



ARL-TR-8227 • DEC 2017



Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission

by Tomasz R Letowski and Angelique A Scharine

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission

by Tomasz R Letowski

Fellow and Researcher Emeritus, ARL

Angelique A Scharine

Human Research and Engineering Directorate, ARL

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<p>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) December 2017		2. REPORT TYPE Technical Report		3. DATES COVERED (From - To) 1 June 2017–29 September 2017	
4. TITLE AND SUBTITLE Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Tomasz R Letowski and Angelique A Scharine				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-HRF-D Aberdeen Proving Ground, MD 21005-5425				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TR-8227	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <p>Speech communication can be assessed in many ways. The objective of this review and analysis is to compare the common methods of evaluating the quality and intelligibility of speech, and detail the merits and limitations of each. The standard speech intelligibility rating scales, perceptual speech intelligibility tests (based on human performance), and technical speech intelligibility predictors (based on the input signal transmitted through a communication system or medium) measurement methods are described and compared. To establish a basis for comparison between the results of these measures, a common intelligibility scale is described. Its use in the comparison of scores obtained for different measures of speech intelligibility is discussed, as well as its use to determine which test is optimal for a given environment. This analysis is intended to serve as a resource for users of standard speech intelligibility measurement methods.</p>					
15. SUBJECT TERMS <p>speech communication, speech intelligibility, speech quality, perceptual measures, objective measures, common intelligibility scale</p>					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 58	19a. NAME OF RESPONSIBLE PERSON Angelique A Scharine
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) 410-278-5957

Contents

List of Figures	iv
List of Tables	iv
1. Introduction	1
2. Characteristics of the Speech Signal	1
3. The Concept of Speech Intelligibility	2
4. Speech Intelligibility and Speech Quality	6
5. Speech Intelligibility Assessment	8
5.1 Speech Intelligibility Rating	9
5.2 Perceptual Tests of Speech Intelligibility	13
5.2.1 Phonetically Balanced Word Test (PBWT)	15
5.2.2 Diagnostic Rhyme Test (DRT)	16
5.2.3 Modified Rhyme Test (MRT)	17
5.3 Technical (Predictive) Tests of Speech Intelligibility	18
5.3.1 Articulation Index (AI)	19
5.3.2 Speech Transmission Index (STI)	20
5.3.3 Speech Intelligibility Index (SII)	22
6. Common Intelligibility Scale	24
7. Selecting a Speech Intelligibility Test and Interpreting Its Score	27
8. Discussion and Conclusions	31
9. References	35
List of Symbols, Abbreviations, and Acronyms	48
Distribution List	50

List of Figures

Fig. 1	Relative distributions of speech power and speech intelligibility across frequency scale (male voice; normal voice effort)	4
Fig. 2	Diagram of a speech communication system (Letowski)	4
Fig. 3	The effect of telephone bandwidth on transmission of vowel /o/ by male talker with voice fundamental frequency $F_0 = 100$ Hz (Rodman 2003)	5
Fig. 4	Relationships between various speech intelligibility scales expressed on the common intelligibility scale (Knight 1994)	25
Fig. 5	CIS and STI values defining STI categories proposed by Houtgast and Steeneken (1984; 1985)	28
Fig. 6	CIS and STI potential relation with category rating (CCR) scale used as the MOS scale in telephone networks quality evaluation.....	28
Fig. 7	Performance intensity functions obtained for the callsign acquisition test (CAT) and the MRT in white noise (Blue-Terry and Letowski 2011)	33
Fig. 8	Theoretical shapes of MRT and CAT performance intensity functions derived for data from the CAT and MRT speech intelligibility measures (Blue-Terry et al. 2012)	33

List of Tables

Table 1	Speech levels (dB A-weighted) of male and female talkers at a 1-meter distance. Levels listed by Berger et al. (2003) and ISO-9921 (ISO 2003) standard (numbers in parentheses)	2
Table 2	Mean opinion score (MOS) and degradation category rating (DCR) scales	10
Table 3	Comparison category rating (CCR) scale used in telephone network quality evaluations	11
Table 4	Ranges of speech intelligibility and corresponding values on the STI and CIS scales	27
Table 5	Required speech intelligibility scores as per ISO 2470:2007 (ISO 2007b)	29
Table 6	Intelligibility criteria for voice communication systems	30
Table 7	Intelligibility score requirements as a function of communication criticality	30

1. Introduction

Speech communication can be assessed in a variety of ways. The objective of this review is to summarize the most common measures of speech intelligibility, stress the strengths and limitations of each of these measures, and establish a basis for comparison between the results of these measures. How speech is measured depends in part on the objective of the measurement—whether it is to establish good speech quality or speech intelligibility—and in part on the restraints and costs associated with the measurement. This review summarizes the different types of measurements of speech quality and speech intelligibility and discusses their relationship. The standard speech intelligibility rating scales, perceptual speech intelligibility tests (based on human performance), and technical speech intelligibility predictors (based on the input signal transmitted through a communication system or medium) are described and compared. The merits and limitations associated with each of these measures are discussed with an aim to outline the criteria on which test selection should be founded. The optimal measure will depend on the absolute need for speech intelligibility, the acoustic environment in which communication will occur, as well as the cost of measurement and regulatory requirements associated with the evaluation. Consequently, a common speech intelligibility scale is described that allows data scores from different measures of speech quality and speech intelligibility to be compared, and allows the user to predict performance on one measure based on the results of another. This analysis is intended to serve as a resource for users of standard speech intelligibility measurement methods and as a guide in selecting speech intelligibility tests and performance criterion for specific applications.

2. Characteristics of the Speech Signal

Speech is a form of language (communication code) that uses vocally produced sounds to convey thoughts, meanings, and feelings. To communicate by speech, speech sounds must be both produced and perceived. Speech production refers to the process by which predetermined vocalized sounds are produced by the talker and organized in sequences forming communication signals. Speech perception is the process by which the listener is able to hear and interpret (understand) the message encoded in the speech signals.

The effective design and use of audio communication systems requires some knowledge of the physical properties of speech and the rules that govern the human perception of speech. The 2 main physical descriptors of speech signal are its sound intensity and spectral content. The long-term average sound intensity levels of

phonated speech produced with various levels of vocal effort are listed in Table 1. However, individual phonemic components of speech vary greatly in their intensity with vowels carrying much greater energy than consonants. The strongest vowel, /aw/, as in word *all*, is about 28 dB more intense than the weakest consonant, /th/, as in word *thin* (Staab 1988). Whispered (unphonated) speech levels are in the order of 40 dB(A) (Traunmüller and Eriksson 2000) but this kind of speech is not used in formal communication.

Table 1 Speech levels (dB A-weighted) of male and female talkers at a 1-meter distance. Levels listed by Berger et al. (2003) and ISO-9921 (ISO 2003) standard (numbers in parentheses)

Vocal Effort	Male Talker	Female Talker
<i>Low (relaxed)</i>	52 (54)	50
<i>Normal</i>	58 (60)	55
<i>Raised</i>	65 (66)	62
<i>Loud</i>	76 (72)	71
<i>Shout</i>	89 (78)	82

A person's vocal level effort depends on the visual and auditory clues stemming from the distance (real or perceived) to the listener and the emotional state of the talker. In noisy environments vocal effort is naturally higher (raised, loud, or shouted) than in quiet (normal) environments, because talkers involuntarily raise their voices to the level needed for them to hear themselves (Lombard effect*; Fairbanks 1954; Summers et al. 1988). Conversely, talkers wearing hearing protectors reduce their vocal efforts by about 3 dB, compared to when unprotected, if the background noise level exceeds 75 dB A (ISO 2003).

The speech levels listed in Table 1 are the levels measured in front of the talker's mouth. However, the vocal source is quite directional and the levels at the talker's back may measure up to 5–7 dB lower. This difference is relatively small at low- and mid-frequencies but sharply increases for spectral content at higher frequencies (consonants).

3. The Concept of Speech Intelligibility

The main criterion for a speech system's effectiveness is its speech intelligibility, that is, the degree to which speech can be understood by a listener. According to Côté (2011, p. 11), speech intelligibility is a measure of how much of a message has been extracted from the recognized phonemes (the smallest units of speech). In other words, it indicates the extent to which words and sentences can be understood

* The Lombard effect, or Lombard reflex, is the involuntary tendency of speakers to increase their vocal effort when speaking in loud noise to enhance the audibility of their voice.

(Viswanathan and Viswanathan 2005). Speech intelligibility is usually expressed as the fraction (percentage) of speech units received correctly. It primarily depends on speech energy in the 1000–6000 Hz band and is poorly correlated with the overall frequency distribution of speech power that extends from about 80 to 8000 Hz (MIL-HDBK-1908B 1999) with the main concentration of average speech spectrum energy in the range of 200–600 Hz. The relationship between the speech power spectrum and the contributions of various spectral regions to speech intelligibility is shown in Fig. 1. Speech power above 1000 Hz is equal to only about 5% of the total speech power but it contributes about 60% to speech intelligibility (Fletcher 1953). Speech energy below 250 Hz contributes only about 2% to overall speech intelligibility (Gerber 1974, p. 244). The spectral centroid (center of gravity) of the speech spectrum below and above of which speech power equally contributes to speech intelligibility in the English language is close to 1500 Hz (Fletcher 1953) [1700 Hz (French and Steinberg 1947)] and varies depending on language.

Similarly to overall speech power, vowels and consonants also differ widely in spectral content. Consonants carry most of the information contained in English speech as well as in other languages. Consonants contain mostly high frequency (above 1500 Hz) speech energy, but this energy is relatively small in comparison to that of the whole speech spectrum (Fig. 1) and thus, consonants are easy to misjudge. Vowels, having much greater power and mostly low- and mid-frequency spectral energy, are usually easily heard. Note that changes in voice effort level not only affect the resulting overall speech level but also speech spectrum. Greater vocal effort results in elevated mid-frequency spectral regions in female and male voices and a higher overall fundamental frequency in male voices (Letowski et al. 1993). Since the mid-frequency spectral region contains mostly vowel energy while consonants are high frequency sounds, an increased vocal effort typically reduces the intelligibility of speech. Note also that in whispered speech the overall sound power is low and the speech spectrum is almost flat, resulting in more frequent vowel confusion than occurs for normal speech.

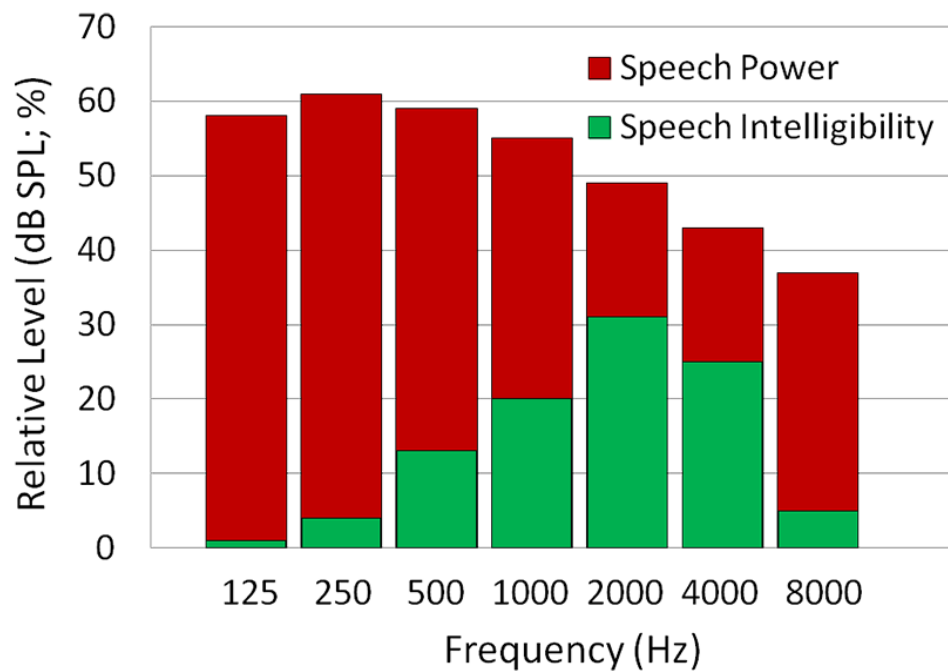


Fig. 1 Relative distributions of speech power and speech intelligibility across frequency scale (male voice; normal voice effort)

In general, the intelligibility of speech signals received by a listener depends on speech articulation by the talker, properties of the surrounding environment or communication system, and the listener's ability to understand incoming speech signals. Figure 2 shows a basic diagram of the relationship between terms related to the intelligibility of speech transmitted by a communication system (communication channel).

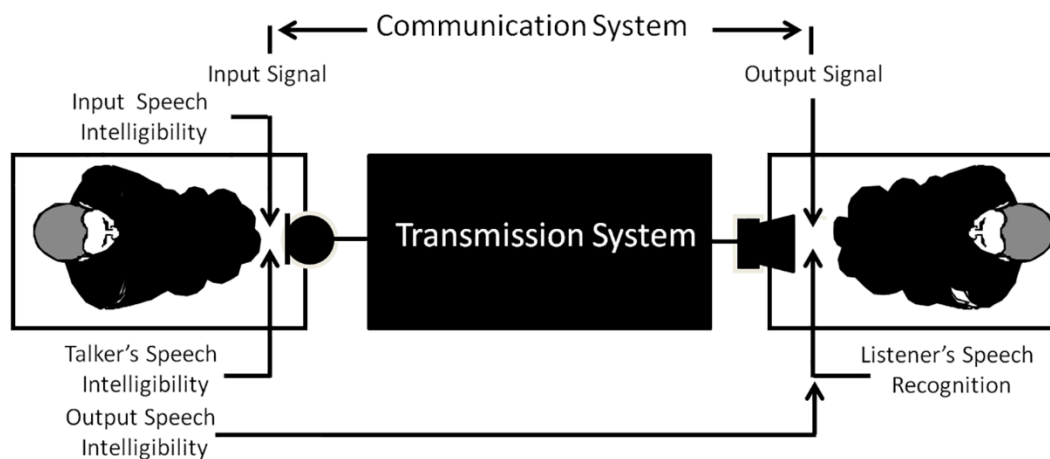


Fig. 2 Diagram of a speech communication system (Letowski)

The communication system shown in Fig. 2 can be any physical environment (air, electroacoustic, optoelectronic, etc.) connecting the talker and the listener whether they are people or machines. When the concept of speech intelligibility is applied to speech degradation caused by a communication system, it should be assumed that both the talker and listener are the ideal sender and receiver, respectively. If this sender and receiver were connected by the ideal speech communication system, all of the speech signals voiced by the talker would be correctly reported by the listener. These speech signals can be said to have 100% speech intelligibility. Therefore, any departure from 100% speech intelligibility in speech transmission through a communication system connecting ideal source (talker) and receiver (listener) is assumed to be due to the limitations of this system. In the case of real talker and real listener, they both contribute to the degradation of speech in the communication channel, and both the limited talker's speech intelligibility and the limited listener's speech recognition ability need to be considered and measured.

