

THE MEASUREMENT OF SPEECH INTELLIGIBILITY

Herman J.M. Steeneken TNO Human Factors, Soesterberg, the Netherlands

1. INTRODUCTION

The draft version of the new ISO 9921 standard on the "Assessment of Speech Communication" defines speech intelligibility as: "a measure of effectiveness of understanding speech". This contribution describes and compares several of these measures for determining the intelligibility of a given speech transmission system. It may include the acoustical environment at the speaker and the listener position.

In general two principally different assessment methods may be applied:

- (1) Subjective assessment, based on the use of speakers and listeners,
- (2) Objective assessment based on physical parameters of the transmission channel.

For a representative estimate of the speech intelligibility at least four speakers and four listeners are required, thus 16 speaker listener pairs. This results in a laborious effort. As the results depend on the individual subject responses, a reproduction of the test results is not obvious and requires at least inclusion of a number of reference conditions.

Objective measurements do not measure intelligibility but determine physical parameters to *predict* intelligibility according to a certain model. One should be aware that such a model might have restrictions that should be considered.

2. SUBJECTIVE INTELLIGIBILITY ASSESSMENT

First of all speech intelligibility should not be confused with speech quality. Speech intelligibility is related to the amount of speech items that is recognized correctly while speech quality is related to the quality of a reproduce speech signal with respect to the amount of audible distortions.

The subjective intelligibility measure might be based on phonemes, words (these may be meaningful words or nonsense words), and sentences. In principle there is a fixed relation between these three different types of speech material. However, although there are conditions where it is much more easily to detect a meaningful word (e.g., a digit or the alphabet) than a nonsense word that consists of a random combination of a consonant, vowel, and consonant (so-called CVC-word).

Various techniques for the presentation of the test material to the subjects and of the type of response are used. With the presentation of test words it is required to embed these words into a carrier phrase. This has the advantage that: the speaker can control his vocal effort, the listener is attended that a test word has to be recognized, and in case of temporal distortion (reverberation, echoes, and automatic gain control) a representative condition with respect to continuous speech is obtained.

The response method might be open or closed. An open response allows the listener to respond to what he/she thinks to have heard. A closed response offers the listener some alternative from which a selection has to be made. This is the case with the modified rhyme test (House et al, 1965) where the listener has to select an initial consonant or a vowel from a group of six alternatives, even if a phoneme outside the alternative list is recognized. This is especially the

case with the Diagnostic Rhyme Test (DRT) which is based on only two alternatives (Voiers, 1977). A closed response paradigm has the advantage that only a simple learning session of the listeners is required, while an open response, especially used with nonsense words, requires extensive training. However, the open response test has the advantage that better discrimination between various transmission conditions is obtained (increased effort pays off). A confusion matrix of the phonemes can be obtained from the scores in case nonsense words with an equally balanced distribution of the phonemes are used. In general a word list is compiled based on a representative selection of initial consonants (C_i), vowels (V), and final consonants (C_f). For the Dutch test 17 initial consonants, 15 vowels and 11 final consonants are used.

Word tests provide both word scores and individual phoneme scores, rhyme test are restricted to phoneme scores with a limited set of alternatives.

For tests with sentences various scoring methods are used. Frequently used is the Mean Opinion Score (MOS) where subjects (minimal 16) are asked to score their impression of the intelligibility on a five point scale. This scale amounts bad, poor, fair, good, and excellent. The MOS is often used for telecommunication assessment (telephone, GSM, etc). A very reproducible test, based on sentence intelligibility provides the Speech Reception Threshold (SRT). For the SRT a sentence that is masked by noise, is presented to a listener. The listener has to recall the sentence precisely. If the listener produces a correct answer, the next sentence is presented with an increased noise level of 2 dB. This continues till the response of the subject is incorrect, than the noise level will be decreased by 2 dB. After a number of presentations, a noise level is obtained for which 50 % of the sentences are responded correctly. The test amounts 13 sentences, the first three sentences guide the listener to the threshold, the noise levels used with the presentation of the last 10 sentences is used to obtain the SRT. The higher the intelligibility of the original speech the more noise can be added for the 50% correct responses (Plomp and Mimpen, 1979).

In Fig 1 the relation between consonant and vowel scores are given for 78 conditions. The conditions are based on three signal-to-noise ratios (0, 7.5, and 15 dB) and 26 band pass conditions. The scatter diagram clearly indicates that a high vowel score can be obtained with a low consonant score en visa versa. Therefore it is recommended to use test material based on both consonants and vowels.

