Proceedings of the Institute of Acoustics

Keynote Lecture at the 40th Anniversary Conference, Birmingham 16 October 2014

FORTY YEARS OF SPEECH INTELLIGIBILITY ASSESSMENT (AND SOME HISTORY)

Herman J.M. Steeneken Former: TNO Human Factors, Soesterberg, The Netherlands
Part time: Embedded Acoustics, Delft, The Netherlands
see also: Acousteen: www.steeneken.nl

1 INTRODUCTION

It took much more than 40 years to progress from handwritten and manually corrected subjective intelligibility tests to automatically controlled assessment methods. The availability of processors and key-boards provided a major improvement and triggered statistical analysis of the results.
Objective assessment methods, either based on a calculation scheme or on-site physical analysis of a communication channel, became generally available and are widely used. This overview highlights some of the mile-stones to the present state-of-the-art.

2 SUBJECTIVE INTELLIGIBILITY ASSESSMENT

Speech is a major means of communication. This was recognized already by Demosthenes (384-322 BC) some 2400 years ago. He wanted to become a great debater and therefore he had to improve his articulation. He started to speak with pebbles in his mouth; the effect was such that finally he became a famous (intelligible) debater in parliament.
Let us skip two millennia to the 1920's – 1960's. New means of communication became available, the telephone (1876) and radio communications for civil and military applications (1919). The development and evaluation of these speech transmission systems required knowledge of how to transfer speech signals for optimal intelligibility. Limited bandwidth, noise and poor electro acoustic transducers (carbon microphones) degraded the speech quality. Studies to relate these limitations to speech quality were initiated by various researchers, for example by Fletcher and Steinberg[9], Egan[7] and many others. Speech intelligibility in an auditorium was also assessed; here reverberation, echoes and noise were the major disturbances. Assessment methods -making use of speakers and listeners- were developed and described by Beranek[4,5], Bolt and MacDonald[6] and Fairbanks[8].

Miller and Nicely[20] focused on the speech material and made a major contribution to the characterization of the different phonemes. Based on their results they proposed five articulatory features: voicing, nasality, affrication, duration and place of articulation (i.e. labial, dental, glottal and nasal).

At the start of the IOA in 1974, three major groups of subjective tests were available:
1 Tests at phoneme and word level:
- monosyllabic words, for example CVC-words (Consonant-Vowel-Consonant), Fletcher and Steinberg[9], Egan[7], and with equally balanced phoneme sets (which made it possible to study confusions), Steeneken[26],
- rhyme tests (selection of a word from a small group of alternatives that generally differ just by the first phoneme), Fairbanks[8], House et al.[12], Voiers[31] and, for Dutch, Steeneken[26],
- closed response set. For a simple test with untrained listeners a closed response set may be used, such as digits or the alphabet,
- initial consonant tests such as the ALcons (articulation loss consonants, Peutz[21]).

2 Tests at sentence level:
- asking a subject to estimate the percentage of words correctly heard,
- SRT (speech reception threshold) a procedure where a presented sentence at a constant level is masked by noise at an adjustable level. In a sequence presentation of different

sentences the noise level corresponding for 50% correct responses is determined (Plomp and Mimpen[22]). The SRT is often used in clinics for optimal adjustment and control of hearing aid performance.

3 Quality rating:

- MOS (Mean Opinion Score) a method to evaluate the listener's impression of the quality of a transmission channel (Goodman and Nash[11]).

In the 1970's the word tests given above were generally performed with a group of speakers and a (different) group of listeners. The listeners wrote down the words they heard. This was evaluated afterwards (by hand!) and delivered a score of correctly reproduced words (the word-score). At the TNO we used word lists of 50 words (embedded in a carrier phrase) to be read by a speaker, ordered within a 5x10 frame. Underneath each word of the reading list a gap was punched, corresponding with a response box on the listener response sheet. The listener had to write his/her responses in the same order. By overlaying the corresponding word list on top of the response sheet all the spoken words together with the responses could easily be checked. In the mid 70's flexible keyboards became available and even confusions between stimuli and responses could be digitally obtained. This was a first step to obtaining diagnostic information from these subjective tests. A confusion matrix between, for example initial consonants, resulted in a display of the relation between these consonants for a certain transmission condition. The Diagnostic Rhyme Test claimed this possibility as well, but within a limited response frame due to the limited response set (Voiers[31]).