A long list of the technical factors that affect speech intelligibility during speech transmission between ideal source and receiver can be found in the ANSI/ASA S3.2 standard (ANSI/ASA 2014). The most dominant of them are limited transmission bandwidth, type and level of background noise (including crosstalk), and oscillations and echoes. The effect of telephone bandwidth (300–3400 Hz) (ITU 2003; TIA 2006) on speech spectrum transmission is shown in Fig. 3.

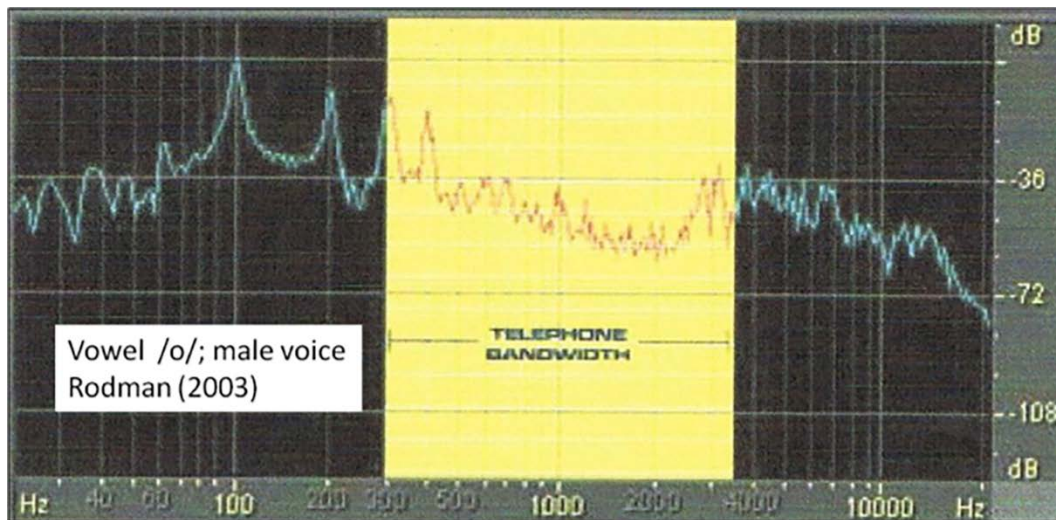


Fig. 3 The effect of telephone bandwidth on transmission of vowel /o/ by male talker with voice fundamental frequency $F_0 = 100$ Hz (Rodman 2003)

The upper limit of frequency bandwidth is considered the most critical technical parameter of a transmission channel with respect to speech intelligibility. Vowels, which are strong wideband sounds, do not contribute much to speech intelligibility at the frequency range below 200–300 Hz. However, excluding the frequency range

below 300 Hz affects the quality and realism of transmitted speech. Consonants, which are low energy speech sounds, are predominantly made up of high frequency spectral content. Therefore, good speech intelligibility requires transmission bandwidth extending up to 7000 Hz. For single syllable words, limiting the upper range of the bandwidth to 3400 Hz results in speech intelligibility as low as 75%. A bandwidth of 7000 Hz is needed to achieve speech intelligibility better than 95% (French and Steinberg 1947). Video conferencing audio connections commonly transmit frequencies up to 6800 Hz. Voice over internet protocol (VoIP) telephony, as well as cellular telephone networks, are also moving to a wider (150–6800 Hz) bandwidth using compressed and uncompressed codec techniques (TIA 2015ab; TIA 2011).

The presence of noise makes the perception of speech difficult due to the strong masking of the high-frequency components of speech. As a result, the consonants, which are low-energy, high-frequency sounds, are more prone to masking than the high-energy, wide-spectrum vowels. Such contaminated speech is still audible but not clear. Thus, speech may be detected but not recognized at low speech-to-noise ratios (SNRs). In general, an SNR of at least 6–10 dB is needed across the frequency range of speech (200–7000 Hz) to make speech sufficiently intelligible for communication (~90% word recognition score) (Moore 1977; NASA/SP-2010-3407 2010).

Similarly, the presence of excessive reverberation (reflected sounds from the boundaries of a space), as well as echoes in the communication system, extend the effective duration of transmitted sounds and cause both concurrent and temporal masking of subsequent sounds. Conversely, nonlinear distortions, such as compression and level limiting, have relatively small effect on speech intelligibility and in some cases may improve speech intelligibility (Moore 1977).

4. Speech Intelligibility and Speech Quality

In making assessments of speech communication systems, speech intelligibility needs to be differentiated from speech quality, which is another perceived characteristic of speech. The judgment of speech quality is an expression of general satisfaction. It expresses the listener's preference for, or discomfort with, the perceived speech. According to Jekosch (2005), the evaluation of speech quality is a comparison of the received speech to the desired speech. Unfortunately, the literature disagrees about the factors that form the main perceptual dimensions of speech quality and their relationship to speech intelligibility. Côté (2011) has analyzed several publications related to this topic and concluded that speech quality has 4 main components: loudness, speech continuity, noisiness, and coloration.

Speech intelligibility was included as another minor component. However, the most prevailing view is that speech quality has 2 main components: naturalness (voice quality) and speech intelligibility (Kraft and Portele 1995; Lewis 2001). Naturalness is the degree to which the speech signal resembles normal clear human speech (Viswanathan and Viswanathan 2005). Voice quality has been defined as the characteristic color of speech (Keller 2005) and can be equated with voice pleasantness or voice naturalness but has to be differentiated from voice timbre, which specifies voice character without qualifying it. Both quality and timbre have multidimensional character and can be evaluated along several perceptual dimensions such as brightness, loudness, nearness, and so on. Most of them are the same in both domains but expressed on qualitative *better-worse* (quality) and quantitative *more-less* (timbre) scales (Letowski 1992).

Since speech intelligibility is frequently considered to be a component of speech quality, some methodologies that are used for assessing speech quality can also be used for assessing speech intelligibility. However, speech quality judgments should not be used *in lieu* of speech intelligibility judgments or vice versa because such judgments may be misleading. Although in some cases speech quality may be highly correlated with speech intelligibility (e.g., Leijon et al. 1991; Sullivan et al. 1988), in other cases it may not be (e.g., Goodman et al. 1978; Punch and Parker 1981; Byrne and Cotton 1988). Payton and Braida (2005) studied the effects of slow-acting wide dynamic range compression (WDRC) on speech intelligibility and speech quality judgments and observed that while WDRC can improve speech intelligibility, it decreases speech quality ratings as the compression ratio increases. The authors concluded that judgments of speech quality should accompany assessments of speech intelligibility to assure that speech transmission optimizes both comfort and intelligibility.

Preminger and Van Tasell (1995) observed that when large differences in speech intelligibility exist they dominate all other perceptual judgments and in such cases speech quality and speech intelligibility are highly correlated. However, this may not be the case when differences in speech intelligibility are relatively small. For example, Kondo (2011) compared results of a speech intelligibility test (DRT) and mean opinion score (MOS) judgments of speech quality and reported the Pearson correlation coefficients of 0.36–0.47 for female and male speech mixed with white and babble noise. In addition, the perceived contributions of voice quality and speech intelligibility to overall speech quality may differ greatly among judges, leading to large variability in the observed data. Unfortunately, in many publications and test reports, the term speech quality is used *sui generis* when the authors mean speech intelligibility. The comment above applies not only to speech quality but also to other similar criteria, such as clarity and comprehensibility,

sometimes used in lieu of speech intelligibility as synonyms. Therefore, in judging speech intelligibility, only the *speech intelligibility* term should be used as an aim of the rating and other similar terms should be avoided. Eisenberg et al. (1998) compared perceptual judgments of speech clarity and speech intelligibility of the same speech material and observed that the ratings were highly related but differed substantially in magnitude with intelligibility ratings consistently exceeding clarity ratings. Recently, Reinhart and Souza (2016) found a high correlation between perceptual judgments of speech clarity and speech recall scores (termed speech intelligibility) but noticed that the recall scores were dependent on the capacity of the listener's working memory while the rating scores were not. In cases like this it is impossible to determine whether the judges equated clarity with perceived intelligibility. Some authors also believe that speech intelligibility differs from speech comprehensibility, which includes contextual information (Yorkston et al. 1996; Hustad 2008; Fontan et al. 2015).

In considering terminological issues, the reader needs to be aware that some authors regard speech intelligibility as having 2 components—audibility and clarity—that contribute to overall intelligibility while other authors consider speech intelligibility as having an even larger number of dimensions. For example, De Bodt et al. (2002) argued for 4 main dimensions: voice quality, articulation, nasality, and prosody. All of these factors indicate that rating studies need to be done very carefully with very clear instructions and with sufficiently trained judges.

5. Speech Intelligibility Assessment

In reference to Fig. 2, speech intelligibility tests may be used for 3 distinct purposes:

1. To assess the talker's ability to articulate speech and to diagnose potential voice/speech disorders;
2. To assess the listener's ability to recognize speech sounds and to assess the listener's hearing deficiencies affecting speech communication (speech audiometry); and
3. To assess the capabilities of a specific transmission channel (e.g., electroacoustic system, reverberant room, etc.) for speech communication and the degree of potential speech degradation.

The focus of this review and analysis is on the third purpose—assessment of transmission channels; however, some limited inferences regarding 2 other purposes—speech articulation and speech recognition—will also be made.

In general, speech intelligibility evaluation methods can be divided into perceptual methods, which require natural or synthetic speech and human listeners, and technical (objective) methods, which use artificial signals and the technical parameters of the communication system to predict speech intelligibility according to some internal criteria. Perceptual methods can be further divided into rating (ordinal scale) and speech recognition measurement (ratio scale) methods based on an estimated or calculated percent correct of listener's responses.

In the case of perceptual tests, speech material should be phonetically (or functionally) balanced to allow equal stress on all elements of transmitted speech. While phonetic balance is especially important for talker and listener assessments, functional balance is important for testing transmission channels used for specific applications.

Transmission channels should be tested under normal operational conditions. For comparison, in speech audiometry, the talker's and listener's ability to produce and receive speech, respectively, should be evaluated under optimal conditions where the SNR ratio is at least 40 dB.

5.1 Speech Intelligibility Rating

Rating is a perceptual categorization procedure by which an object or various objects are assessed according to their perceived grade or rank using a verbal or numeric scale of successive categories. The number of successive categories can be freely specified but it should be small enough to be manageable and large enough to allow meaningful differentiation of the objects. The speech material can be operational phrases, sentences or connected speech that are representative of the specific application. They may be prerecorded, synthetic, or delivered live but must be presented in a way consistent with the purpose of the assessment.

The most widely used statistical rating system for speech intelligibility is the MOS. In a MOS test, a number of people rate their perceptions on a 5-step scale from 1 (bad) to 5 (excellent) and their ratings are statistically averaged. The MOS is an absolute category rating system in which a listener makes a rating after listening to a single item and no comparisons are made. Only integer numbers should be used for ratings (ITU-T 1996). Other rating scales exist that use 7-, 9-, and 11-steps, but these are only recommended for highly trained judges (Miller 1956; Osgood et al. 1957; Eisler 1966; Warr and Knapper 1968).

The MOS was originally developed for the assessment of speech quality in telephone networks. Since then its use has been extended to various audio and video signals and systems where a degree of departure from an ideal reference signal needs to be known. When the MOS is used to assess the quality of transmission systems, it is sometimes referred to as the “transmission quality” or “quality of service (QoS) rating system. Obviously, the same rating system may be used for assessment of other signal or system properties, such as speech intelligibility, if only the assessment criterion is clearly stated.

Since the MOS can be used for an assessment of various systems and signals, it has a number of qualifiers appended to its name to indicate its specific application. When the MOS is used to assess listening quality of a signal it is referred to as the MOS-LQS (or MOS-LQ), that is, the MOS-Listening-Quality Scale (ITU-T 2003). The MOS-LQS scale is commonly used in the telephone industry to rate performance of telephone network connections and is described in ITU-T Recommendation P.800 (ITU 1996). It can be used alone or together with 2 supporting scales: the listening-effort scale (MOS-LES) and the loudness-preference scale (MOS-LPS)—both being similar 5-step scales—to assess various aspects of the telephone transmission. Other versions of the MOS include the MOS-CQS (MOS conversational-quality scale) or the MOS-PQS (MOS picture-quality scale). In case of speech intelligibility ratings, the MOS scale may be qualified as the MOS-SIS (MOS speech intelligibility scale).

The MOS scale in its original form, and in the paired comparison form known as the degradation category rating (DCR) scale, are both shown in Table 2.

Table 2 Mean opinion score (MOS) and degradation category rating (DCR) scales

Rating	MOS	DCR
5	Excellent	Inaudible
4	Good	Audible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

In the paired comparison judgments (DCR scale), the reference condition is always presented first and the reference condition must always be better than the tested condition. This limits the use of the DCR.

Goldberg and Riek (2000, p. 118) observed that most speech coding systems score between 3 and 4 on the MOS scale and that a coder scoring above 4 provides very good speech quality. More recently, Waidyanatha et al. (2012) suggested that MOS values equal to or greater than 4 should be the desired speech quality score for

telephone network connection and for emergency communication systems. Typical modern high quality codec systems such as the G.711 score about 4.4 on the MOS scale while the G.729, which involves greater signal compression, scores 4.1

It has to be stressed that despite its name, the MOS is a rating (assessment) system and not *sensu stricto* a scoring (measurement) system. The MOS scale, however, meets all of the criteria for a well-designed rating scale, having an odd number of categories and clearly defined levels. The 5-item category range is large enough to be sensitive to differences and small enough to be easily memorized and result in fairly repeatable judgments.

Absolute category ratings used in MOS testing are generally less sensitive than estimates obtained by a paired comparison technique (e.g., Munson and Karlin 1962; Eisenberg et al. 1997). When appropriate, the International Telecommunication Union (ITU) recommends that MOS ratings be combined with paired comparison techniques to create an extended 2-directional version of MOS scaling. The resulting scale is a 7-step extended MOS scale called the Comparison Category Rating (CCR) scale and is shown in Table 3. Such 2-directional comparative scales are considered less taxing for the human information processing system than single-directional absolute rating scales (e.g., Osgood et al. 1957). The CCR scale is used in telephone network quality evaluations. The listener compares a test sample to a standard reference sample that can be of a higher or lower quality than the test sample (e.g., Rothauser and Urbanek 1965). The resulting speech quality value is referred to as the Comparison Mean Opinion Score.