Some tests are only based on consonants such as the Diagnostic Rhyme Test (DRT, Voiers, 1977) and the articulation loss of consonants (Al_{cons} , Peutz 1971). As these tests are normally used within a limited area of applications (DRT for speech coders, and Al_{cons} in room acoustics) there might be a unique relation with results obtained in similar conditions. However, for application in a wider range of distortions there might be a different relation for each field of application and no unique criteria can be applied.

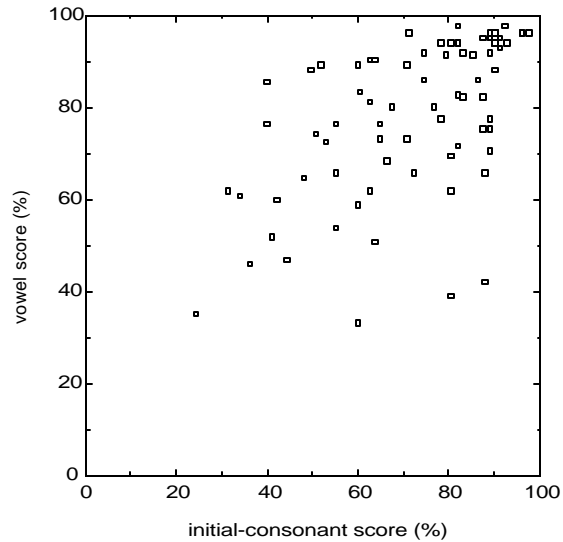


Fig. 1 Relation between consonant and vowel score for 78 conditions based on three signal-to-noise ratios and 26 bandwidth limitations.

In Fig. 2 a qualification and the relation between various subjective intelligibility scores and the subjective STI (Speech Transmission Index) is given. The qualification intervals are also related to a specific speech-to-noise ratio for a noise with a frequency spectrum equal to the speech spectrum. The graph shows that a ceiling effect is obtained for sentence scores. Meaningful PB words (Anderson and Kalb, 1987) also show a ceiling effect but the equally balanced CVC provides a wider range of qualifications.

Barnett (1995, 1999) proposed to use a reference scale, the Common Intelligibility Scale (CIS). The idea is to determine for each test method a unique relation with the CIS. The advantage is that criteria expressed in CIS scores are easy convertible to other measures. Barnett based the CIS on a mathematical relation with STI ($CIS = 1 + \log(STI)$), this resulted in a compressed relation with the five qualification intervals. Also the relation with the speech-to-noise ratio is not linear. Therefore, a suggestion was made to redefine the CIS and to use a linear relation with respect to the speech-to-noise ratio.

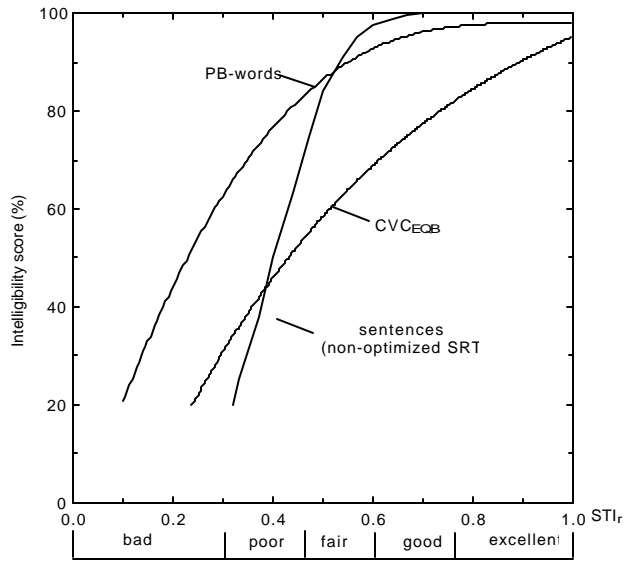


Fig. 2 Qualification and relation between various intelligibility scores and the STI (Houtgast and Steeneken, 1984)