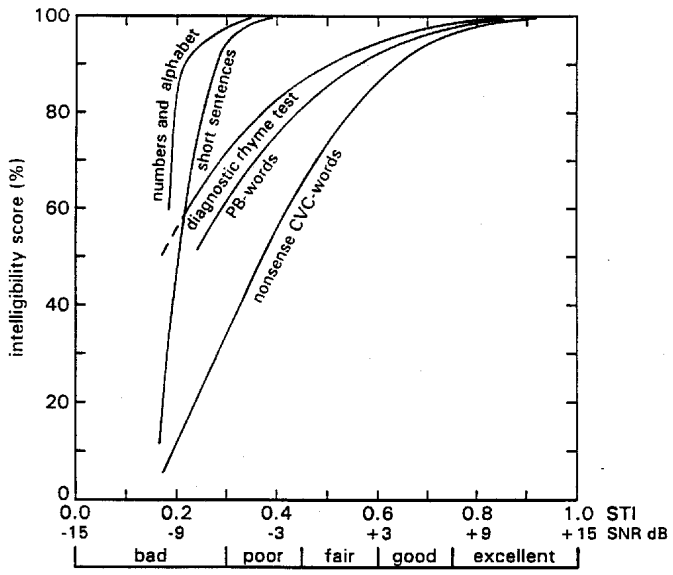

Fig.1. Relation between some intelligibility measures as a function of the signal-to-noise ratio (based on noise with a long-term speech spectrum).

The major differences between the various test methods are related to the size of the vocabulary and the complexity of the speech material used for the test. Fig.1. gives -for five intelligibility measures- the score as a function of the signal-to-noise ratio of speech masked by noise (Steeneken[27]). This gives an impression of the effective range of each test. The given relation between intelligibility scores and the signal-to-noise ratio is valid only for noise with a frequency spectrum similar to the long-term speech spectrum. This makes the signal-to-noise ratio the same for each frequency band, as is the case with voice-babble. A signal-to-noise ratio of 0 dB then means that speech and noise have an equal spectral density.

As can be seen from the figure, the CVC-nonsense words discriminate over a wide range, while meaningful test words[1] have a slightly smaller range (Anderson and Kalb[2]). The digits and the alphabet give saturation at a signal-to-noise ratio of -5 dB. This is due to: (a) the limited number of test words and (b) the fact that recognition of these words is controlled mainly by the vowels rather than by the consonants. Vowels have an average level of approximately 5 dB above the average level of consonants, and are therefore more resistant to noise. On the other hand non-linear distortions, such as peak-clipping, will have a greater impact on vowels than on consonants. Therefore the use of the digits and the alphabet, for which recognition is mainly based on vowels, may give misleading results.

[1] Meaningful test words are normally phonetically balanced (PB); hence the frequency distribution of the phonemes is representative for the language used.

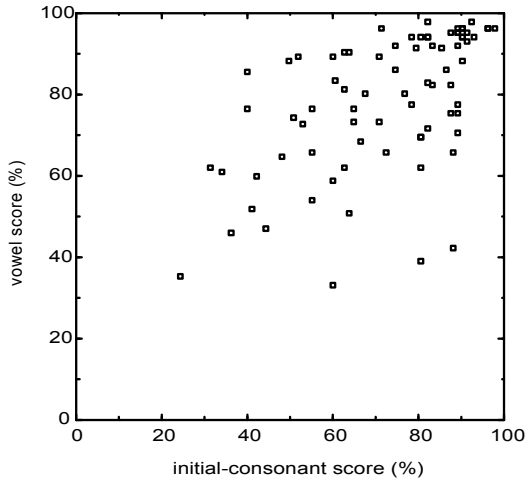

Fig. 2. Initial consonant score versus vowel score obtained from CVC-words for 78 transmission conditions.

This is indicated in Fig. 2. In this figure the initial-consonant score is given versus the vowel score as obtained from CVC-word tests for 78 different transmission conditions. The graph shows that a high vowel score and a low consonant score can be obtained for one type of channel (e.g. band-pass limiting) while conversely a low vowel score combined with a high consonant score can be obtained for another type of channel (e.g. peak clipping). This indicates that the exclusive use of either consonants (ALcons) or vowels in a subjective test may lead to an incorrect evaluation of the transmission quality. A combination of consonants and vowels, as with CV or CVC words, is required.