Table 3 Comparison category rating (CCR) scale used in telephone network quality evaluations

CCR Rating	Quality of the second stimulus compared to the first one
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

The main advantages of the MOS rating system for assessment of speech communication systems are: 1) comprehensive assessment of the system, 2) face validity of the data, and 3) simplicity of testing. However, MOS tests require trained listeners to provide meaningful scores and to avoid large data variability. Further, MOS ratings do not indicate what types of problems were encountered by

the listeners unless additional attributes are also rated. Thus, the test results are open to interpretation by equipment manufacturers and installers since they provide no contextual information about the causes of poor speech quality.

When used for speech intelligibility assessment, the MOS scale, as well as all other rating scales, is not sufficiently precise for some applications. That is, the degree to which it can predict speech intelligibility scores is quite limited in comparison to other scoring methods. Further, it is a low-level, ordinal scale that limits the types of statistical analyses that can be performed on the data. Conversion of ordinal-scale numbers into ratio-scale numbers is possible, but requires a large sample size to assure small confidence intervals, making data collection both costly and time-consuming.

Despite the limitations described above, categorical ratings of speech intelligibility, such as those made with the MOS scale, have been reported to be “valid measures of speech recognition” (Purdy and Pavlovic 1992, p. 254). Several authors have reported good agreement between speech intelligibility ratings and intelligibility measured as percent correct speech recognition (e.g., Payton and Braida 2005). For example, in studies conducted at the Harvard Psycho-Acoustic Laboratory during World War II, intelligibility ratings obtained from a small number of trained listeners correlated well with speech recognition scores obtained with the same listeners (Stevens et al. 1944). Therefore, numerical scores produced by the various speech intelligibility tests are frequently equated to MOS categories when evaluating their results.

Using a percentage rating scale (0–100%) with assumed ratio scale properties allows one to avoid the limitations of ordinal scale. A percentage rating scale requires that the judged range has 2 physical ends and the judged quantity can be expressed as a fractional number. In the case of speech intelligibility, this scale is referred to as the Speech Intelligibility Rating (SIR) scale. Judgments made on this scale express the listener’s belief that a certain percentage of the message was properly heard. Percentage scale scores for both syllables (Rankovic and Levy 1997) and sentences (Speaks et al. 1972; Cienkowski and Speaks 2000) showed very good agreement with speech recognition scores even when using untrained listeners. Because SIR tests are much less time consuming than speech recognition tests (described later), they can be reliably applied in many situations where data collection needs to be completed in a short time and with ad-hoc listeners.

A well-documented, freely available, sentence intelligibility test involving SIR scale has been published by Cox and McDaniel (1989). The test consists of several, specially-written connected discourse passages that are read (presented) to a listener who subsequently judges, on a scale from 0 to 100%, how intelligible the

passage was (see also McDaniel and Cox 1992). The test was initially developed for hearing aid selection but has since been used for judging distortions in speech transmission systems and the speech production and perception of cochlear implant users (e.g., Allen et al. 2001; Zhou et al. 2013).

There are relatively few studies comparing speech intelligibility rating data with intelligibility test scores but all of them indicate that intelligibility ratings closely approximate objective scores on a sentence transcription task when performed by normally hearing listeners (e.g., Yorkston and Beukelman 1978; Cox et al. 1991). However, Cox et al. (1991) reported that hearing impaired persons have a tendency to rate speech intelligibility as lower than indicated by their objective scores.

5.2 Perceptual Tests of Speech Intelligibility

Perceptual tests of speech intelligibility directly measure the percentage of correctly understood speech items and several such test methods are included in international and national standards. In these tests speech recognition responses are collected from human listeners presented with some forms of human speech: *phonemes*, *syllables*, *words*, *phrases*, *sentences*, or *paragraphs*. Speech material can be live, recorded, or synthetic. In addition to assessment of overall speech intelligibility, collected data can be analyzed to provide information about type of errors and about specific deficiencies that were encountered. However, their validity is limited to the speech material and speech complexity level (speech unit).

The speech unit most commonly used to assess speech intelligibility of the communication systems is the word, although phonemes, syllables, and sentences have also been used. For example, IEEE Recommended Practice for Speech Quality Measurements (IEEE 1969; Rothaus et al. 1969), lists 72 lists of 10 phonetically balanced sentences (Harvard Sentences) for communication systems testing. The internet Open Speech Repository makes available standardized and repeatable recordings of female and male voices reading the IEEE sentences [http://www.voiptroubleshooter.com/open_speech/american.html]. However, as the most common speech intelligibility test used to measure speech transmission over communication systems is some form of word test, this will be the focus of this section.

The preferred status of word tests in evaluation of speech communication systems may be explained by historically easy access to word level tests, attractiveness of the meaningful short test items to the listening panel, and easy mathematical treatment (e.g., by binomial model) of the data that are scored as either correct or incorrect and not by parts; for example, sentences scored as percentages of words correctly repeated (Thornton and Raffin 1978; Raffin and Schaffer 1980;

Boothroyd and Nitttrouer 1988). It is also important to mention in this context the assumption that speech recognition (speech intelligibility) is a single construct (Bilger 1984); “Because all speech-recognition tests evaluate the same construct, scores on all tests must be related and, therefore, scores on one speech-recognition test should be predictive of scores on other tests,” (Olsen et al. 1997, p. 183). This assumption has been validated in some studies within 6–12% error for phoneme, word, and sentence tests using the same basic speech material presented both at low and high SNRs and regardless of the hearing status of the listener (Boothroyd and Nitttrouer 1988; Olsen et al. 1997). However, it has to be remembered that the differences in speech material and in number of test items used in various speech tests call for caution in making any generalization regarding test equivalency.

The predictability and reliability of word-level tests depend on the test vocabulary size and the number of response alternatives offered to the listener. Larger vocabulary sizes make the test more predictive and reliable. In terms of the listener’s response alternatives, all tests can be categorized as either closed-set or open-set response alternatives. An open-set test assumes an unlimited set of available alternatives and the listener must make an unconstrained guess of the presented word. The intelligibility (INT) score is calculated as

$$INT(\%) = \frac{100}{T} (R), \quad (1)$$

where T is number of items in the test and R is number of correct responses. In a closed-set test, the listener is given a set of response alternatives (words) and must select one of them in response to the presented item. The INT for a closed-set test is usually calculated as

$$INT(\%) = \frac{100}{T} \left(R - \frac{W}{N-1} \right), \quad (2)$$

where W is the number of wrong responses and N is the number of listener response alternatives. The second part of Eq. 2 is a correction for guessing.

Assuming a binomial distribution of the test scores (Thornton and Raffin 1978), its standard deviation (SD) can be calculated as

$$SD = 100 \sqrt{\left[\frac{e(1-e)}{n} \right]}, \quad (3)$$

where n is the number of words and e is the word error rate (percent error).

Three word-level speech intelligibility tests are included in the American National Standards Institute (ANSI/ASA 2014) standard for use in the evaluation of speech transmission systems:

- 1) Phonetically Balanced Word Test (PBWT)
- 2) Diagnostic Rhyme Test (DRT)
- 3) Modified Rhyme Test (MRT)

Other similar tests include the Central Institute for the Deaf (CID) W-22 Auditory Test (Hirsh et al. 1952), the Northwestern University Auditory Test No. 6 (Tillman and Carhart 1966), the California Consonant Test (Owens and Schubert 1977), and the US Army Research Laboratory (ARL) Callsign Acquisition Test (CAT) (Rao and Letowski 2006).

5.2.1 Phonetically Balanced Word Test (PBWT)

The PBWT is a semi-open word test consisting of 1000 monosyllabic words divided into 20 phonetically balanced lists of 50 words. Each word is embedded in the same carrier phrase such as “Would you write ____ now.” The words of each list are presented in a different random order each time the list is used. In a single test trial, a word from the list is presented to the listeners who write down what they think the word was. The intelligibility score (PB) is calculated as it would be in the case of an open-set test. The test, originally developed by Egan (1948) at the Harvard Psycho-Acoustic Laboratory, was originally known as the Harvard PB-50 test. The minimum test length is a one-word list but using 2–3 lists is recommended since this results in more reliable data.

The phonetic balance of speech sounds is very important for all speech tests since increasing the frequency of a sound's occurrence increases the probability of its recognition and selection of a related word. Similarly, words that appear more frequently than others in the test are easier to recognize. Rubenstein and Pollack (1963) refer to it as the “word predictability” and stated that within certain limits, word intelligibility “is a simple power function of [word] probability,” (p.157). They also noted that successive repetition of the same word improves its intelligibility. Miller et al. (1951) observed that repeating the word 3 times had the same effect as improving SNR by 2 dB. Hilkhuisen et al. (2012) presented short sentences up to 8 times and observed an improvement in recognition during the first 5 presentations (from 36 to 51%).

The PBWT is difficult to accurately administer and requires a long training time (typically 10–20 h) before listener responses stabilize at their highest level. The more distortions within a communication channel, the greater their effects on the recognition of unfamiliar words relative to familiar words (Epstein et al. 1968). Therefore, a long training time is required to familiarize listeners well with all the words (1000) and their pronunciations. However, the PBWT is the most accurate

of all the standardized tests and is recommended for use when high data accuracy and sensitivity are required. It is particularly sensitive to the SNR; a relatively small change in SNR causes a large change in the intelligibility score. In ANSI/ASA S3.2 the PBWT uses 1000 words but there are also shorter versions, including one with 256 words for testing less critical systems.

5.2.2 Diagnostic Rhyme Test (DRT)

The DRT uses monosyllabic English words and consists of 96 rhyming word pairs (Voiers 1977, 1983). Almost all of the words have the consonant-vowel-consonant (CVC) structure. The words in a pair differ only in their initial consonant sounds (i.e., veal-feel) following the concept of the rhyme test developed by Fairbanks (1958). In a single test trial, the listener is visually shown a pair of words and asked to identify which one of these 2 words is subsequently spoken by the talker. The correct word is always one of the 2 words presented visually. Since it is a closed-set test, the closed-set correction for guessing given in Eq. 2 is used. A carrier sentence is not used.

The main strengths of the DRT are its simple administration, short training time (1–2 h), and task simplicity. Because the words are presented without a carrier phrase, the test administration is very time efficient. It is also a phonetically balanced test that takes into account both word similarity and phonetic frequency. In addition, the test results can be reviewed in terms of distinctive features of speech derived from the Jacobson et al. (1952) distinctive feature system. To calculate the speech intelligibility score, the listener's responses are categorized according to 6 phonemic features: voicing, nasality, sustention, sibilation, graveness, and compactness. These category scores provide diagnostic information about transmission system deficiencies. At the final stage of score calculation, the 6 scores are averaged together to provide an overall measure of system intelligibility.

Unfortunately, the DRT measures speech intelligibility for only initial consonants. It has been established that initial consonants of normal speech are more easily recognizable than final consonants (Nye and Gaitenby 1973). Further, the DRT has been criticized as having relatively poor sensitivity: a 2-alternative closed-set test being “too easy to produce meaningful results,” (Webster and Allen 1972, p. 36). However, due to its providing information about the distribution of distinctive feature errors, the DRT has been extensively used in testing speech coders for both military (e.g., by Department of Defense Digital Voice Processor Consortium) and civilian applications (e.g., Schmidt-Nielsen 1992).

5.2.3 Modified Rhyme Test (MRT)

The MRT is a monosyllabic word test developed by House et al. (1965) that uses 300 monosyllabic English words divided into 50 6-word lists of rhyming or similar-sounding words (i.e., pin, sin, tin, fin, din, win). Almost all words of the test have a CVC structure. The words in each group differ in either the sound of the initial (25 groups), or final (25 groups), consonant. The spoken words are typically presented in a carrier phrase “You will mark ____ now.” A listener is shown 6 rhyming response alternatives and asked to identify the spoken word. The MRT has structural similarity to the rhyme test developed by Fairbanks (1958) who used 5-word groups of rhyming words; however, words in the rhyme test and the DRT only differed in their initial consonant sound.

The MRT has gained wide popularity and is typically used to measure the communication performance of military and aviation communication systems. As a closed-set (N=6) test, the MRT is easy to administer and it does not require as much training as for the PBWT. Further, listeners show very little evidence of learning during repeated tests (House et al. 1965; Williams and Hecker. 1967; Nye and Gaitenby 1973). MRT scores tend to be lower for final than for initial consonants, which is the reason why MRT is frequently preferred over the DRT for speech communication system assessment. Logan et al. (1989, p.579) used the MRT to compare speech intelligibility of several speech synthesizers and reported that the “MRT can be used as reliable measure of the segmental intelligibility of synthetic speech produced by a variety of text-to-speech systems.” Normally, the MRT requires only a short training time of 1–2 h but in the case of synthetic speech, the learning effect during tests is easily noticeable (Nye and Gaitenby 1973).

Limitations of the MRT include “an imbalance in the number of times various consonants are presented, an uneven representation of initial and final consonants (with 2 consonantal phonemes omitted entirely), and an incomplete use of vowel environments” (Nye and Gaitenby 1973, p.90). In sum, in the phonemes system, /ə/, /a/, /u/, /aI/, /av/, and /ov/ are not used at all, the phonemes /tʃ/ and /z/ are not used in the initial and the phoneme /ʃ/ is not used in the final position (Nye and Gaitenby, 1973, p.78). In some later versions of MRT word list, the types of missing phonemes change but the test still remains unbalanced.

In addition, it has to be stressed that the PBWT is an open-set test while both the MRT and the DRT are closed-set tests. Closed-set tests are preferable to open-set tests because they are easier and less expensive to administer and require less training, that is, they are less affected by participants’ learning and practice (Schmidt-Nielsen 1992). They also result in smaller inter-listener variability of the

test results and thus require fewer listeners to achieve the same level of data reliability. However, both MRT and DRT tasks focus on one specific phoneme at a time while the PBWT focuses attention on the whole word and thus its open-set structure is “more likely to match the processing characteristics of the natural speech environment” (Greenspan et al. 1998, p. 210). Another advantage of PBWT over the MRT and DRT is that its scores are usually lower than the scores of the other 2 tests so less possibility exists for the ceiling effect to affect collected data in case of testing good communication systems (Schmidt-Nielsen 1992).

