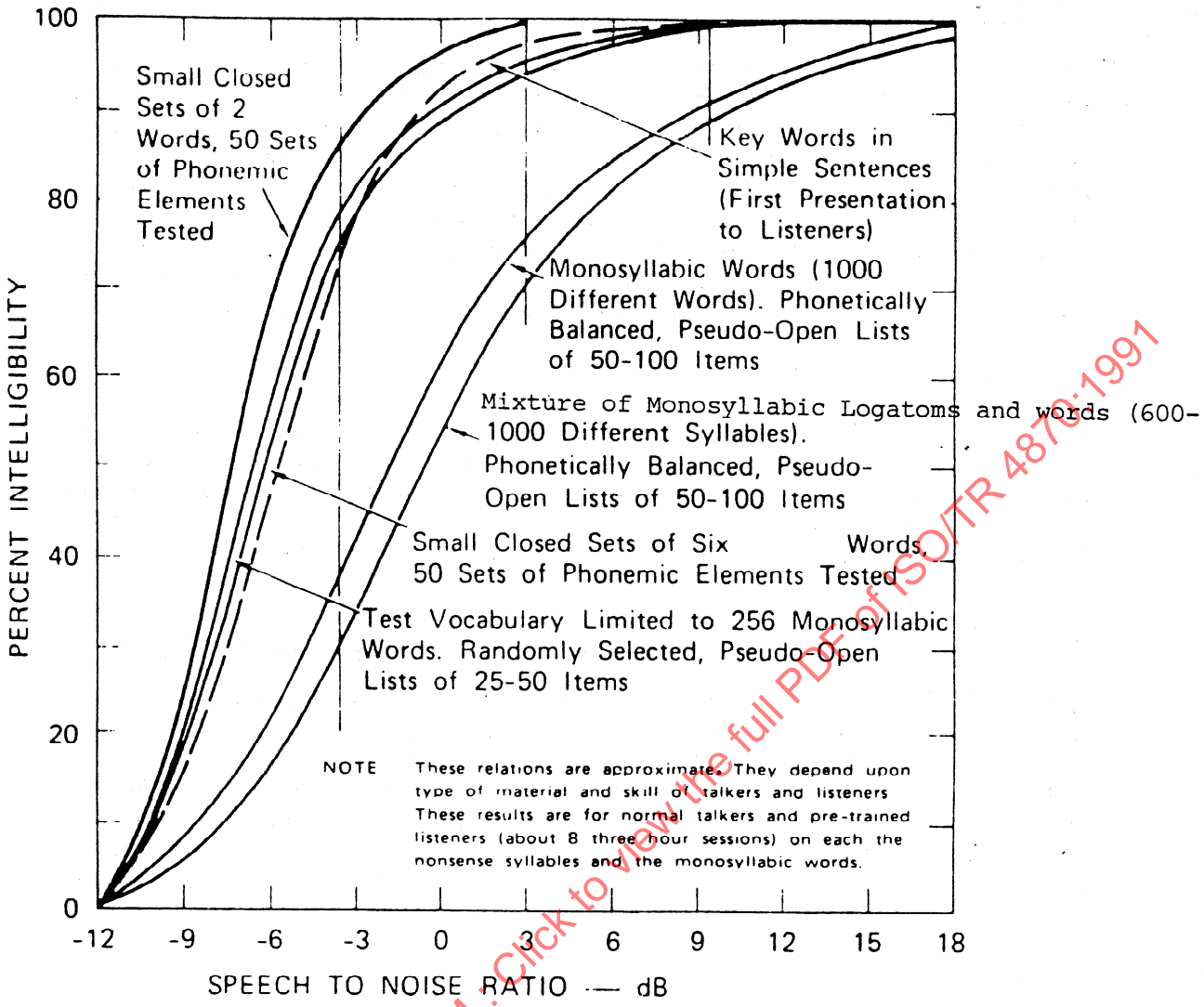

Figure 3 Examples for certain English Language tests, of the general relation between percent intelligibility and wide-band speech-to-noise ratio.