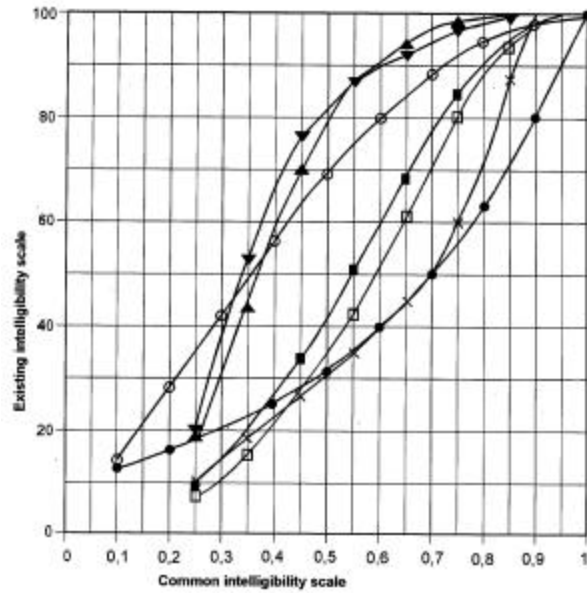


Fig. 3 Common Intelligibility Scale after Barnett (1995). Legend: \circ = PB words (256), \square = short sentences, \triangle = STI, \diamond = A_{cons} , \blacklozenge = PB words (1000), \bullet = 1000 syllables, \times = AI.

3. OBJECTIVE INTELLIGIBILITY ASSESSMENT

The assumption that the intelligibility of a speech signal is based on the sum of the contributions of individual frequency bands was proposed between 1925 and 1930 by Fletcher and modeled by French and Steinberg in 1947. They described that the specific information content of a speech signal is not equally distributed along the frequency range of a speech signal and developed a model of twenty contiguous frequency bands that provided an equal contribution to a defined index, the so-called Articulation Index (AI). This was the beginning of the development and the application of objective measures that *predict* intelligibility for various types of transmission channels.

Two frequently used objective measures are the STI (Speech Transmission Index, Steeneken and Houtgast, 1980, 1998), and the SII (Speech Intelligibility Index).

The STI is a measure that is based on the generation and analysis of an artificial test signal that replaces the speech signal. The result of the analysis is an index that ranges from 0 to 1. The STI accounts correctly for band-pass limiting, noise, reverberation, echoes, and non-linear distortion. STI is standardized by IEC standard 60268-16 (version 2, 1998).

The SII (former AI) is an objective measure that is obtained by calculation taking in account the physical properties of the transmission channel. The SII accounts for band-pass limiting and noise. The effect of temporal and non-linear distortions is not directly included. SII is standardized by ANSI standard S3.05 (1997).

In the STI concept the intelligibility of speech is related to the preservations of the spectral differences between consecutive speech elements, the phonemes. This can be described by the envelope function. An example of this envelope function for a 10s speech sample and for the octave band of 250 Hz is given in Fig. 4A.

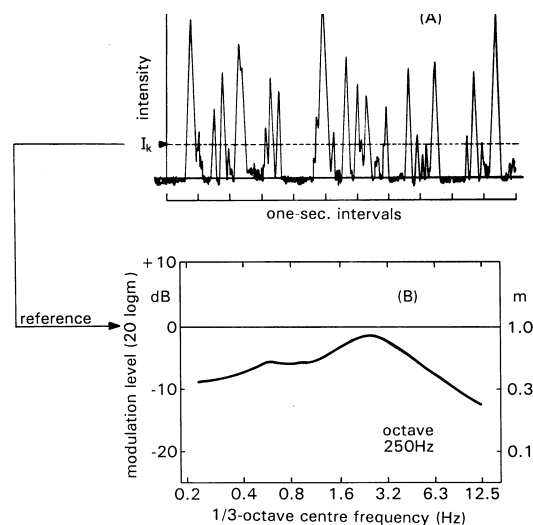


Fig. 4 Envelope function and envelope spectrum for the octave band 250 Hz of a 10 s speech sample.

The envelope function is determined by the specific sequence of phonemes of a specific utterance. A general description is offered by the frequency spectrum of the envelope function, the so-called envelope spectrum. This is given in Fig. 4B. The envelope spectrum is normalized with respect to the average intensity. The envelope spectrum has a maximum at the syllable repetition rate (3 Hz) and ranges between 0.2 Hz and 20 Hz. Fig. 5 shows the effect of temporal distortion and of noise on the envelope function and on the corresponding envelope spectrum.