One common drawback of all perceptual speech intelligibility tests is “the correlation between mean values and the variance of scores contributing to these means” (Studebaker et al. 1995, p. 174). This is also called “the ceiling effect” and affects scores on both very low and very high ends of the scale; however, a nonlinear transformation of the scores, such as arcsine transformation, can alleviate the problem (Studebaker 1985; Studebaker et al. 1995).

5.3 Technical (Predictive) Tests of Speech Intelligibility

The main advantage of all perceptual ratings and intelligibility tests is that because these methods are based on the perception of speech by listeners, they do not have any limitations in respect to the characteristics of the communication system or those of the test environment. However, their common limitation is that they can be time-consuming and costly, especially when the tests are conducted within an iterative design process. In addition, to verify that the obtained results are reliable, the test should be repeated at least once and some measure of repeatability derived from the data.

Technical measures of speech intelligibility use synthetic speech-like signals, or the relevant average speech and noise spectra, and the technical parameters of the communication system to calculate speech intelligibility. They were developed as fast and inexpensive substitutes to perceptual measures for assessment of speech intelligibility and include real-time and on-line applications. They do not actually measure speech intelligibility but predict it on the basis of the transmittability of the input signal. Their values typically vary from 1.0 (perfect) to 0 (poor) and they are assumed to be univocally related to speech intelligibility scores from 100% to 0%. Such measures are very useful in the design phase of the communication systems to ensure that potential problems with speech intelligibility are identified and resolved early. They also are widely used in the case of existing communication systems since they are more cost-effective and time-effective than perceptual tests. However, they do have their own limitations and they are not substitutes for perceptual speech intelligibility testing of the operating systems unless they a) have

been proven to sufficiently account for all the types of distortions caused by the signal processing techniques used in the tested system or b) their relative scores are sufficiently higher than the corresponding human tests intelligibility scores required for the tested system. When speech intelligibility is predicted by technical measures, one needs to consider that speech compression and peak-clipping operations may be seen by the technical measures as decreasing SNR and therefore results in artificially low speech intelligibility predictions, although in fact they may enhance speech intelligibility.

The 3 most commonly used quantitative speech intelligibility predictors standardized in ANSI/ASA S3.5 and IEC 60268-16 are:

- 1) Articulation Index (AI),
- 2) Speech Transmission Index (STI), and
- 3) Speech Intelligibility Index (SII).

Several other measures, such as articulation loss of consonants (ALcons) (Peutz 1971), speech interference level (Beranek 1947), or speech audibility index (Mueller and Killion 1990), also exist and are included in some comparative studies. For example, Barnett (1997) compared speech intelligibility scores measured with ALcons with scores obtained with some form of the STI test. However, these other methods are either less general in their applications or are less widely used.

In addition to the above predictors of speech intelligibility, there is the group of measures (algorithms) recommended by the ITU for prediction of speech quality, and more specifically for prediction of the MOS number. The most advanced among them is the Perceptual Evaluation of Speech Quality (PESQ) algorithm. Some of the other algorithms are the Perceptual Analysis and Measurement System (PAMS) and Perceptual Speech Quality Measurement (PSQM). All of these measures calculate the overall amount of signal distortion introduced by the transmission system and transform it through a psychoacoustic model to estimate the speech quality of the output signal. PESQ scores are reported to be well correlated with speech intelligibility ratings under some conditions ($R=0.91-0.99$) (Beerends et al. 2004; 2005; 2009). However, since they are, by definition, estimates of speech quality and not speech intelligibility, they are not discussed here.

5.3.1 Articulation Index (AI)

The AI is based on the concept that the overall speech intelligibility of the communication system results from independent intelligibilities calculated from the

peak-speech-to-root-mean-square noise ratio measured in selected frequency bands from 200 to 7000 Hz (Fletcher and Steinberg 1929; Fletcher 1953; French and Steinberg 1947; Kryter 1962ab). Three methods of calculating AI are offered: a) critical band ($n = 20$) method (most sensitive), b) third-octave band method, and c) octave band method (easiest). The AI assigns a value between 0 (poor) and 1 (perfect) to speech intelligibility. An AI score of 0.3 or below is considered unsatisfactory, greater than 0.3 to 0.5 is satisfactory, greater than 0.5 to 0.7 is good, and greater than 0.7 is very good to excellent.

The AI accounts for system noise and linear distortions and its calculations are described in the ANSI S3.5 (ANSI 1969) speech intelligibility standard. It is a measure similar to a Shannon's channel capacity (Shannon 1948) defining the maximum amount of information that the system can transmit without error (Allen 2003). However, the AI does not sufficiently account for the effects of reverberation, compression, and nonlinear signal distortions (e.g., Humes et al. 1986; Knight 1994; Payton et al. 1994) and has since been superseded by SII in the newer version of the same standard (ANSI/ASA 2017).

5.3.2 Speech Transmission Index (STI)

The STI (developed in early 1970s at the Netherlands Organisation for Applied Scientific Research (TNO) [Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek] in the Netherlands) uses a concept similar to AI, in that the prediction of speech intelligibility is based on a weighted contribution for a number of frequency bands but it also employs the concept of modulation transfer functions (MTF) and uses a synthetic speech-like signal based on a number of amplitude modulation processes (Steeneken and Houtgast 1980). The analysis is performed using 7 octave bands, from 125 to 8000 Hz, and 14 modulation frequencies. The reduction of the depth of modulation, expressed as the MTF, measured at the output of the communication channel is considered to represent loss of intelligibility. Speech intelligibility measured on the STI scale is considered excellent at and above 0.75, good between 0.75 and 0.6, fair between 0.6 and 0.45, poor between 0.45 and 0.3, and bad below 0.3.

The STI was originally developed to predict the effects of background noise level and room reverberation on the intelligibility of male speech, heard directly or passing through a communications channel. In the case of noisy and reverberant speech signals it was reported to predict PBWT values with standard deviations of about 5.0% (Steeneken and Houtgast 1980; Jacob et al. 1991). However, the original STI was shown to severely underestimate speech intelligibility for systems with extremely limited bandwidths (e.g., horn loudspeakers for which the frequency response starts at 1000 Hz), does not account for nonlinear distortions

produced by the transmission system, and underpredicts the intelligibility of female voices. Therefore, the original STI was extensively revised (for a summary, see Steeneken and Houtgast 2002) and its revised version (STI_r) alleviates several of the original version's limitations. The revised STI accounts for some nonlinear and time domain distortions (e.g., peak clipping, automatic gain control, presence of echoes), implements redundancy correction for adjacent frequency bands (signal-to-noise ratios in adjacent bands are no longer considered independent), and redefines frequency weighting functions to account for both male and female voices.

The current STI calculation procedures, a direct calculation method based on MTF and an indirect calculation method (Schroeder 1981) based on impulse response (for comparison see Zhu et al. 2014a), were standardized internationally in the IEC 60268-16 standard (IEC 2011). However, despite its wide popularity and the steady improvement of its algorithms, the STI test is criticized as an unreliable predictor when nonlinear processes and compression are involved (Ludvigsen et al. 1993; Goldsworthy and Greenberg 2004). In the case of room acoustics, Onaga et al. (2001) reported that the STI does not discriminate properly between early reflection energy, which contributes to speech intelligibility and reverberant energy and echoes, which decrease intelligibility. The STI overestimates speech intelligibility in environments that are both reverberant and noisy (Payton et al. 1994). There have also been concerns voiced by some practitioners that STI does not account for the effects of perceptual masking and fluctuating background noise. For example, van Schoonhoven et al. (2017) reported that when the indirect STI calculation method is used, a minimum impulse-to-noise ratio of 25 dB in fluctuating noise is needed to get meaningful results. Mechergui et al. (2017, p. 1471) regarded both the direct and indirect methods as “not adequate for real time intelligibility monitoring”. Please note, however, that the STI algorithm has been gradually modified during the last 25 years; thus, some of these early criticisms may no longer be true.

Since the STI measure is computationally intensive, several simplified versions of STI have been developed to be used for less stringent transmission requirements. The Rapid Speech Transmission Index (RASTI), implemented in equipment produced by Brüel & Kjær (Denmark) and popular in some European countries, limits its calculations to the 2 octave bands centered at 500 and 2000 Hz and a limited set of 9 modulation frequencies (4 in the 500 Hz band and 5 in the 2000 Hz band). However, the RASTI frequently overestimated intelligibility and has since been abandoned. Another simplified version, the Sound Transmission Index Public Address (STIPA), uses all 7 STI octave bands but only a subset of 7 modulation frequencies (one frequency per band). It requires less computational power and can be implemented on simple handheld devices. It is commonly used for measuring

the intelligibility of public address (PA) systems in airports and railway stations. According to studies conducted by Gomez et al. (2007) and Zhu et al. (2014a) there are no substantial differences between the STI and STIPA metrics regardless whether the direct or indirect STI measurement method is used. The difference between values seldom exceeds 0.03, equivalent to 1 “just noticeable difference” (JND) for STI scores (Bradley et al. 1999).

5.3.3 Speech Intelligibility Index (SII)

The SII, described in the ANSI/ASA S3.5 standard (ANSI/ASA 2017), was developed as a combination of the AI measure with some version of MTF that accounts for the effects of room reflections. The concept of the SII is based on measuring the SNR in a number of frequency bands, and weighting each SNR by band-importance functions that are dependent on speech material. Four methods of calculating the SII are offered, each using a different number and size of frequency bands. In an order of their accuracy, they are: a) critical bands ($n = 21$), b) one-third octave bands ($n = 18$), c) equally-contributing bands ($n = 17$), and d) octave band ($n = 6$) variants. However, Zhu et al. (2014b) compared SII scores obtained with the octave- and one-third-octave band methods and have found only small differences in the resulting scores. In addition, the SII includes corrections for speech spectrum changes due to changes in vocal effort and to account for upper spread of masking (for more information see ANSI/ASA (2017) and Hornsby (2004)). SII calculations have been reported to account for the effects of reverberation and stationary noise and correlate well with the results of perceptual tests and STI. The score can be adjusted for a person wearing hearing protection devices and also for the presence of visual cues. Since the SII is based on the long-term average spectrum of noise, it is a poor predictor of the effects of nonstationary noise. Similarly, it does not account well for peak-clipping and other similar nonlinear distortions. Several authors proposed various modifications of SII to mitigate these limitations, but they were only partially successful (Kates and Arehart 2005; Rhebergen and Versfeld 2005; Rhebergen et al. 2006).

Unfortunately, the ANSI/ASA standard includes no information about the relationship of SII-derived values to intelligibility scores obtained by human listeners. However, Lyzenga and Rhebergen (2010) reported that the speech recognition threshold corresponds to SII values of approximately 0.22 and to 50% sentence recognition. In addition, Bradley (2003) reported that SII values are about 0.07 higher than AI values calculated for the same speech material and transmission conditions when the AI score is below 0.55.

According to Larm and Hongisto (2006, p. 1117), “the SII and STI are so similar, when used for room acoustical purposes, that the need for many different quantities

is questionable.” Similarly, Zhu et al. (2014b) compared SII and STI metrics in 2 different rooms and at several SNR ratios and found no large differences. However, the authors concluded that “relative loose restriction on measurement conditions in ANSI/ASA S3.5 may reduce the comparability of different measurements” (Zhu et al. 2014b, p. 7). In addition, the SII has been criticized for insufficiently accounting for nonlinear distortions and upward spread of masking (Robinson and Casali 2000; Valimont 2006; Schlesinger and Boone 2010).

In general, technical methods of predicting speech intelligibility are fast and convenient to apply; however, all computed measures have some limitations dependent on the algorithm implemented. Such measures should be used carefully if the communication system uses nonlinear signal processing such as signal compression, level limiting, or phase shift. Therefore, the use of these measures requires operating personnel with significant experience and analytical skills with the ability to identify sources of potentially inaccurate and misleading data.

MIL-STD-1474E (2015), as well as its predecessor MIL-STD-1474D, states that technical measures shall not be used to predict intelligibility of synthetic speech because some key acoustic features are not present in non-human “speech”. Instead, the standard stipulates that the intelligibility of synthetic speech shall be measured using perceptual intelligibility tests. However, it should be noted that Chen and Loizou (2011a) evaluated performance of several speech intelligibility predictors in relation to synthetic speech assessment and found that STI prediction, although not AI prediction, was highly correlated with listeners’ performance ($r = 0.92$). This may suggest that STI, and potentially SII, may be used for the assessment of synthetic speech transmission systems.

All 3 of the technical measures discussed (AI, STI, SII) have also been criticized for being too restrictive and for not taking into account the effects of nonstationary character of noise on speech communication; however, other available measures do account for these effects. For example, the Extended SII algorithm, proposed by Rhebergen and colleagues (2005; 2006), provides provisions for nonstationary noise and seems to be more suitable than the original SII for the prediction of real-world speech intelligibility. Other measures, such as the Dau Model and Glimpse Proportion Metric, are intended to handle the effects of nonstationary noise, but they have not been yet sufficiently tested (Tang and Cooke 2011). It must be stressed that the presence of continuous noise in a communication system represents the worst case scenario; thus, in applications where communication is critical and a conservative estimate of minimum speech intelligibility is required, nonstationary noise algorithms are not appropriate.

In addition, all perceptual and technical measures of speech intelligibility discussed here assume monaural or diotic (the same signal at both ears) speech reception. They have been criticized for not taking into account properties of the binaural (dichotic) hearing (different signals delivered to each ear) although some binaural implementations of perceptual tests have already been tried and some binaural extensions of STI (Schlesinger and Boone 2010) and SII (Beutelmann and Brand 2006) have already been proposed. It has to be noted that binaurally presented speech typically results in greater intelligibility than diotic speech due to the “spatial release from masking” resulting from the spatial separation of the speech source from the noise. For example, Pollack and Picket (1958) reported that, depending on the relative positions of speech and noise sources, stereophonic reproduction of speech in noise provided a 5.5 to 12 dB advantage in comparison to monophonic reproduction. However, in the case of the communication system, the monaural or diotic transmission is the worst case scenario and monaural (diotic) testing should be used as a conservative estimate for critical systems.

6. Common Intelligibility Scale

Equipment designers and users are often confused about the relative merits of the large number of perceptual and technical speech intelligibility measures described in the literature and offered by equipment and software developers. In addition, existing industry and government standards frequently refer to different criteria and test methods, resulting in multiple measurements taken using different scales and tests. If speech intelligibility metrics for 2 different communication systems were obtained using different intelligibility tests, the systems cannot be directly compared unless there is an established relationship between those test scores. Therefore, to establish relationships between the scores obtained with different tests and measures, several authors have proposed or derived graphical or numerical relationships based on the results of studies using 2 (or more) speech intelligibility measures. Some of these relationships have been developed on the basis of a common physical measure (e.g., SNR). Unfortunately, such an approach is very limited in its generalizability since the observed relationships are not transferable. For example, 2 tests may differ in their relationship at a specific SNR if the speech or noise characteristic changes and the same tests may result in different score difference for specific speech and noise conditions if the SNR changes. Thus, the relational functions developed to facilitate comparison of various tests and allow conversion of test scores obtained under specific physical test conditions cannot be translated to other test conditions and test materials.