The comparison above is based on developments of individual studies. However, a comparative experiment, to evaluate the objective intelligibility assessment method RASTI (Room Acoustic Transmission Index), also delivered a comparison between subjective test methods (Houtgast and Steeneken[15]). The results of this multilingual evaluation of subjective intelligibility assessment methods applied in an auditorium triggered the standardization process. Eleven laboratories participated in a round-robin test in which 16 transmission conditions (combinations of: noise, PA-system, microphone position and AMBIO-system) were compared. The choice of the subjective assessment method (word-type, rhyme test, carrier phrase, presentation) was free. Each laboratory used their own available method as given in Table I.

Table I. Description of the various test methods[15].

Country	Test-material	Carrier phrase	Speakers-listeners	Presenta-tion	Score
1 England	Monosyllabic words	No	1m/1f - 20	Loudspeaker	Word-score
2 Finland	Two syllables	Yes	2m/2f - 8	Loudspeaker	Word-score
3 France	One/two syllables	Yes	1m/1f - 8	Headphone	Word-score
4 Germany	6 alternative rhyme	Preceding	1m - 12	Headphone	CV-score
5 Hungary	One/two syllables	No	1m/1f - 10	Loudspeaker	Syllable-score
6 Italy	5 alternative rhyme	Preceding	1m - 18	Headphone	Mean-score
7 Netherlands	Nonsense CVC	Yes	1m - 5	Headphone	Cons-score
8 New Zealand	Meaningful words	Yes	1m - 30	Loudspeaker	Word-score
9 Poland	One/two syllables	No	1m - 21	Loudspeaker	Word-score
10.1 Sweden	Meaningful words	Preceding	1m - 12	Headphone	Word-score
10.2 Sweden	Nonsense CVC	Yes	1f - 12	Headphone	Word-score
11 Yugoslavia	Nonsense CVCV	No	2m - 30	Loudspeaker	Cons-score

Regarding the different types of test-methods a mutual relation similar to Fig. 1 may be expected. Therefore the relation between the 16 scores of the 11 participants was based on a Spearman rank-order correlation. The correlation coefficients between all participants and the mean correlation coefficient for each test are given in Table II. Scores based on just consonants or a limited response set (rhyme words) show a low correlation with the tests based on (CVC) words.

The experimental results of the international round-robin assessment experiment also delivered a view on subjective qualification boundaries[15]. For an acoustic consultant these boundaries are essential rather than the different scores obtained with different assessment material. Advice given

on whether the delivered system is useful or not must be indisputable. For this reason Barnett and Knight[3] proposed an intelligibility scale that relates various intelligibility measures with each other. This scale was also introduced in some editions of the IEC standard 60268-16[2].

Table II. Spearman rank-order correlation coefficients between the various test-scores[15]. The bottom row presents the mean correlation for that particular test with all other tests

	GBR	FIN	FRA	GER	HUN	ITA	NLD	NZL	POL	SWE	YUG
England	-										
Finland	0.86	-									
France	0.89	0.94	-								
Germany	0.79	0.93	0.86	-							
Hungary	0.93	0.94	0.87	0.85	-						
Italy	0.51	0.81	0.81	0.82	0.68	-					
Netherlands	0.40	0.73	0.75	0.79	0.54	0.92	-				
New Zealand	0.92	0.95	0.93	0.85	0.91	0.67	0.60	-			
Poland	0.92	0.91	0.82	0.76	0.95	0.59	0.45	0.91	-		
Sweden	0.87	0.98	0.95	0.90	0.92	0.82	0.73	0.93	0.90	-	
Yugoslavia	0.86	0.85	0.82	0.71	0.91	0.59	0.48	0.89	0.90	0.81	-
Mean corr.	0.79	0.89	0.86	0.83	0.85	0.72	0.64	0.85	0.81	0.88	0.78
RASTI	0.83	0.97	0.96	0.92	0.89	0.88	0.82	0.91	0.83	0.96	0.82

3 OBJECTIVE INTELLIGIBILITY ASSESSMENT

The advantage of subjective intelligibility tests is that it may be representative if the various test parameters are chosen in accordance with the application. However no, or only little, diagnostic information is obtained concerning the effect of various transfer conditions on the final score. Moreover subjective tests are laborious and require a large number of speakers and listeners. Therefore various researchers investigated the relation between the physical properties of a transmission channel and the transmission quality. Fletcher and Steinberg[9] and French and Steinberg[10] provided information on the importance of the frequency transfer of a speech signal and the dynamic range in relation to the signal-to-noise ratio. Various hearing aspects such as masking effects were also considered. All of this information resulted in a prediction model for the intelligibility based on the physical transfer parameters -the Articulation Index (AI)- as described and compiled in a calculation sheet by Kryter[16], and the later development of the SII (Speech Intelligibility Index, ANSI).