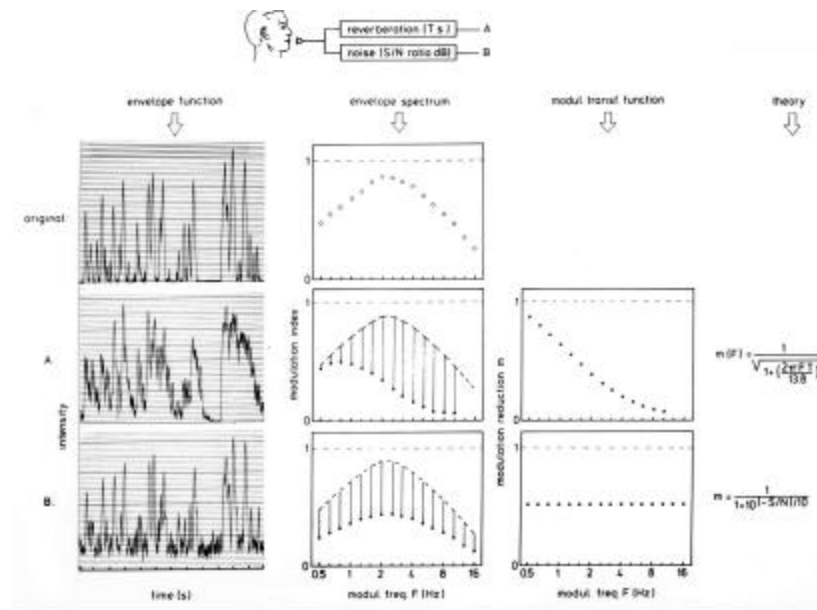


Fig. 5 Effect of reverberation (A) and of noise (B) on the envelope function and the envelope spectrum. These effects can be described by the MTF (Modulation Transfer Function).

Fig. 5A shows the effect of reverberation. The fast, highly peaked, envelopes are smeared due to the effect of reverberation. This is reflected in the envelope spectrum as a low-pass filter function. This filter response, the Modulation Transfer Function (MTF), is the difference between the original envelope spectrum and the envelope spectrum of the reverberated signal. For stationary noises the average intensity is increased, that results in a shift of the MTF. The effect of a single echo (not show) results in a rippled MTF related to the delay and the relative level of the echo. For the determination of the MTF in case of reverberation or echoes, the impulse response of the room can be used. However, if combinations of other types of distortions are effective than a specific, speech like test signal, is required. The STI is based on the determination of the effective signal-to-noise ratio in all 7 octave bands. This also includes the effect band-pass limitation, noise, temporal distortion and non- linear distortion. A simplified description of this test signal is given in Fig. 6.

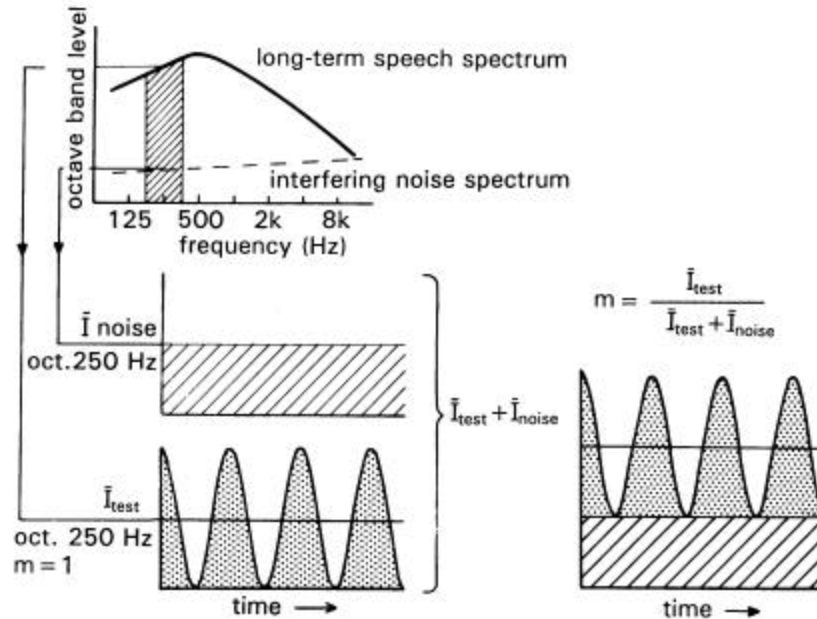


Fig. 6 Simplified description of the STI test signal.