To facilitate the general comparison of scores obtained with different intelligibility tests under different test conditions, a common denominator is needed. One such common denominator, the Common Intelligibility Scale (CIS), was proposed in the mid-1990s by Barnett and Knight (Barnett and Knight 1994; Knight 1994; Barnett 1999). The CIS corresponds to STI as follows, $CIS = 1 + \log (STI)$. The authors took STI and 6 other common speech intelligibility tests (measures) and using established relationships between them that were published, among others, by Kryter (1970, p. 76) and Steeneken and Houtgast (1980), and expressed them all on a common abstract scale, which they named the CIS. The relationships between these 7 scales expressed as a function of the CIS are shown in Fig. 4. The concept of the CIS is that if under some test conditions a specific intelligibility test results in a given score, the scores for other tests conducted under the same test conditions can be predicted by looking at their values at the same CIS coordinate. In addition, knowing the relationship between one of the speech intelligibility tests plotted in Fig. 4 and a new test is sufficient to predict the relationship of this new test with any of the other tests plotted already in the graph.

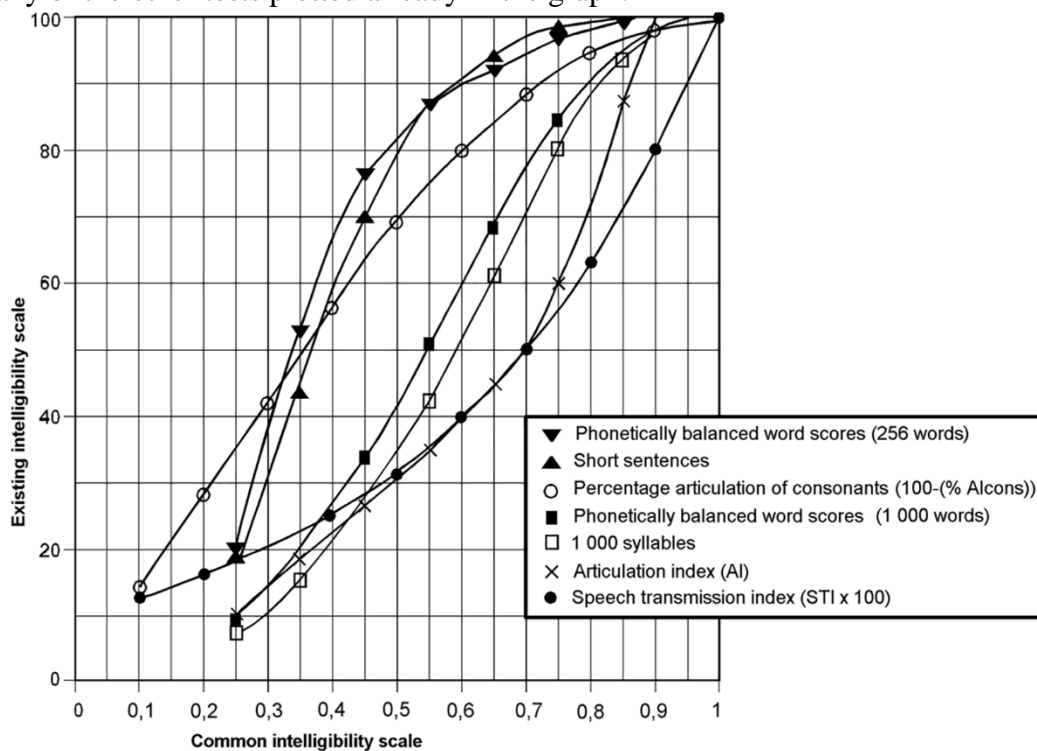


Fig. 4 Relationships between various speech intelligibility scales expressed on the common intelligibility scale (Knight 1994)

According to Knight (1994, p.64), the CIS “was designed to make the chart both readable and usable and renders the word score data as straight lines indicating a linear relationship between the CIS and perceived intelligibility”. Thus, the CIS assumes a linear relationship with the true perceived intelligibility and monotonic nonlinear relation with all test scales represented in Fig. 4. An example of this nonlinear relationship can be observed in the relationship between CIS and STI, which can be expressed mathematically as

$$CIS = 1 + \log(STI), \quad (4)$$

Knight (1994) warns, however, that where the gradients of curves representing specific tests are shallow, the resulting CIS values are not reliable (Knight 1994). Therefore, while CIS may allow an approximate prediction of one score on the basis of another, it cannot be considered an accurate translation procedure in all cases. Further research is needed to improve accuracy and reliability of translations.

In using CIS it should be remembered that most of the data used for deriving the functions shown in Fig. 4 were based on speech-in-noise experiments since noise is the main factor affecting speech intelligibility. More studies with focus on the effects of reverberation and nonlinear speech processing are needed to either fine-tune the existing curves or to derive alternative sets of curves for various environments. Such studies may help to account for some discrepancies between STI and AI (SII) functions shown in Fig. 4 and results of some indirect comparisons between these measures obtained in complex environments (e.g., Payton et al. 1994; Larm and Hongisto 2006; Foster 2015).

More research also needs to be done to establish unique relationships between various test scores for specific languages and to determine whether determined relationships vary as functions of specific speech distortions and listening conditions (Williams and Hecker 1968). For example, Galburn and Kitapci (2016) examined the impact of room acoustics on speech intelligibility of 4 languages (English, Polish, Arabic and Mandarin) and observed large differences among languages in results of both listening and predictive speech intelligibility tests. “English was the most intelligible language under all conditions, and differences with other languages were larger when conditions were poor” (p. 79). Polish and Arabic were particularly sensitive to room conditions.

7. Selecting a Speech Intelligibility Test and Interpreting Its Score

The curves shown in Fig. 4 may guide one in the selection of a speech intelligibility test for a specific application. First, all other things being equal, the ideal test is the one with the greatest discrimination power (i.e., a steep gradient and a small standard deviation) along the CIS in the range of intelligibility being investigated. Thus, according to Fig. 4, the STI offers very good discrimination at high values of speech intelligibility ($CIS > 0.8$) while the ALcons and PBWT (1000 or 256 words) scores have good discrimination at low speech intelligibility ($CIS < 0.5$). In selecting a test for a particular application one should also consider the degree to which the speech materials correspond to the context in which communication will occur, as well as whether reference scores obtained with similar systems are available. The target words should be embedded in a carrier phrase when perceptual assessments of speech intelligibility involve room acoustics to capture the effects of reflections and reverberation (IEC 1998).

Finally, it is useful to interpret the intelligibility score in terms of qualitative categories such as good or bad. Houtgast and Steeneken (1984; 1985) proposed a system of 5 qualitative ranges of speech intelligibility—Excellent/Good/Fair/Poor/Bad—and assigned the STI values to each of them (see also Steeneken and Houtgast 2002). Later on, Knight (1994) assigned the CIS values to these 5 categories. However, these values disagree with the CIS values calculated according to Eq. 4 expressing the relationship between CIS and STI. Table 4 shows the qualitative scale described above with the CIS and STI values assigned to them.

Table 4 Ranges of speech intelligibility and corresponding values on the STI and CIS scales

Speech intelligibility	STI (Houtgast and Steeneken 1984)		CIS (Knight 1994)
	STI	$CIS = 1 + \log(STI)$	
Excellent	> 0.75	> 0.88	> 0.90
Good	0.60–0.75	0.78–0.88	0.70–0.90
Fair	0.45–0.60	0.65–0.78	0.60–0.70
Poor	0.30–0.44	0.48–0.65	0.50–0.60
Bad	< 0.30	< 0.48	< 0.50

The differences between the 2 CIS values given in Table 4 are minimal at the ends of the scale but larger in the middle. When CIS (Knight 1994) values are converted to STI, the differences between these values and the values listed by Houtgast and Steeneken are larger than 0.03 and exceed the uncertainty of STI instruments

(STI=0.02-0.03) (Wijngaarden and Verhave 2002; 2006) and the JND value for the STI scores (Bradley et al. 1999). The greatest difference is for Good/Fair boundary and exceeds 3 JNDs.

Since the relationship expressed in Eq. 4 was the foundation of the CIS, the CIS values assigned to each qualitative range ought to be based on this function, or conversely, a new set of STI ranges should be derived from the Knight's (1994) proposal. A review of literature (e.g., IEC 1998; NFPA 2010) indicates that the former approach is commonly accepted. This relationship is shown graphically in Fig. 5.

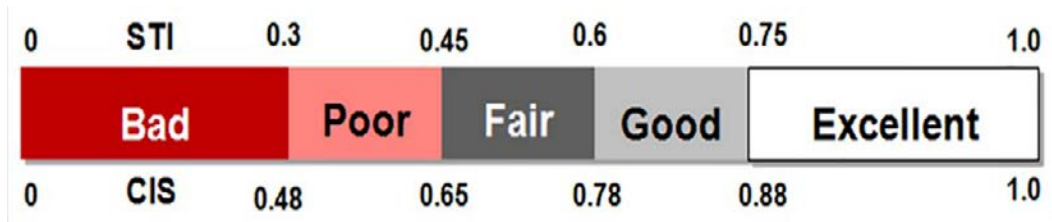


Fig. 5 CIS and STI values defining STI categories proposed by Houtgast and Steeneken (1984; 1985)

Note that the qualitative scale shown in Table 4 and Fig. 5 is actually a MOS scale with references to the CIS (Table 4) and STI scales. This relationship is shown in Fig. 6.

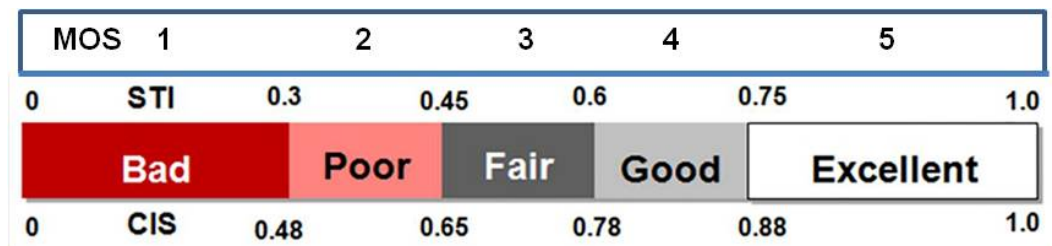


Fig. 6 CIS and STI potential relation with category rating (CCR) scale used as the MOS scale in telephone networks quality evaluation

On the basis of the framework shown in Fig. 6, all other intelligibility scales included in Fig. 4 can be added to the Fig. 6 to show qualitative categories along their extent. New scales can be added in the future. This approach allows one to select a speech intelligibility test for a particular application on the basis of its appropriateness for a given environment, potential speech distortions, test availability (instrumentation, listeners), experience of the test personnel and external restrictions (time, budget) rather than on the basis of past history of tests in similar environments and limited literature references.

The final decision when using a speech intelligibility test for a particular application is the choice of a threshold value of speech intelligibility for a pass/fail decision in determining the “health” of a given environment (instrumentation, human being) from a speech intelligibility point of view. Alternatively, several such threshold values may be needed to classify a given environment to one of the predetermined quality classes or categories.

A commonly recommended criterion for spoken warning alerts in public spaces is a minimum speech intelligibility corresponding to the STI score 0.5 (0.7 CIS). If speech intelligibility varies across the listening area, it is recommended that the average STI score minus one standard deviation be 0.5 STI (0.7 CIS) or greater and at least 90% of the area have STI values of not less than 0.45 (0.65 CIS) (IEC 1998; NFPA 2010). The threshold STI value of 0.5 is also recommended in the ISO 7240 standard, replacing IEC 60849[†]. This standard also lists mean and minimal criterion values for speech intelligibility as measured for other speech tests. These values are shown in Table 5 and all of them correspond to 0.7 CIS (Fig. 4).

Table 5 Required speech intelligibility scores as per ISO 2470:2007 (ISO 2007b)

Measurement method	Required intelligibility score	
	Mean intelligibility	Minimum intelligibility
STI or STIPA	0.50	0.45
PB 256 words (%)	94	91
PB 1000 words (%)	77	68
MRT (%)	94	90
SII	0.50	0.45

Note: Minimum Intelligibility applies to a single location within the space.

It seems like the “mean” specifications in Table 5 represent the normal acceptable speech intelligibility for public communication. Some applications, such as military communication, air traffic control or “red telephone lines,” may require better speech intelligibility to avoid human losses, international conflicts, or public confusion. Therefore, current Federal Aviation Administration (FAA) standard HF-STD-001B (Ahlstrom 2016) and the US military standard MIL-STD-1472 (DOD 1999b; 2015) list values for High, Normal, and Minimal speech intelligibility criteria to be used depending on the criticality of the application. These standards recommend use of the PB, MRT, and AI tests for measuring speech intelligibility and do not provide references to either the CIS or STI scales. A summary of the proposed values is given in Table 6.

[†] The CIS value is no longer listed in ISO 7240 (2007ab).

Table 6 Intelligibility criteria for voice communication systems

Communication requirement	Required intelligibility score		
	PBWT 1000	MRT	AI
High	90	97	0.7
Mean (Normal)	75	91	0.5
Minimal	43	75	0.3

A comparison of Table 6 and Fig. 4 shows that the values of PBWT 1000 (1000 words) and AI scores listed in Table 6 for the High, Normal, and Minimal intelligibility thresholds agree with 0.8 CIS, 0.7 CIS, and 0.5 CIS values, respectively, for the respective test curves shown in Fig. 4. This agreement serves as a basis for establishing the CIS values for High, Normal, and Minimal threshold criteria from the other speech tests shown in Fig. 4 as well as for adding MRT values to the CIS and Fig. 4.

It is tempting to expand Fig. 4 to include other common speech intelligibility tests but it creates a chart that is difficult to follow. Therefore, it seems useful to create an expanded table of recommended High, Mean, and Minimal speech intelligibility threshold scores for many common speech intelligibility measurement scales. Table 7 presents a proposed table created from the data in Figs. 4–6, Tables 4–6, and some additional literature (e.g, Ariöz and Günel 2016).