The first description of the use of a computational method for the prediction of the intelligibility of speech and its realization in a measuring device was given by Licklider et al.[18]. They described a system which could measure the spectral correspondence between speech signals at the input and at the output of the transmission channel under test, the so-called Pattern Correspondence Index (PCI). This PCI shows a remarkable similarity with the AI (Articulation Index), although the approach is quite different. Five years later Kryter and Ball[17] described a system called the Speech Communication Index Meter (SCIM), which was based on the AI.

Houtgast and Steeneken[13], developed a system based on the use of an artificial test signal which was transmitted over the channel-to-be-tested and which was analysed at the output. The test signal was an amplitude-modulated noise signal with a square-wave amplitude modulation. Hence the signal level alternated between two values. The difference between these two levels was 20 dB and the switching rate was 3 Hz. The noise carrier had a frequency spectrum corresponding to the long-term speech spectrum. It was the first approach in which speech-related phenomena, concerning dynamic variations and temporal variations, were included in an artificial test signal and after analysis used to calculate a transmission index.

[2] The many discussions with Peter Barnett, and others, during meetings of the “reproduced sound” group in Windermere were very much appreciated.

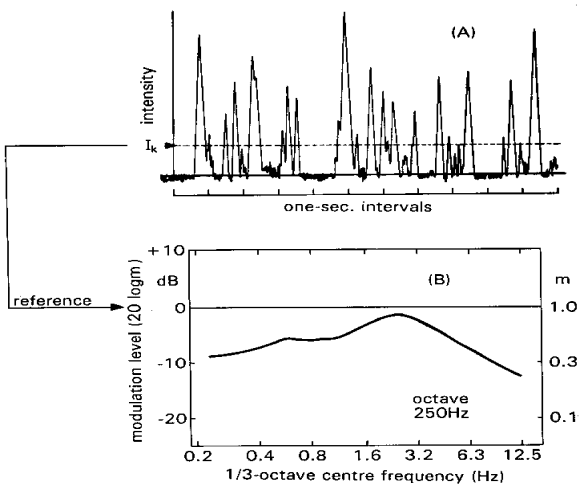

Fig. 3. Envelope function (panel A) of a 10s speech signal filtered for the octave band 250 Hz. The corresponding envelope spectrum (panel B) is normalized with respect to the mean signal intensity.

The next step was to go back to the speech signal itself as the intelligibility is related to the preservation of the spectral differences between successive speech sounds, the phonemes. This can be described by the envelope function[14,24] for a number of frequency bands. An example of this envelope function for a 10s speech sample and for the octave band of 250 Hz is given in Fig. 3. 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 panel B. 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. 4 shows the effect of temporal distortion and noise on the envelope function and on the corresponding envelope spectrum. 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 (panel A). 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 just the average intensity is increased. That results in a shift of the MTF (panel B). The effect of a single echo (not shown) results in a rippled MTF related to the delay and the relative level of the echo.

The MTF's obtained for seven octave bands (125Hz-8kHz) are the basis of a prediction model of the corresponding speech intelligibility. Further development resulted in using an optimized artificial test signal rather than a speech signal. The benefit of this approach was a better reproducibility and a lower sensitivity for speech-like disturbances. Also the effect of non linear distortion could be better accounted for. Calculation parameters such as the contribution of each frequency band, the range of the MTF, hearing related masking and thresholds were obtained by correlation with subjective CVC-word based scores obtained for many reference channels[25,28].

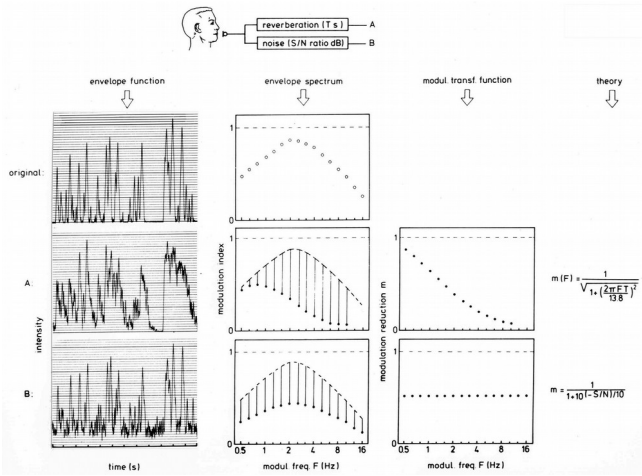

Fig. 4. Reductions -in an octave-band specific envelope function- for two transmission conditions: (A) with reverberation, (B) with additive noise. The modulation transfer is obtained by comparison with the original envelope spectrum.

