

# STANDARDISATION OF PERFORMANCE CRITERIA AND ASSESSMENTS METHODS FOR SPEECH COMMUNICATION

*Herman J.M. Steeneken*

Convenor of ISO/TC159/SC5/WG3, CEN/TC122/WG 8,  
Project leader IEC/TC100 60268-16

## ABSTRACT

The purpose of standardisation of the performance of speech communications is to assure a certain level of speech communication quality for various applications. The quality of speech communications is assessed in case of warning, danger, or information messages for work places, public areas, meeting rooms, and auditoria. In many applications direct communication between humans is considered while in other applications the use of electro-acoustic systems (e.g., PA systems) will be the most convenient means of informing and instructing people present.

An *international* standard on this subject, under the responsibility of ISO and CEN, has a status of draft international standard and has been ratified in a voting procedure (ISO-FDIS-9921). Various *international* and *national* standards are already accepted or recently initiated.

Also assessment methods are described in standards and technical reports. Some objective assessment methods are standardised at international and national level.

## 1. INTRODUCTION

Safe operation of warning and alert systems requires a certain level of speech communication quality. Therefore, general accepted criteria and performance measures have to be used. This is the goal of working groups and committees who work at international or national level.

International standardisation of assessment methods for speech communications is covered by ISO (International Standardisation Organisation), CEN (Commission European Normalisation), IEC (International Electrotechnical Commission), and ITU (International Telecommunication Union).

Under responsibility of ISO and CEN, international standardisation is performed by two special workgroups. These workgroups are referred to as ISO/TC 159/SC 5/WG 3 and CEN/TC 122/WG 8.

A recently revised International Standard (ISO 9921) specifies *criteria* for speech communication quality in case of verbal alert and danger signals, information messages, and speech communications in general. Methods to *predict* and to *measure* the performance in practical applications are addressed and examples are given. For this purpose both subjective and objective assessment methods may be applied.

In comparison with visual alert and warning signals, auditory signals are omni-directional and may therefore be preferred in many situations (smoke, out of line-of-sight). It is required however that, in case of verbal messages, a sufficient intelligibility is offered. If this cannot be achieved synthetic warning signals may be considered (see ISO 7731). Some of these standards are also disseminated at national level in the appropriate language. Some nations use their own standardization network such as ANSI (American National Standards Institute).

The communications to be involved may be directly between humans, through public address or intercom systems, or by using prerecorded messages. In order to obtain optimal performance for a specific application, three items are essential:

- 1: performance criteria,
- 2: development and predictive tools,
- 3: assessment methods.

These items have to be covered in a general purpose document in which not only “high technology” solutions are offered but also methods and tools which are simple to apply and generally available.

In an international workgroup as mentioned above various nations are represented. The task of the work group is to draft or to revise a standard or Technical Report (TR). After completion, the document is presented to all supporting countries for comments, possible suggestions for improvement and a vote for approval. If these comments are accepted, the working group will include these in the draft version. Then the voting procedure is started and at this phase, the nations can only respond by acceptance or rejection. If two third of the supporting nations have voted in favor for the standard, it is accepted and will be published.

A standard generally consists of normative and informative information, while a technical report consists of informative information only. A technical report is not subjected to a voting procedure.

## 2. SELECTION OF CRITERIA FOR SPEECH COMMUNICATION QUALITY

A requirement for the understanding of spoken messages is a correct recognition of each utterance. In technical terms it means that a sentence-intelligibility score of 100% is required for simple sentences. However, there are many situations for which a better performance is required. If we consider alert and warning situations it is sufficient to fully understand a short message under adverse conditions even if it requires some effort of the listener to understand the message correctly. In a meeting room, auditorium, or at work places where speech communication is a part of the task or where people are normally present for a longer period of time, a more relaxed speaking and listening condition is required. For the speaker this may be reflected by the vocal effort that is required to be understood (quantified as relaxed, normal, raised, loud, and very loud). For the listener the listening effort may be primarily related to the speech quality offered at the listening position. In Table I normative criteria for various types of applications are given. Five qualification intervals (excellent, good, fair, poor, bad) are used and related in an informative table to various subjective and objective measures (see section 4, Table II).