The noise is present at the input to the communication system. The speech signal is undistorted, at a level of 65 dB, A-weighted sound pressure level. The noise is steady, with a speech-shaped spectrum and measured with A-weighting based on Ref. 1.

Annex A
(informative)

Supplemental Information Relative to the Construction of a Speech Intelligibility Test and Illustrative Examples

A.1 Acoustical and Non-acoustical Factors in Intelligibility Testing and Speech Communications

It is to be recognized that the understanding of the elements of speech that is based primarily upon acoustical features as measured by a given intelligibility test, is influenced by the listener's concept of the set, frequency of occurrence, and arrangement of the different elements included in the test. Concern for this psychological factor within the listener is important to the proper construction and interpretation of speech intelligibility tests.

Note 1: The purpose of the pseudo-open test lists, as distinct from the open test lists, is to make convenient the obtainment and maintenance of some desired condition of phonemic balance and difficulty during a comparative testing series. This phonemic balance and difficulty of test items generally permits the administration of a fewer number, for valid results, of test lists per test condition than is possible with the use of the open test list format. This convenience and reduction, for a given degree of validity and discrimination, of testing time with the pseudo-open lists has a detrimental aspect in that after prolonged periods of testing of the same listeners, the listeners will tend to learn the associations among words of the lists, even though the order of the items within each list are "scrambled" a number of times. This delayed learning, when it occurs, reduces subsequently reliability and discrimination of the pseudo-open test lists with those listeners.

Note 2: The alternative to the pseudo-open test list is the open test list. As indicated in 2.14, in the open test list, the test items are drawn randomly from the total set of test items, without replacement of items following the construction of each list. This procedure requires the use for each test condition of more lists than are generally required with the pseudo-open set test lists because the phonemic equivalence of the randomly constructed, or truly open test lists, cannot be assured by this selection process. Accordingly, it is necessary, in order to be certain that each test condition is evaluated with phonemically comparable speech material, to administer a greater number of open set lists than pseudo-open set lists.

Note 3: The purpose of repeating a small number of items within each list is to reduce the propensity of listeners to reject the perception of an item that appears later in a test list because it had been perceived (perhaps falsely) as appearing earlier in the list. The fact that some number of unspecified items are repeated in each list is, of course, to be known by the listeners.

Note 4: The purpose of list scrambling is to prevent the listeners from memorizing sequence contingencies for the items and to reduce the possible effect on item responses than can come from the listeners learning, with repetitions of the lists, that individually the test lists are not truly open to the total message set. It is appropriate to also change the repeat test items for different scramblings of each test list.

Note 5: In order that the closed-set test be a reasonably sensitive measure or index of the intelligibility of normal speech it is necessary that each set contain as many items as may be confusable one with the other under normal conditions of speech communications and that as many message sets as are required to test all, or nearly all, phonemic sounds of the language are included on the test lists.

Note 6: The real message set size, but not the apparent, for each test item of tests of the open or pseudo-open list format varies significantly as functions of the noise and speech signal conditions present during each intelligibility test. For example, under adverse noise or frequency distortion conditions the number of items containing confusable phonemes increases relative to the number that are auditorially confusable under less adverse noise conditions.

Note 7: Because of the constraints on responses available to the listener with the small closed-set list format, the real message set size and the apparent set size remain about the same regardless of noise and speech signal conditions. The difficulty and discriminability of intelligibility tests is, for a given type of test material, a direct function of the real message set size, the larger the size the greater the difficulty and discriminative powers of the test.

Note 8: Because of: (1) normal language and test-length restraints upon the number of polysyllables suitable for intelligibility testing, and (2) the contingencies by the phonemic structures within and among the polysyllables to be employed, polysyllables tend to be relatively easy and nondiscriminative test items. That is to say, the apparent correct perception of one phonemic element is based upon the perception of some other element in

the polysyllable and the knowledge that certain phonemic elements are constrained to appear together in some polysyllable contexts; thereby, with polysyllables, the intelligibility is determined to some extent by non-acoustical factors and not solely on how well, as measured by intelligibility, the communication system articulates the relevant acoustical information.

Note 9: In some languages there are phonemic elements that appear only as parts of polysyllables, and to the extent that these elements may discriminate among communication systems in ways not done by monosyllabic words or logatoms, such polysyllables should be included as items in the intelligibility test.