The original measuring procedure required a measuring time of 10 minutes, therefore several simplified screening methods were developed each with a restricted set of applications. The first one was RASTI[15] (1979) a portable device developed for *person-to-person* communications. It was made commercially available by Brüel and Kjaer (Denmark). STIPA[29] (2002) provides a wider range of applications including public address systems but also has some restrictions. The MTF can also be obtained from an impulse response measurement. However the effect of background noise and of non-linear distortion is not directly available.

Tools to predict the MTF in an auditorium were developed and these are mostly based on ray-tracing (Ahnert[1], van Rietschote[23], Schroeder[24]).

The effect of non-native speech was studied by van Wijngaarden[32]. In general, a decrease of the effective signal-to-noise ratio of -4dB was found (for speaker or listener).
A goal for the future is to arrive at a measuring scheme (as we started with in 1970) by comparison of the input and output speech signal suitable for all types of distortion including the contribution of directional hearing.

4 STANDARDISATION

The availability of flexible test methods led to standardization of minimal requirements for alert and warning messages and other verbal communication applications. International standardisation is coordinated by ISO (International Standardisation Organisation), CEN (Commission European Normalisation), IEC (International Electrotechnical Commission), and ITU (International Telecommunication Union). A recently revised International Standard (ISO 9921) specifies criteria for speech communication quality in the cases of verbal alert and danger signals, information messages, and speech communications in general. Such a standard offers application related criteria. In Table III normative criteria are given.

Table III. Recommended minimum criteria for the intelligibility for six categories of applications.

Application	Intelligibility	LSA -L LN dB (SIL)	STI	effective SNR dB	Maximum vocal effort
Alert and warning	poor/fair	10	0.45	-1.5	Loud
Person-to-person (critical)	poor/fair	10	0.45	-1.5	Loud
Person-to-person (relaxed	good	15	>0.60	3	Relaxed
Public address in public areas	fair/good	11	0.50	0	Normal
Personal communication	fair/good	11	0.50	0	Normal

Table IV. Qualification and relation between various intelligibility measures.

Qualification	Sentence score %	CVC_{EB} nonsense word score %	Meaningful PB-word score %	STI	$L_{SA} - L_{LN}$ dB (SIL)	SII
Excellent	100	>81	> 98	>0.75	21	
Good	100	70-81	93-98	0.60-0.75	15 - 21	>0.75
Fair	100	53-70	80-93	0.45-0.60	10 - 15	
Poor	70-100	31-53	60-80	0.30-0.45	3 - 10	<0.45
Bad	<70	<31	<60	< 0.30	< 3	

Five qualification intervals (excellent, good, fair, poor, and bad) are used and related to various subjective and objective measures (Table IV). The methods used in a standard to predict and to measure the performance in an application differ in complexity. Simple objective methods just compare the speech spectrum and the noise spectrum at the listener position (SIL). Advanced methods also take into account the effect of temporal distortion, non-linear distortion and hearing aspects (STI, SII).
The SIL-method (Speech Interference Level, Beranek[5]) is based on the A-weighted speech level and the mean noise level within four octave bands. This can be achieved with relative simple measuring equipment. The STI and SII require more complex measuring procedures, these are well described in various standards (IEC 60268-16, 4th edition[19], ANSI 3.5). Subjective measures may also be used. In general, all measures should be described in a standard, be reproducible, produce results which comply with the required qualifications and be applicable to the (acoustical) conditions covered by the standard.

5 THE PAST AND THE FUTURE

In the last four decades major developments were made in both subjective and objective assessment of speech communication. Data collection and processing improved due to the availability of processing tools.
Also the range of applications extended from telephone and (military) radio communications to a variety of commonly used facilities such as: alert and warning systems, coders, public address, classrooms, meeting rooms and hearing-aid adjustment. Some of the required minimum intelligibility scores were standardized by ISO, CEN, IEC and ANSI.
Objective assessment methods became available, based either on a calculation scheme or on-site physical analysis of a communication system. Hand-held analysis equipment simplified common use and is widely available on the market.
The ultimate goal for the future is to use a representative, binaural speech-token and to determine the corresponding intelligibility score[27]. Such a procedure allows a dynamic adaptive system to deliver optimal performance. We are close!!!