Table I. Normative criteria for speech intelligibility and vocal effort. The criteria are according FDIS ISO 9921.

Application	Minimum intelligibility rating	Maximum vocal effort
Alert and warning situations (correct understanding of simple sentences)	Poor	Loud
Alert and warning situations (correct understanding of critical words)	Fair	Loud
Person-to-person communications (critical)	Fair	Loud
Person-to-person communications (prolonged normal communication)	Good	Normal
Public address in public areas	Fair	Normal
Personal communication systems	Fair	Normal

An international standard focused on the technical design of systems applied for warning signals is released by IEC (IEC 60849). This standard introduces a common intelligibility scale (CIS) that was proposed by Barnett and Knight (1995). The normative criterion for the intelligibility of a warning signal according to this standard is  $CIS = 0.7$ . This is equal to a  $STI = 0.5$ . The same criterion is used for the national USA fire alarm code (NFPA 72, National Fire Protection Association).

### **3. METHODS FOR PREDICTION OF THE PERFORMANCE OF SPEECH COMMUNICATION SYSTEMS**

The prediction of the performance with respect to the intelligibility of speech communication channels is generally based on the effective signal-to-noise ratio at the listener position. Various methods are developed to calculate this effective signal-to-noise ratio derived from the vocal effort and acoustic aspects of the speaker, the transfer of the speech signal by electro-acoustic systems, and the acoustical aspects at the speaker and listener position.

The various methods differ in complexity. Simple methods just compare the speech spectrum and the noise spectrum at the listener position. Advanced methods also take into account the effect of temporal distortion, non-linear distortion and hearing aspects.

The SIL-method (Speech Interference Level, Beranek, 1947) is based on the A-weighted speech level and the mean noise level within in four octave bands. A predictive measure of the intelligibility is obtained by subtracting the mean noise level from the estimated speech level (noise level represents mean value of the levels of the octave-bands 500-4000Hz,  $SIL = L_{SA} - L_{LN}$ ).

The STI (Speech Transmission Index), and SII (Speech Intelligibility Index, ANSI 3.05 1998) take into account the speech and noise spectrum and additionally the bandwidth, speech production and hearing aspects (Fletcher and Steinberg, 1929; Steeneken and Houtgast, 1980, 1992, 1999).

The SII is designed to predict various subjective intelligibility measures such as: nonsense syllables, phonetically balanced words, monosyllables, DRT words (Diagnostic Rhyme Test), short passages of easy reading material and monosyllables of speech in presence of noise (House et al., 1965). A slightly different calculation scheme is used in order to predict the scores related to the various subjective measures. SII takes also into account hearing aspects such as masking and hearing disorders.

The STI is designed for prediction of nonsense syllables, and gives a qualification of the predicted speech intelligibility. Additionally to the other methods the STI accounts for temporal distortions by making use of the so-called Modulation Transfer Function (MTF), male and female speech signals, and for non-linear distortions.

None of the predictive intelligibility measures take into account the ability of a person to focus on speech sounds from a specific direction (directional hearing). Directional hearing might improve, under certain conditions, the intelligibility. This may be related to an improvement of the effective signal-to-noise ratio by approximately 3 dB.

The STI and SII are well described in various standards [IEC 60268-16 2<sup>nd</sup> edition, ANSI 3.5].

#### 4. ASSESSMENT METHODS

Quantification of the speech quality requires specification of qualification intervals that cover the potential use, selection of measuring methods should comply with these qualification intervals. Also the selected measures must be applicable by the potential users of the standard. Therefore, the selected measures should cover the following specifications:

- 1 described in a standard or at least published (with peer review) and generally accepted,
- 2 reproducible,
- 3 producing results which comply with the required qualifications, such that scores from one method can be converted to another without ceiling effects (saturation),
- 4 including subjective and objective measuring methods,
- 5 applicable to the (acoustical) conditions covered by the standard.