Note 10: Lists can be constructed to contain various classes of phonemes, but not necessarily in the proportions found in the language as normally used for communications. It is appropriate to score such test lists according to the number of correct answers obtained for each class of phonemes and to then multiply or correct these scores in accordance with the proportions each class of phonemes is normally found in the given language.

A.2 Practical Compromises and Considerations in Test Development, Standardization and Usage

The choice of speech materials to be used for intelligibility testing must usually be a compromise to satisfy the following three goals: (1) ideally the speech test should present to the speech communication system units, for scoring, of speech material that are at least as difficult to correctly perceive as are the more difficult units (usually individual words) of speech material to be perceived from usage of the system in real life; (2) the speech test must be reliable so that the results of an evaluation of one system or condition of use can be compared to those found for other systems or conditions of use tested as part of the same test program or as tested in a separate test program; and (3) the training time and costs required for conducting the test are not prohibitive.

A.3 Selection of a Speech Intelligibility Test

Most speech intelligibility tests represent the compromises between some of the principles specified in clause 3 of the main text and various practical considerations related to the costs and time required to train the test subjects and to administer and score the tests. The monosyllabic, small closed set word test (see example below) appears to be the most practical of these tests, because of its ease of usage with untrained listeners and to be also particularly suited, for this reason, for the evaluation of the integrity of the sound transmission and transduction components of the ear. However, to be able to more precisely discriminate among the capabilities of different systems to transmit intelligible speech under a wide variety of conditions, the large open or pseudo-open set lists of monosyllabic words or logatoms are the best.

A.4 Diagnostic Articulation Test

Attention is also invited to a type of speech test called "phoneme discrimination", or "diagnostic" speech tests (Refs. 4 and 5). In these tests the listeners are presented a small-closed set of syllables with minimal phonemic associations in a format that permits the scoring of specific confusions between two selected phonemes that differ with respect to but one, so-called, distinctive feature. However, under conditions of everyday speech communications, confusions in the perception of phonemes are not restricted to those related to but one distinctive feature, and, accordingly, these tests tend to be inordinately easy for the listener. For example, with only a two-choice diagnostic test format, the listener will score 50% correct on the average, merely by guessing, whereas with normal speech a greater choice of possible speech sounds spring into the listeners' minds as possibilities and, therefore, cause a reduced intelligibility.

Accordingly, these diagnostic speech tests do not provide an index to the intelligibility of speech communications that is adequate to discriminate among speech transmission systems that differ in intelligibility under normal usage. Obviously, as outlined previously in the text, all structured intelligibility tests utilize speech material from restricted (compared to real-life communications) message sets but not to the extent to which this is done in the so-called diagnostic speech test.

A.5 Sentence Intelligibility Tests

The proper measurement of speech intelligibility, as herein defined, requires the use of phonemic contexts and balances and acoustic noise conditions representative of normal speech communication. On first thought, it might appear that phrases or sentences would provide appropriate test material for this purpose. However, there are two major reasons why phrases or sentences do not usually lend themselves for general speech intelligibility tests: first, as mentioned previously, the correct understanding of the speech sounds and words of phrases or sentences is significantly influenced by the knowledge the listener has of the grammar, syntax, and meaning of the ideas involved; inasmuch as these factors are present to different degrees in different listeners, their operation in a speech intelligibility test can obscure and confuse the evaluation of the abilities of different transmission systems to transmit and articulate understandable speech sounds. In brief, the purpose of the present intelligibility tests is not to measure the contribution of grammar, syntax, meaning, etc., to speech communications, but merely the contribution to speech communications that is made possible by the transmission characteristics, in terms of understandable speech sounds, to speech communications.

The second reason sentence intelligibility tests are generally not used is the practical problem of creating a sufficient number of lists of sentences that are phonetically representative of speech in general and yet of equal difficulty or familiarity to typical members of test listening crews. Specifically, test sentences, because of learning of the sentence structures, cannot be repeated with the same listening crew in order to obtain reliable or properly discriminative, comparative intelligibility measures of