6 REFERENCES

1. Ahnert, W, Bourrilet, C., and Feistel, S., (2001) “Phase presentation in the Acoustic Design program EASE”. Proc. 110Th AES convention Amsterdam.
2. 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.
3. Barnett, P. W. and Knight, R.D. (1995). “The Common Intelligibility Scale”, Proc. I.O.A. Vol 17, part 7.
4. Beranek, L.L., (1947). "The design of speech communication systems,” Proc. of the Institute of Radio Engineers 35, 880-890.
5. Beranek, L.L. (1954). *Acoustics* (McGraw-Hill, New York).
6. Bolt, R.H., and MacDonald, A.D., (1949). "Theory of speech masking by reverberation," J. Acoust. Soc. Am. 21, 577-580.
7. Egan, J.P. (1944). "Articulation testing methods," OSRD report No. 3802.
8. Fairbanks, G., (1958). "Test of phonetic differentiation: The Rhyme Test," J. Acoust. Soc. Am. 30, 596-600.
9. Fletcher, H., and Steinberg, J.C. (1929). Bell Sys Tech. J. 8, 806.
10. French, N.R., and Steinberg, J.C., (1947). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. 19, 90-119.
11. Goodman, D.J., and Nash, R.D., (1984). "Subjective quality of the same speech transmission conditions in seven different countries," IEEE Trans Comm. 30, 642-654.

12. 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.
13. Houtgast, T., and Steeneken, H.J.M., (1971). "Evaluation of speech transmission channels by using artificial signals", Acustica 25, 355-367.
14. 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.
15. 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.
16. Kryter, K.D., (1962). "Methods for the calculation and use of the articulation index", J. Acoust. Soc. Am. 34, 1689-1697.
17. Kryter, K.D. and Ball J.H.,(1964).”A meter for measuring the performance of speech communication systems”, Techn. Doc. Report ESD-TDR-64-674.
18. Licklider, J.C.R., Bisberg, A., and Schwartzlander, H., (1959). "An electronic device to measure the intelligibility of speech," Proc. Natl. Electronic Conf. 15, 329-334.
19. Mapp, P., (2002) “Further thoughts on Speech Transmission Index (STI)”. IOA Reproduced Sound 18. Proc. IOA Vol. 24 Pt 8.
20. Miller, G.A., and Nicely, P.E., (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352.
21. 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).
22. Plomp, R., and Mimpen, A.M., (1979). “Improving the reliability of testing the speech reception threshold for sentences". Audiology 8, 43-52.
23. Rietschote, H.F. van, Houtgast, T., Steeneken, H.J.M., (1981). “Predicting speech intelligibility in rooms from the modulation transfer function. IV. A ray-tracing computer model”. Acustica 49(1981) 245-252. 1981 829.
24. Schroeder, M.R., (1981) “Modulation Transfer Functions. Definition and measurement”. Acustica 49,pp. 179-182.
25. Steeneken, H.J.M., and Houtgast, T., 1980. "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. 67, 318-326.
26. Steeneken, H.J.M., 1982 “Ontwikkeling en toetsing van een Nederlandstalige diagnostische rijmtest voor het testen van spraakkommunikatiekanalen”. Report IZF 1982-29 TNO Institute for perception, Soesterberg NL.
27. Steeneken, H.J.M., (1992) “On measuring and predicting speech intelligibility”. Soesterberg: TNO Institute for Perception, 162 p. ISBN: 90-6743-209-1.
28. 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.
29. Steeneken, H.J.M., Verhave, J.A., McManus, S., and Jacob, K.D., (2001) “Development of an Accurate, Handheld, Simple-to-use Meter for the Prediction of Speech Intelligibility”, Proceedings IoA 2001, Reproduced sound (17). Stratford-upon-Avon, UK.
30. Van Gils, Bastiaan J., and Sander J. van Wijngaarden., (2005) “Objective Measurement of the Speech Transmission Quality of Vocoders by Means of the Speech Transmission Index”. TNO Human Factors,
31. 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, Stroudburg), 374-384.
32. Wijngaarden, S.J. van, Steeneken, H.J.M., Houtgast, T. (2002) “Quantifying the intelligibility of speech in noise for non-native listeners”, J. Acoust. Soc. Am. 111 (4), 2002.
33. Wijngaarden, S.J. van, Steeneken, H.J.M., Houtgast, (1999) “Objective prediction of speech intelligibility at high ambient noise levels using the Speech Transmission Index”. Proc. Eurospeech99, Budapest, 2639-2642.