Table 7 Intelligibility score requirements as a function of communication criticality

Communication requirement	Common intelligibility scale (CIS)	Speech intelligibility tests		Predictors of speech intelligibility		
		MRT	PBWT	AI	STI	SII
High intelligibility; separate syllables understood; MOS ≥ 4	0.78	97%	90%	0.70	0.62	0.72
Normal intelligibility; about 98% of sentences correctly heard; digits understood; MOS ≥ 3.5	0.70	91%	77%	0.50	0.504	0.57
Minimal intelligibility; about 90% of sentences correctly heard; standardized phrases understood; MOS ≥ 2.2	0.50	75%	43%	0.30	0.32	0.37

Test values for SII have been added to those of the CIS, MRT, PBWT, STI, and AI, as SII replaces the AI in the ANSI/ASA standard. The values shown were

determined from several literature studies that compare these tests with others (e.g., Bradley 2003; 2004).

The values provided in Table 7 apply to contexts in which speech comprehension is critical to the user. In contrast, speech privacy is an important design criterion for open offices; therefore, unwanted speech, if still audible, should have low intelligibility, with AI, SII, and STI scores no greater than 0.15, 0.20, and 0.20, respectively (Bradley 2003; Pop and Rindel 2005).

8. Discussion and Conclusions

The transmission of speech within a specific acoustical context or over a communications system depends on the fidelity of the signal reaching the listener. Factors that affect this fidelity include characteristics of the transmitted speech signal itself, the noise masking within the communication medium, and the effects of room acoustics and electronic distortions. The measurement of transmission effectiveness depends on the objective, whether it be to describe its perceived quality, or to assess its intelligibility. Speech quality assessments can be somewhat subjective, whereas intelligibility is more easily quantified as the percentage of speech material recognized. These 2 speech parameters are frequently, but not always, highly correlated. For example, peak-clipping of speech has a detrimental effect on speech quality but it does not affect, and can even slightly improve, speech intelligibility. Therefore, the specific objective, quality or intelligibility, should be clearly stated and communicated to evaluators using perceptual scaling. It is not uncommon to find evaluations for which the intended measure was speech intelligibility, but the resulting assessment only gives estimates of speech quality.

In the case of speech intelligibility scoring, the percentage of speech material properly recognized is highly dependent on the complexity and predictability of the speech materials used, whether they are phonemes, syllables, words, or sentences. Further, the test difficulty depends on the size of the test vocabulary and whether the response set is open or closed. Although perceptual measures of speech intelligibility such as the MRT and PBWT are preferred, because they allow one to directly measure the effects of nonlinear factors on human speech recognition, they are costly in terms of the time and number of listeners required to evaluate a communications system. Technical measures such as AI, STI, and SII provide quick, inexpensive methods for evaluating speech intelligibility that are based on measuring the changes to signal input as a function of the communications medium; however, these are only more or less successful in accounting for temporal factors such as reverberation and distortion. It is suggested that persons measuring speech intelligibility use a common intelligibility scale that cross-references speech

intelligibility scores obtained within the same communication medium for several standard measures. The CIS may not allow a precise translation of speech intelligibility scores from one test to another; however, it does allow one to estimate their approximate relationship. Further, by providing information about each test's relative sensitivity as a function of the test acoustical environment, the CIS may serve as a tool that enables the user to select the optimal test for a given context.

To assess the utility of the CIS, let us consider the following example. There are 2 ANSI/ASA standards that specify methods for measuring speech intelligibility, ANSI/ASA S3.2 (2014; perceptual) and ANSI/ASA S3.5 (2017; predictive). One of the criticisms of ANSI/ASA S3.2 is that the 3 speech sets it specifies do not correspond well to the vocabularies used in operational environments. Currently, the speech intelligibility measure most commonly used by the military is the MRT, and its common use makes it possible to compare current measures with previously made assessments. If the test and evaluation process incorporates the use of untrained, but otherwise representative listeners, such as using Soldiers for evaluations of tactical communications equipment, the use of speech materials containing obscure terminology may result in frustration and a sense that the testing lacks relevancy. Alternative speech intelligibility measures, such as the Callsign Acquisition Test (Rao and Letowski 2006) have been developed with the aim of assessing speech intelligibility using materials that are representative of military communications. ANSI/ASA S3.2 allows for the use of alternative speech materials; however, their use may limit the ability to compare the resulting data to previous assessments. By using the established relationship of these measures to a standardized measure such as the MRT and the CIS, data from these alternative measures can be interpreted and approximate equivalencies posited (Blue-Terry and Letowski 2011; Blue-Terry et. al 2012; Figs. 7 and 8). Since communications systems are often chosen on the basis of many factors, only one of which is speech intelligibility, these alternative measures may have a valid role to establish sufficient performance for use, or to screen among systems under consideration. Thus the ability to relate these scores to more widely used measures can serve as a practical cost saving measure.

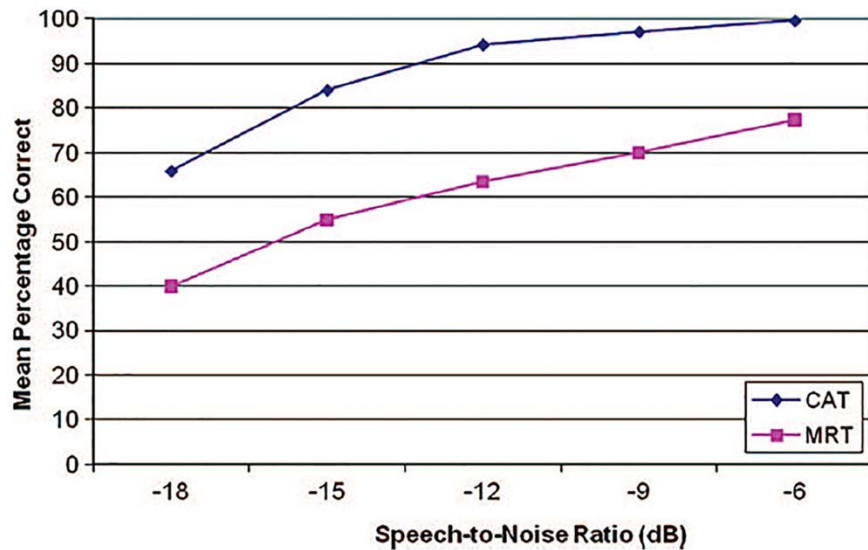


Fig. 7 Performance intensity functions obtained for the callsign acquisition test (CAT) and the MRT in white noise (Blue-Terry and Letowski 2011)

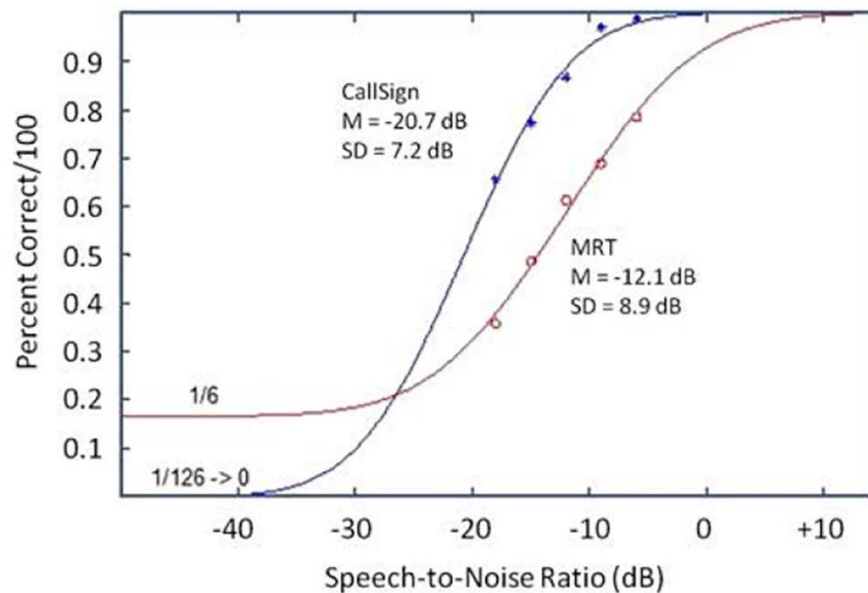


Fig. 8 Theoretical shapes of MRT and CAT performance intensity functions derived for data from the CAT and MRT speech intelligibility measures (Blue-Terry et al. 2012)

Additionally, and importantly, the utility of the CIS is that it frequently allows one to compare results of 2 tests with no need to refer to the test conditions. For example, comparing 2 speech intelligibility tests in 2 different types of noise will most likely result in a different relationship between these tests as a function of SNR for each type of noise. Replacing the SNR with CIS allows making this relationship unique and independent of noise type. With respect to the CIS, the difference “X” between 2 tests of speech intelligibility will always be the same,

although they may refer to different SNRs due to characteristics of the noise masker. It does not mean that the CIS should serve as a replacement of SNR when presenting data, but it provides a context for the data that may enable decision-making in some cases.

Alternative measures, such as CAT, have a valid role in establishing adequate intelligibility for use, or to screen among systems under consideration. Thus the ability to relate their scores to more widely used measures, and to do this in a generalized form, can serve as a practical cost saving measure and may enable good decisions.

9. References

- Ahlstrom, V. Human factors design standard (DOT/FAA/HF-STD-001B). Atlantic City (NJ): Federal Aviation Administration William J. Hughes Technical Center; 2016.
- Allen C, Nikolopoulos TP, Dyar D, O'Donoghue GM. Reliability of a rating scale for measuring speech intelligibility after pediatric cochlear implementation. *Otol Neurotol*. 2001;22(5):631–633.
- Allen JB. The articulation index is a Shannon channel capacity. In: Pressnitzer D, de Cheveigne A, McAdams Collet SL, editors. *Auditory signal processing*. New York (NY): Springer Verlag; 2003. p. 314–320.
- ANSI S3.5-1969. (R1986) Methods for the calculation of the articulation index. New York (NY): American National Standards Institute; 1969.
- ANSI/ASA S3.2-2009 (R2014). American national standard method for measuring the intelligibility of speech over communication systems. Melville (NY): Acoustical Society of America; 2014.
- ANSI/ASA S3.5-1997 (R2017). American national standard methods for calculation of the speech intelligibility index. Melville (NY): Acoustical Society of America; 2017.
- Ariöz, U, Günel, B. Evaluation of hearing loss simulation using a speech intelligibility index. *Turk J Electr Eng Co*. 2016;24:4193–4207.
- Barnett PW, Knight RD. The common intelligibility scale. *Proceedings of the Institute of Acoustics*. 1994;17(7):201–206.
- Barnett PW. Conversion of RASTI to ALcons: a robust and reversible process? *Proceedings of the Institute of Acoustics*. 1997;19(6):115–133.
- Barnett PW. Overview of speech intelligibility. *Proceedings of the Institute of Acoustics*. 1999;21(5):1–16.
- Beerends J, Larsen E, Iyer N, van Vugt J. Measurement of speech intelligibility based on the PESQ approach. *Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN)*. Prague Czech Republic; 2004.
- Beerends J, van Wijngaarden S, van Buuren R. Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with vocoders. *Proceedings of the Workshop New Directions for Improving Audio*

- Effectiveness RTO-MP-HFM-123. Neuilly-sur-Seine (France): RTO, 2005. p. 10.1–10.6.
- Beerends J, van Buuren RA, van Vugt JM, Verhave JA. PESQ based speech intelligibility measurements. Proceedings of the International Conference on Acoustics including the German 35th Annual Conference on Acoustics NAG/DAGA. Rotterdam (Holland): TNO; 2009.
- Beranek LL. The design of speech communication systems. Proceedings of the Institute of Radio Engineers. 1947;35:880–890.
- Berger EH, Royster LH, Royster JD, Driscoll DP, Layne M. The noise manual. 5th ed. Fairfax (VA): AIHA Press; 2003.
- Beutelmann R, Brand, T. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J Acoust Soc Am. 2006;120(1):331–342.
- Bilger RC. Speech recognition test development. In: Elkins E, editor. ASHA Reports, No. 14: Speech recognition by the hearing impaired. Rockville (MD): American Speech-Language-Hearing Association; c1984. p. 2–7.
- Blue-Terry M, Letowski T. Effects of white noise on Callsign acquisition test and modified rhyme test scores. Ergonomics. 2011;54(2):139–145.
- Blue-Terry, M, McBride M, Letowski T. Effects of speech intensity on the Callsign acquisition test (CAT) and modified rhyme test (MRT) presented in noise. Arch Acoust. 2012;37(2):199–203.
- Boothroyd A, Nitttrouer S. Mathematical treatment of context effect in phoneme and word recognition. J Acoust Soc Am. 1988;84 (1):101–114.
- Bradley JS. The acoustical design of conventional open plan offices. Can Acoust. 2003;31(2): 23–30.
- Bradley JS. Acoustical design for open-plan-offices. Construction Technology Update No. 63. Ottawa (Canada): NRC-CNRC, 2004.
- Bradley JS, Reich R, Norcross SG. A just noticeable difference in C50 for speech. Appl Acoust. 1999;58(2):99–108. [https://www.nrc-cnrc.gc.ca/ctu-sc/ctu_sc_n63].
- Byrne D, Cotton S. Evaluation of the National Acoustic Laboratories' new hearing aid selection procedure. J Speech Hear Res. 1988;31(2): 178–186.

- Chen F, Loizou PC. Predicting the intelligibility of vocoded speech. *Ear Hearing*. 2011a;32(2):331–338.
- Chen F, Loizou PC. Predicting the intelligibility of vocoded and wideband Mandarin Chinese. *J Acoust Soc Am*. 2011b;129(5):3281–3290.
- Cienkowski KM, Speaks C. Subjective vs. objective intelligibility of sentences in listeners with hearing loss. *J Speech Lang Hear R*. 2000;43(5):1205–1210.
- Côté N. Integral and diagnostic intrusive prediction of speech quality. Berlin (Germany): Springer Verlag; 2011.
- Cox RM, McDaniel DM. Development of the speech intelligibility rating (SIR) test for hearing aid comparisons. *J Speech Hear Res*. 1989;32:347–352.
- Cox RM, Alexander GC, Rivera IM. Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners. *J Speech Hear Res*. 1991;34:904–915.
- De Bodt MS, Nernandez-Diuaz Huici ME, Van De Heyning PH. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*. 2002;35:283–292.
- Egan JP. Articulation testing methods. *Laryngoscope*. 1948;58(9):955–991.
- Eisenberg LS, Dirks DD, Gornbein JA. Subjective judgments of speech clarity measured by paired comparisons and category rating. *Ear Hearing*. 1997;18(4):2954–306.
- Eisenberg LS, Dirks DD, Takayanagi S, Martinez AS. Subjective judgments of clarity and intelligibility for filtered stimuli with equivalent speech intelligibility index predictions. *J Speech Lang Hear R*. 1998;41(2):327–339.
- Eisler H. Measurement of perceived acoustic quality of sound-reproducing systems by means of factor analysis. *J Acoust Soc Am*. 1966;39(3):484–492.
- Epstein A, Giolas TG, Owens E. Familiarity and intelligibility of monosyllabic word lists. *J Speech Hear Res*. 1968;11: 435–438.
- Fairbanks G. Systematic research in experimental phonetics. I. A theory of the speech mechanism as a servosystem. *J Speech Hear Disord*. 1954;19(2):133–139.
- Fairbanks G. Test of phonemic differentiation: the rhyme test. *J Acoust Soc Am*. 1958;30(7):596–600.