In Table II three subjective methods are compared (Miller and Nicely, 1955; Plomp and Mimpen, 1979; Steeneken, 1992). Some of these are described in standard ISO 4870. It is of great importance that the test material for a subjective test used in reverberating environments makes use of test words embedded in a carrier phrase in order to make sure that a representative reverberation is present during the presentation of the test word.

Also three objective methods are given in Table II (Lazarus, 1990; Steeneken and Houtgast, 1980, 1999). These methods are similar to those discussed for prediction purposes.

The intelligibility rating scale is similar to the scaling used with the Mean Opinion Score (MOS) as standardized by ITU (P800) and also proposed by Houtgast and Steeneken (1984).

Table II. Intelligibility rating and relations among six intelligibility measures. The sentence score refers to simple sentences, CVC<sub>EQB</sub>-nonsense words with an equally balanced phoneme distribution, and the PB-word score (related to the phonetically balanced Harvard list).

Intelligibility rating	Sentence score %	Meaningful PB-word <sup>1</sup> score %	CVC <sub>EQB</sub> -non-sense word Score %	STI	SIL <sup>2</sup> dB	SII <sup>3</sup>
Excellent	100	> 98	> 81	> 0,75	21	
Good	100	93 – 98	70 – 81	0,60 - 0,75	15 – 21	> 0,75
Fair	100	80 – 93	53 – 70	0,45 - 0,60	10 – 15	
Poor	70 – 100	60 – 80	31 – 53	0,30 - 0,45	3 – 10	< 0,45
Bad	< 70	< 60	< 31	< 0,30	< 3	

We compared, for a number of noise conditions, the scores of the STI, SII and SIL. In Fig. 1 the relation between STI and SII is given for 40 noise conditions with a random spectrum distributed over a wide range of octave levels. The rank-order correlation between both measures is  $r = 0.93$ . The figure shows a slight saturation of the SII at higher values.

<sup>1</sup> According to Anderson and Kalb (1987)[1].

<sup>2</sup> Qualification of the SIL-method is converted to a five-point scale rather than the original seven-point scale.

<sup>3</sup> The SII procedure does not provide qualification intervals. The ANSI standard does provide two benchmarks: good > 0.75, poor < 0.45.

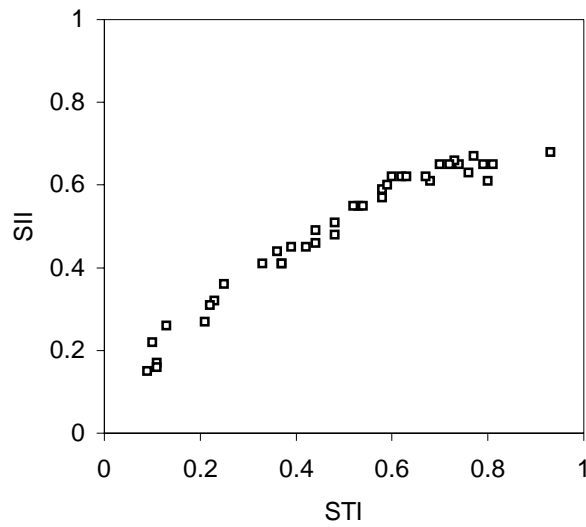


Fig.1. Relation between STI and SII. The correlation coefficient amounts  $r = 0.93$ .

The relation between SIL and STI / SII is given in Fig. 2. Here the rank order correlation of SIL and STI amounts  $r = 0.97$  and between SIL and SII  $r = 0.95$ . It should be noted that SIL is only applicable for noise conditions and undistorted speech signals.

The field of application of the objective methods depends on the ability to cope with the distortions, which are relevant for a specific application. Possible distortions are: background noise, reverberation, echoes, increased vocal effort of the speaker. In case that electronic means are used also band-pass limiting and non-linear distortions have to be included.