The test signal consists of 7 separate octave band signals from which six bands consists of an artificial speech signal (required for the generation of possible non-linear distortion components) and one octave band that consists of a test signal. In the graph the test signal for the octave band with center frequency 250 Hz is shown. A modulated signal with a well-defined sinusoidal intensity envelope is used to determine the MTF. The frequency of this modulation is varied within the range of the fluctuations in speech. The graph describes the addition of an interfering noise, this is reflected in the modulation index "m". In this way a full matrix for seven octave bands and 14 modulation frequencies (0.63-12.5 Hz) is obtained. From this the effective SNR for each octave band is derived. This calculation also includes the effect of auditory masking and the reception threshold. A weighted summation of the seven octave contributions result in the STI value. The measurement of a full STI requires 10 minutes. Therefore some simplifications were applied for measurements under specific conditions. For example the RASTI (Room Acoustics STI, developed in 1979 with a simple microprocessor) was restricted to person-to-person communications but often used for assessment of PA-systems. Hence, band-pass limiting and non-linear distortions were not accounted for correctly. STITEL is a fast method for telecommunication systems, this method does not account for temporal distortion. The advantage is that a measurement can be performed in 15 seconds.

Some commercial available methods predict the STI value from data based on various objective measures (such as the impulse response, ray tracing results, or other predictive measures). This might in conflict with the basic concept of STI. The STI model determines the *effective* signal-to-noise ratio for all types of distortions in a generic relation to predict intelligibility. The standard IEC 60268-16 describes these various applications in detail.

4. CONCLUSION

Present signal processing technologies, integrated in personal computers, allow us to perform advanced measurements on public address systems and telecommunication channels used for alert and warning messages, professional use, and entertainment.

5. BIBLIOGRAPHY

- Anderson, B.W., and Kalb, J.T. 1987. "English verification of the STI method for estimating speech intelligibility of a communications channel," J. Acoust. Soc. Am. 81, 1982-1985.
- Barnett, P. W. and Knight, R.D. (1995). "The Common Intelligibility Scale", Proc. I.O.A. Vol 17, part 7.
- Barnett, P. W. (1999). "Overview of speech intelligibility" Proc. I.O.A Vol 21 Part 5.
- French, N.R., and Steinberg, J.C., 1947. "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. 19, 90-119.
- House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (1965). "Articulation testing methods: Consonantal differentiation with a closed response set", J. Acoust Soc. Am. 37, 158-166.
- Houtgast, T., and Steeneken, H.J.M., 1973. "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica 28, 66-73.
- Houtgast, T., and Steeneken, H.J.M., 1985. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. 77, 1069-1077.
- IEC 1998. "Sound system equipment- Part 16: Objective rating of speech intelligibility by speech transmission index". IEC standard 60268-16 second edition 1998.
- Kryter, K.D., 1962. "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. 34, 1689-1697.
- Pavlovic, C.V., 1987. "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. 82, 413-422.
- Peutz, V.M.A., (1971). "Articulation loss of consonants as a criterion for speech transmission in a room". J. Aud. Eng. Soc. 19, 12 (Dec 1971).
- Plomp, R., and Mimpen, A.M., (1979). "Improving the reliability of testing the speech reception threshold for sentences". Audiology 8, 43-52.
- Steeneken, H.J.M., and Houtgast, T., 1980. "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. 67, 318-326.
- Steeneken, H.J.M., 1992a. "Quality evaluation of speech processing systems," Chapter 5 in *Digital Speech Coding: Speech coding, Synthesis and Recognition*, edited by Nejat Ince, (Kluwer Norwell USA), 127-160.
- Steeneken, H.J.M., Verhave, J.A., Houtgast, T. 1993. Objective assessment of speech communication systems; introduction of a software based procedure. Proc. Eurospeech 93, 3rd Conference on Speech Communication and Technology, Berlin Germany, p. 203-206.
- Steeneken, H.J.M., and Houtgast, T., 1999. "Mutual dependency of the octave-band weights in predicting speech intelligibility," Speech Communication 28 (1999), 109-123.
- Steeneken, H.J.M., and Houtgast, T. 2002. "Phoneme-group specific octave-band weights in predicting speech intelligibility," Accepted for publication in Speech Communication (2002).
- Steeneken, H.J.M., and Houtgast, T., 2002. "Validation of the STI_r method with the revised model," accepted for publication in Speech Communication (2002).
- Voiers, W.D. (1977). "Diagnostic evaluation of speech intelligibility." In *Speech Intelligibility and Speaker Recognition*, Vol 2. Benchmark papers in Acoustics, edited by M.E. Hawley (Dowden, Hutchinson, and Ross, Stroudsburg), 374-384.