- Fletcher H. Speech and hearing in communication. 2nd Ed. Oxford (England): D Van Nostrand; 1953.
- Fletcher H, Steinberg JC. Articulation testing methods. *Bell Syst Tech J*. 1929;9(4):806–854.
- Fontan J, Tardieu J, Gaillard P, Woisard V, Ruiz R. Relationship between speech intelligibility and speech comprehension in babble noise. *J Speech Lang Hear R*. 2015;58:977–986.
- Foster J. Speech intelligibility in cockpit voice recorders [master's thesis]. [Baltimore (MD)]: John Hopkins University; 2015.
- French NR, Steinberg JC. Factors governing the intelligibility of speech sounds. *J Acoust Soc Am*. 1947;19(1):90–119.
- Galbrun L, Kitapci K. Speech intelligibility of English, Polish, Arabic and Mandarin under different room acoustic conditions. *Appl Acoust*. 2016;114:79–91.
- Gerber SE. Introductory hearing science: physical and psychological concepts. Philadelphia (PA): WB Saunders Company; 1974.
- Goldberg R, Riek L. The practical handbook of speech coders. Boca Raton (FL): CRC Press; 2000.
- Goldsworthy RL, Greenberg JE. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J Acoust Soc Am*. 2004;116 (6):3679–3689.
- Gomez L, Nestoras C, Dance S, Murano S. Comparison of speech intelligibility measurements in a diffuse space. *Proceedings of the 14th International Congress on Sound and Vibrations*. Cairns (Australia): IISV, 2007. p. 1–8.
- Goodman DJ, Goodman JS, Chen M. Intelligibility and ratings of digitally coded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978; 26(5):403–409.
- Greenspan SL, Bennett RW, Syrdal AK. An evaluation of the diagnostic rhyme test. *Int J Speech Tech*. 1998;2:201–214.
- Hilkhuyzen G, Lloyd J, Huckvale, MA. Effects of replay on the intelligibility of noisy speech. *Proceedings of the 46th AES International Conference on Audio Forensics–Recording, Recovery, Analysis, and Interpretation*; 2012 June 14–16; Denver (CO). New York (NY): Audio Engineering Society; c2012. p. 50–56.

- Hirsh IJ, Davis H, Silverman SR, Reynolds EG, Eldert E, Benson WR. Development of materials for speech audiometry. *J Speech Hear Disord*. 1952;17:321–337.
- Hornsby BWY. The speech intelligibility index: what is it and what's it good for? *Hearing J*. 2004;57(10):10–17.
- House AS, Williams CE, Hecker MHL, Kryter KD. Articulation testing methods: consonantal differences with a closed-response set. *J Acoust Soc Am*. 1965;37(1):158–166.
- Houtgast T, Steeneken HJM. A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria. *Acustica*. 1984;54(4):185–199.
- Houtgast T, Steeneken HJM. The modulation transfers function in room acoustics. Technical Review No 3. Marlborough (MA): Brüel & Kjær Instruments; 1985.
- Humes LE, Dirks DD, Bell TS, Ahlstrom C, Kincaid GE. Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners. *J Speech Hear Res*. 1986 Dec;29(4): 447–462.
- Hustad KC. The relationship between listener comprehension and intelligibility scores for speakers with dysarthia. *J Speech Lang Hear R*. 2008;51(3):562–573.
- IEEE No. 297-1969. IEEE recommended practice for speech quality measurements. New York (NY): 1969. Institute of Electrical and Electronics Engineers; c1969.
- IEC 60849:1998. Sound systems for emergency purposes. Geneva (Switzerland): International Electrotechnical Commission; 1998. [withdrawn and replaced with ISO 7240 m parts 16 and 19].
- IEC 60286-16:2011. Sound system equipment – part 16: objective rating of speech intelligibility by speech transmission index. Geneva (Switzerland): International Electrotechnical Commission; 2011.
- ISO 9921:2003. Ergonomics – assessment of speech communication. Geneva (Switzerland): International Organization for Standardization; 2003.
- ISO 7240-16:2007. Fire detection and alarm systems – part 16: sound system control and indicating equipment. Geneva (Switzerland): International Organization for Standardization; 2007a.

- ISO 7240-19:2007. Fire detection and alarm systems – part 19: design, installation, commissioning, and service of sound systems for emergency purposes. Geneva (Switzerland): International Organization for Standardization; 2007b.
- ITU-T Recommendation P.800. Methods for subjective determination of transmission quality. Geneva (Switzerland): International Telecommunication Union; 1996.
- ITU-T Recommendation P.800.1. Mean opinion score (MOS) terminology. Geneva (Switzerland): International Telecommunication Union; 2003.
- ITU-T Recommendation P.310. Transmission characteristics for telephone-band (300-3400 Hz) digital telephones. Geneva (Switzerland): International Telecommunication Union; 2003.
- Jacob KD, Birkle TK, Icker CB. Accurate prediction of speech intelligibility without the use of in-room measurements. *J Audio Eng Soc.* 1991;39(4):232–242.
- Jakobson R, Fant CG, Halle M. Preliminaries to speech analysis: the distinctive features and their correlates. MIT Acoustics Laboratory technical report no. 13. Cambridge (MA): The MIT Press; 1952.
- Jekosch U. Voice and speech quality perception: assessment and evaluation. Secaucus (NJ): Springer-Verlag New York Inc; 2005.
- Kates J, Arehart K. Coherence and the speech intelligibility index. *J Acoust Soc Am.* 2005;117:2224–2237.
- Keller E. The analysis of voice quality in speech processing. In: Chollet G, Esposito A, Faundez-Zanuy M, Marinaro M, editors. *Nonlinear speech processing and applications. Lecture notes in computer science; Vol 3445.* Berlin (Germany): Springer-Verlag Berlin Heidelberg; 2005. p. 54–73.
- Knight RD. The common intelligibility scale. *Proceedings of the Institute of Sound and Communication Engineers.* 1994;5:60–67.
- Kondo K. Estimation of speech intelligibility using perceptual speech quality scores. In: Ipsic I, editor. *Speech and language technologies.* Rijeka (Croatia): InTech; 2011 June. p. 154–176. [Available from: <http://222intechopen.com/books/speech-and-language-technologies/estimation-of-speech-intelligibility-usingperceptual-speech-quality-scores>].
- Kraft V, Portele T. Quality evaluation of five German speech synthesis systems. *Acta Acustica.* 1995;3:51–365.

- Kryter KD. Methods for the calculation and use of the articulation index. *J Acoust Soc Am.* 1962a;34(11):1689–1697.
- Kryter KD. Validation of the articulation index. *J Acoust Soc Am.* 1962b;34(11):1698–1702.
- Kryter KD. *The effects of noise on man.* New York (NY): Academic Press, 1970.
- Larm P, Hongisto V. Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index. *J Acoust Soc Am.* 2006;119(2):1106–1117.
- Leijon A, Lindkvist A, Ringdahl A, Israelsson B. Sound quality and speech reception for prescribed hearing aid frequency responses. *Ear Hearing.* 1991;12(4):251–260.
- Letowski T. Timbre, tone color, and sound quality: concepts and definitions. *Arch Acoust.* 1992;17(1):17–30.
- Letowski T, Frank T, Caravella J. Acoustical properties of speech produced in noise. *Ear and Hearing.* 1993;14(5):332–338.
- Lewis JR. The revised mean opinion scale (MOS-R): preliminary psychometric evaluation. West Palm Beach (FL): IBM Voice Systems; 2001 Mar 27. IBM Report No.: TR 29.3414.
- Logan JS, Greene BG, Pisoni DB. Segmental intelligibility of synthetic speech produced by rule. *J Acoust Soc Am.* 1989;86 (2):566–581.
- Ludvigsen C, Elberling C, Keidser G. Evaluation of a noise reduction method – comparison of observed scores and scores predicted from STI. *Scand Audiol.* 1993;38:50–55.
- Lyzenga J, Rhebergen KS. Auditory model for the speech audiogram. Paper presented at: 2nd Workshop on Speech in Noise: Intelligibility and Quality; 2010 Jan 7–8; Amsterdam (The Netherlands). [http://www.phon.ucl.ac.uk/events/quality2010/talks/Johannes_Lyzenga.pdf]. [accessed on 2013 July 1].
- McDaniel DM, Cox RM. Evaluation of the speech intelligibility rating (SIR) test for hearing aid comparisons. *J Speech Hear Res.* 1992;35:686–693.
- Mechergui N, Djaziri-Larbi S, Jaidane M. Speech based transmission index for all: an intelligibility metric for variable hearing ability. *J Acoust Soc Am.* 2017;141(3):1470–1484.

- MIL-HDBK-1908B. Department of Defense handbook: definitions of human factors terms. Redstone Arsenal (AL): US Army Aviation and Missile Command. DOD; 1999 Aug 16.
- MIL-STD-1472F. Department of Defense design criteria standard: human engineering. Redstone Arsenal (AL): DOD; 1999b.
- MIL-STD-1472G. Department of Defense design criteria standard: human engineering. Redstone Arsenal (AL): DOD; 2012.
- MIL-STD-1474E. Department of Defense design criteria standard: noise limits. Redstone Arsenal (AL): DOD; 2015.
- Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test material. *J Exp Psychol.* 1951;41:329–335
- Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev.* 1956;63: 81–97.
- Moore BCJ. An introduction to the psychology of hearing. 2nd Ed. London (England): Academic Press; 1977.
- Mueller HG, Killion MC. An easy method for calculating the articulation index. *Hearing J.* 1990;43(9):1–4.
- Munson WA, Karlin JE. ISO-preference method for evaluating speech transmission circuits. *J Acoust Soc Am.* 1962;34 (6):762–774.
- NASA/SP-2010-3407. Human integration design handbook (HIDH). Washington (DC): National Aeronautical and Space Administration; 2010.
- NFPA 72. National fire alarm and signaling code handbook. Quincy (MA): National Fire Protection Association; 2010.
- Nye PW, Gaitenby JH. Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). Haskins Laboratories status report on speech research. New Haven (CT): Haskins Laboratories; c1973. SR-33, p. 77–91.
- Olsen W, Van Tassell DJ, Speaks CE. Phoneme and word recognition for words in isolation and in sentences. *Ear Hearing.* 1997;18 (3):175–188.
- Onaga H, Furuy Y, Ikeda T. The disagreement between speech intelligibility index (STI) and speech intelligibility. *Acoust Sci Tech.* 2001;22(4):265–271.
- Osgood C, Suci G, Tannenbaum P. The measurement of meaning. Chicago (IL): University of Illinois Press; 1957.

- Owens E, Schubert ED. Development of California consonant test. *J Speech Hear Res.* 1977;20(3):463–474.
- Payton KL, Uchanski RM, Braida LD. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J Acoust Soc Am.* 1994;95(3):1581–1592.
- Payton KL, Braida LD. Effect of slow-acting wide dynamic range compression on measures of intelligibility and ratings of speech quality in simulated-loss listeners. *J Speech Lang Hear R.* 2005;48(3):702–714.
- Peutz VMA. Articulation loss of consonants as a criterion for speech transmission in a room. *J Audio Eng Soc.* 1971;19 (11):23–27.
- Pollack I, Pickett JM. Stereophonic listening and speech intelligibility against voice babble. *J Acoust Soc Am.* 1958;30 (2):131–133.
- Pop CB, Rindel JH. Perceived speech privacy in computer simulated open-plan offices. *Proceedings of the 2005 Congress and Exposition on Noise Control Engineering.* Rio de Janeiro (Brazil): Ince; 2005. p. 1–7.
- Preminger JE, Van Tasell DJ. Qualifying the relation between speech quality and speech intelligibility. *J Speech Hear Res.* 1995;38(5):714–725.
- Punch JL, Parker CA. Pairwise listener preferences in hearing aid evaluation. *J Speech Hear Res.* 1981;25:366–374.
- Purdy SC, Pavlovic CV. Reliability, sensitivity and validity of magnitude estimation, category scaling and paired-comparison judgements of speech intelligibility by older listeners. *Audiology.* 1992;31(5):254–271.
- Raffin MJM, Schafer D. Application of a probability model based on the binomial distribution to speech-discrimination scores. *J Speech Lang Hear R.* 1980;23:570–575.
- Rankovic CM, Levy RM. Estimating articulation scores. *J Acoust Soc Am.* 1997;102(6):3754–3761.
- Rao MD, Letowski T. Callsign acquisition test (CAT): speech intelligibility in noise. *Ear Hearing.* 2006;27(2):120–128.
- Reinhart P, Souza P. Intelligibility and clarity of reverberant speech: effects of wide dynamic range compression release time and working memory. *J Speech Lang Hear R.* 2016;59(6):1543–1554.