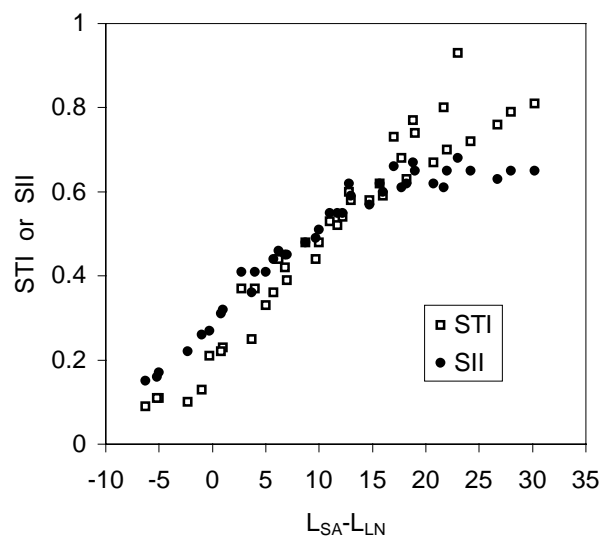


Fig. 2. Relation between SIL, STI and SII. The correlation coefficients between SIL and STI is  $r=0.97$  and between SIL and SII  $r=0.95$ .

## 5. CONCLUSIONS

This overview gives criteria for adequate quality of speech communications for several applications. Alert and warnings are primarily considered but also conditions for more relaxed communication conditions (Table I + II).

The standards are adapted to this view and give criteria for acceptable speech communication quality in various conditions from warning and alert conditions to the more relaxed meeting room. Additional to the criteria, methods to assess the performance of existing situations or to predict the performance for applications under development are given. After the “11 September” incident the standardization of verbal alert and warning signals is accelerated, therefore the availability of criteria and assessment methods is required. Both international standards and national standards are available these are summarized below.

Some international standards are:

- International Standardization Organization ISO 9921 “Ergonomics, Assessment of speech communication”, revised version developed by ISO-TC159/SC5/WG3, passed voting procedure (96% of voting members in favor), to be disseminated 2003. Will also be disseminated as a CEN standard.
- International Standardization Organization Technical report, ISO/TR4870 “Acoustics – The construction and calibration of speech intelligibility (1991).
- International Electrotechnical Commission IEC 60268-16 (1998-03) “Sound system equipment – Part 16: Objective rating of speech intelligibility by Speech Transmission Index”, ratification procedure of revised version in progress 2002.
- International Electrotechnical Commission IEC 60849 (1998-02) “Sound systems for emergency purposes”, ratification procedure of revised version in progress 2002.
- International Telecommunication Union, ITU P800 (08/96) “Methods for subjective determination of transmission quality”.

Some interesting national standards are:

- American National Standards Institute publication ANSI S3.5 (1998) on Speech Intelligibility Index
- American National Standards Institute publication ANSI S3.2 (1989) on Modified Rhyme Test.
- National Fire Protection Association (USA) with fire alarm code NFPA72.
- BSI (UK), BS 5839-8 Fire detection and alarm systems for buildings. Code of practice for the design, installation and servicing of voice alarm systems.
- BS 7827 (1996) Code of practice for designing, specifying, maintaining, and operating emergency sound systems at sports venues.
- NEN (NL), NEN 3438 “Ergonomie - Geluidhinder op de arbeidsplaats - Streefwaarden voor geluidniveau en nagalmtijden met betrekking tot verstoring van communicatie en concentratie.”

## 6. REFERENCES

- 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.
- Beranek, L.L. (1947) “Airplane quieting II specification of acceptable noise levels”. Trans. Amer. Soc. Mech. Engrs.

69:97-100.

- Fletcher, H., and Steinberg, J.C. (1929). "Articulation testing methods", Bell Sys Tech. J. **8**, 806.
- 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. (1984). "A multi-lingual evaluation of the RASTI-method for estimating speech intelligibility in auditoria," *Acustica* **54**, 185-199.
- Lazarus, H., (1990). "New methods for describing and assessing direct speech communication under disturbing conditions". *Environment International*, **16**, pp. 373-392.
- Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338-352.
- 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. (1992). "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., and Houtgast, T. (1999). "Mutual dependence of the octave-band weights in predicting speech intelligibility". *Speech Communication* 1999, vol. **28**, 109-123.