- Rhebergen KS, Versfeld NJ. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J Acoust Soc Am*. 2005;117(4):2181–2192.
- Rhebergen KS, Versfeld NJ, Dreschler WA. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J Acoust Soc Am*. 2006;120(6): 3988–3997.
- Robinson GS, Casali JG. Speech communication and signal detection in noise. In: Berger EH, Royster LH, Royster DP, Driscoll DP, Layne M, editors. *The noise manual*. Fairfax (VA): American Industrial Hygiene Association; 2000. p. 567–600.
- Rodman J. The effect of bandwidth on speech intelligibility [white paper]. San Jose (CA): Polycom Inc; 2003 Jan 16 [accessed 2013 July 1]. <http://suntelnetworks.com/HDVoiceStudy.pdf>.
- Rothausen EH, Urbanek GE. New reference signal for speech-quality measurements (A). *J Acoust Soc Am*. 1965;38(5): 940.
- Rothausen EH, Chapman WD, Guttman N, Hecker MHL, Nordby KS, Silbiger HR, Urbanek GE, Weinstock M. IEEE recommended practice for speech quality measurements. *IEEE Trans Audio Electroacoust*. 1969;17(3): 225–246; also *IEEE 297*: 1969; (withdrawn).
- Rubenstein H, Pollack I. Word predictability and intelligibility. *J Verb Learn Verb Be*. 1963;2:147–158.
- Schlesinger A, Boone MM. (2010). Speech intelligibility assessment in binaural and nonlinear hearing aids. Poster presented at: 2nd Workshop on Speech in Noise: Intelligibility and Quality; 2010 Jan 7–8; Amsterdam (The Netherlands). [http://www.phon.ucl.ac.uk/events/quality2010/posters/Anton_Schlesinger.pdf]. [accessed on 2013 July 1].
- Schmidt-Nielsen A. Intelligibility and acceptability testing for speech technology. Washington (DC): Naval Research Laboratory; 1992 May 22. Report No.: NRL/FR/5530-92-9379. Also: Schmidt-Nielsen A. Intelligibility and acceptability testing for speech technology. In: Syrdal A, Bennett R, Greenspan S, editors. *Applied speech technology*. Boca Raton (FL): CRC Press; 1994. Chapter 5; p 501–520.
- Schroeder MR. Modulation transfer functions: definition and measurement. *Acustica*. 1981;49(3):179–182.

- Shannon CE. The mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423 (parts I & II) and 623-656 (part III).
- Speaks C, Parker B, Harris C, Kuhl P. Intelligibility of connected discourse. *J Speech Hear Res*. 1972;15:590–602.
- Staab W. Significance of mid-frequencies in hearing aid selection. *Hearing J*. 1988 June;4:6–23.
- Steeneken HJM, Houtgast T. A physical method for measuring speech-transmission quality. *J Acoust Soc Am*. 1980;67(1):318–326.
- Steeneken HJM, Houtgast T. Validation of the revised STI_r method. *Speech Commun*. 2002;38:413–425.
- Stevens SS, Abrams MH, Goffard SJ, Miller J. Subjective ratings of intelligibility of talkers in noise. In: *Speech in noise: a study of the factors determining its intelligibility*. OSRD Report No. 4023. Cambridge (MA): Harvard Psycho-Acoustic Laboratory; 1944.
- Studebaker GA. A rationalized arcsine transform. *J Speech Hear Res*. 1985;28(3):455–462.
- Studebaker GA, McDaniel DM, Sherbecoe RL. Evaluating relative speech recognition performance using the proficiency factor and rationalized arcsine differences. *J Am Acad Audiol*. 1995;6:173–182.
- Sullivan JA, Levitt H, Hwang JY, Hennessey AM. An experimental comparison of four hearing aid prescription methods. *Ear Hearing*. 1988;9(1):22–32.
- Summers WV, Pisoni DB, Bernacki RH, Pedlow RI, Stokes MA. Effects of noise on speech production: acoustic and perceptual analyses. *J Acoust Soc Am*. 1988;84(3):917–929.
- Tang Y, Cooke M. Subjective and objective evaluation of speech intelligibility enhanced under constant energy and duration constraints. *Interspeech 2011. Proceedings of the 12th Annual Conference of the International Speech Communication Association*; 2011 Aug 28–31; Florence (Italy).
- Thornton AR, Raffin MJM. Speech-discrimination scores modeled as a binomial variable. *J Speech Hear Res*. 1978;21:507–518.
- TIA-810-B. Transmission requirements for narrowband digital telephones. Arlington (VA): Telecommunications Industry Association; 2006.

- TIA-920.000-B. Overview of transmission requirements for digital interface communications devices. Arlington (VA): Telecommunications Industry Association; 2015a.
- TIA-920.110-B. Transmission requirements for digital interface communications devices with handsets. Arlington (VA): Telecommunications Industry Association; 2015b.
- TIA-920.130-A. Transmission requirements for wideband digital wireline telephones with headset. Arlington (VA): Telecommunications Industry Association; 2011.
- Tillman TW, Carhart R. An expanded test for speech discrimination utilizing CNC monosyllabic words (Northwestern University auditory test no. 6). Brooks Air Force Base (TX): USAF School of Aerospace Medicine; 1966 Jun 1–12. Report No.: SAM-TR-66-55.
- Traunmüller H, Eriksson A. Acoustic effects of variation in vocal effort by men, women, and children. *J Acoust Soc Am*. 2000;107(6):3438–3451.
- Valimont RB. Active noise reduction versus passive designs in communication headsets: speech intelligibility and pilot performance effects in an instrument flight simulation. Doctoral dissertation. Blacksburg (VA): Virginia Polytechnic Institute and State University; 2006.
- van Schoonhoven J, Rhebergen KS, Dreschler WA. Towards measuring the speech transmission index in fluctuating noise: accuracy and limitations. *J Acoust Soc Am*. 2017;141(2):818.
- Viswanathan M, Viswanathan M. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Comput Speech Lang*. 2005;19:55–83.
- Voiers WD. Diagnostic evaluation of speech intelligibility. In: Hawley ME, editor. *Speech intelligibility and speaker recognition*. Stroudsburg (PA): Dowden, Hutchinson & Ross; c1977. p. 374–387.
- Voiers WD. Evaluating processed speech using the diagnostic rhyme test. *Speech Tech*. 1983;1(4):30–39.
- Waidyanatha N, Wilfred T, Perera K, Silva M. Mean opinion score performance in classifying voice-enabled emergency communication systems. ICCIS 2012. Proceedings of the 2012 International Conference on Computer & Information Science; 2012 June 12–14; Kuala Lumpur (Malaysia): New York (NY): IEEE; c2012. p. 676–682.

- Warr PB, Knapper Ch. The perception of people and systems. New York (NY): Wiley & Sons Book Co.; 1968.
- Webster JC, Allen CR. Speech intelligibility in naval aircraft radios. San Diego (CA): Naval Electronics Laboratory Center; 1972 Aug. Report No.: NELC-TR-1830.
- Wijngaarden SJ, Verhave JA. Measurement and predictions of speech intelligibility in traffic tunnels using the STI. In: Houtgast T, Steeneken, HJM, van Wijngaarden, editors. Proceedings of the International Symposium on STI, TNO Human Factors; 2002; Soesterberg (The Netherlands). Past, present and future of the transmission index, p. 113.
- Wijngaarden SJ, Verhave JA. Predictions of speech intelligibility for public address systems in traffic tunnels. *Appl Acoust.* 2006;67:306–323.
- Williams CE, Hecker MHL. Selecting an intelligibility test for communication system evaluation (A). *J Acoust Soc Am.* 1967;42(5):1198.
- Williams CE, Hecker MHL. Relation between intelligibility scores for four test methods and three types of speech distortion. *J Acoust Soc Am.* 1968;44(4):1002–1006.
- Yorkston KM, Beukelman DR. Comparison of techniques for measuring intelligibility of dysarthric speech. *J Commun Disord.* 1978;11:499–512.
- Yorkston KM, Strand E, Kennedy M. Comprehensibility of dysarthric speech: investigations for assessment and treatment planning. *Am J Speech-Lang Pat.* 1996;51(3):55–66.
- Zhou H, Chen Z, Shi H, Wu Y, Yin S. Categories of auditory performance and speech intelligibility ratings of early-implanted children without speech training. *PLoS ONE.* 2013;8(1).
- Zhu P, Mo F, Kang J. Experimental comparison between direct and indirect measurement methods for the objective rating of speech intelligibility. Proceedings of the 21st International Congress on Sound and Vibrations. Beijing (China): IISV; 2014a. p. 1–8.
- Zhu P, Mo F, Kang J. Experimental comparison between STI and SII metrics for the objective rating of speech intelligibility. Proceedings of the 21st International Congress on Sound and Vibrations, Beijing (China): IISV; 2014b. p. 1–7.

List of Symbols, Abbreviations, and Acronyms

AI	articulation index
ALcons	articulation loss of consonants
ANSI	American National Standards Institute
ASA	Acoustical Society of America
CAT	callsign acquisition test
CCR	comparison category rating
CID	Central Institute for the Deaf
CIS	common intelligibility scale
CQS	conversational-quality scale
CVC	consonant-vowel-consonant
dB	decibel
dB(A)	A-weighted decibels
DCR	degradation category rating
DOD	Department of Defense
DRT	diagnostic rhyme test
FAA	Federal Aviation Administration
INT	intelligibility
ITU	International Telecommunication Union
JND	just noticeable difference
LES	listening-effort scale
LQS	listening-quality scale
LPS	loudness-preference scale
MOS	mean opinion score
MOS-LQS	MOS-listening-quality scale
MRT	modified rhyme test

MTF	modulation transfer function
PA	public address
PAMS	perceptual analysis and measurement system
PB	phonetically balanced
PBWT	phonetically balanced word test
PESQ	perceptual evaluation of speech quality
PQS	picture-quality scale
PSQM	perceptual speech quality measurement
QoS	quality of service
RASTI	rapid speech transmission index
SD	standard deviation
SII	speech intelligibility index
SIR	speech intelligibility rating
SIS	speech intelligibility scale
SNR	speech-to-noise ratio/signal-to-noise ratio
STI	speech transmission index
STIPA	sound transmission index public address
STI _r	speech transmission index-revised
TNO	Netherlands Organisation for Applied Scientific Research (Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek)
WDRC	wide dynamic range compression
VoIP	voice over internet protocol

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIR ARL
(PDF) IMAL HRA
RECORDS MGMT
RDRL DCL
TECH LIB

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 ARL
(PDF) RDRL HRB B
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

8 ARL
(PDF) SFC PAUL RAY SMITH CTR
RDRL HRO COL H BUHL
RDRL HRF J CHEN
RDRL HRA I MARTINEZ
RDRL HRR R SOTTILARE
RDRL HRA C A RODRIGUEZ
RDRL HRA B G GOODWIN
RDRL HRA A C METEVIER
RDRL HRA D B PETTIT
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 USA ARMY G1
(PDF) DAPE HSI B KNAPP
300 ARMY PENTAGON
RM 2C489
WASHINGTON DC 20310-0300

1 USAF 711 HPW
(PDF) 711 HPW/RH K GEISS
2698 G ST BLDG 190
WRIGHT PATTERSON AFB OH
45433-7604

1 USN ONR
(PDF) ONR CODE 341 J TANGNEY
875 N RANDOLPH STREET
BLDG 87
ARLINGTON VA 22203-1986

1 USA NSRDEC
(PDF) RDNS D D TAMILIO
10 GENERAL GREENE AVE
NATICK MA 01760-2642

1 OSD OUSD ATL
(PDF) HPT&B B PETRO
4800 MARK CENTER DRIVE
SUITE 17E08
ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

12 ARL
(PDF) RDRL HR
J LOCKETT
P FRANASZCZUK
K MCDOWELL
K OIE
RDRL HRB
D HEADLEY
RDRL HRB C
J GRYNOVICKI
RDRL HRB D
C PAULILLO
RDRL HRF A
A DECOSTANZA
RDRL HRF B
A EVANS
RDRL HRF C
J GASTON
RDRL HRF D
A MARATHE
A SCHARINE

1 TSPID
(PDF) US ARMY – NATICK SOLDIER
RSRCH & DEV CTR
M MARKEY

1 TSPID
(PDF) US ARMY – NATICK
SOLDIER
RSRCH & DEV CTR
D LEE

1 US ARMY - NATICK SOLDIER
(PDF) RSRCH & DEV CTR
A CHISHOLM

1 US ARMY - NATICK SOLDIER
(PDF) RSRCH & DEV CTR
S GERMAIN

1 TSPID
(PDF) US ARMY - NATICK SOLDIER
RSRCH & DEV CTR
J KRUSZEWSKI

1 RSRCH PSYCHOLOGIST
(PDF) DIRECTOR, AIRCREW
PROTECTION DIV
US ARMY AEROMEDICAL
RSRCH LAB
W AHROON PHD

1 WALTER REED NATIONAL
(PDF) MILITARY
MEDICAL CTR
AUDIOLOGY AND SPEECH
CTR
D BRUNGART

1 SENIOR ENGR RSRCH
(PDF) PSYCHOLOGIST
BATTLESPACE
ACOUSTICS BR
B SIMPSON

1 BATTLESPACE
(PDF) ACOUSTICS BR
AFRL WPAFB US
N IYER

1 BATTLESPACE
(PDF) ACOUSTICS BR
ER THOMPSON

1 BATTLESPACE
(PDF) ACOUSTICS BR
G ROMIGH

1 US AFRL
(PDF) R MCKINLEY

1 STATE COLLEGE PA
(PDF) T LETOWSKI

1 VIRGINIA TECH UNIV
(PDF) J CASALI

1 SENIOR RSRCH
(PDF) ADMINISTRATOR
DOD HEARING CTR
OF EXCELLENCE
T HAMMILL

1 PROGRAM MGR
(PDF) DOD HEARING CTR
OF EXCELLENCE
K BUCHANAN

1 ARMY PUBLIC HEALTH
(PDF) CMND (PROVISIONAL)
LTC M ROBINETTE

1 DRDC TORONTO ARMY
(PDF) LIAISON OFC
HUMAN-COMPUTER
INTERACTION GROUP
G ARRABITO

1 UNIV OF COLORADO
(PDF) K AREHART

1 NASA
(PDF) J ALLEN

1 FOOD AND DRUG ADM-DEPT
(PDF) OF HEALTH
AND HUMAN SERVICES
V DASIKA

1 DEFENSE HEALTH AGCY
(PDF) C ESQUIVEL

1 US NAVAL SUBMARINE
(PDF) RSRCH LAB
J FEDERMAN

1 OHIO STATE UNIV
(PDF) R GODFREY

1 US ARMY (RET)
(PDF) M GRANTHAM

1 BRIGHAM YOUNG UNIV
(PDF) R HARRIS

1 ARMY MEDICAL
(PDF) RSRCH AND MATL CMND
S HOLLONBECK

1 US ARMY PUBLIC HEALTH
(PDF) CMND
C JOKEL

1 AUBURN UNIV
(PDF) S KRISHNAMURTI

1 US NAVAL SUBMARINE
(PDF) RSRCH LAB
B LAWSON

1 US ARMY PUBLIC HEALTH
(PDF) CMND
M MAKASHAY

1 UNIV OF MINNESOTA
(PDF) P NELSON

1 MICHIGAN STATE UNIV
(PDF) B RAKERD

1 UNIV OF MINNESOTA
(PDF) R SCHLAUCH

1 US ARMY PUBLIC HEALTH
(PDF) CMND
L SHERLOCK

1 US NAVAL SEA SYS CMND
(PDF) J ZIRIAX