Improvements of the STI method: frequency weighting, gender, level dependent masking, and phoneme specific prediction

Herman J.M. Steeneken

1. INTRODUCTION

In the early 1970s, the STI was used in many applied research projects. This provided experience for application in many practical conditions and also showed us some restrictions. For example, the validity of STI for systems with extremely limited bandwidth gave us the impression of underestimation of the intelligibility. We also recognised that extension with female speech for the prediction of the intelligibility was required. Therefore, in 1989, a study was started to cover these issues and in the course of this study some other issues were raised. The following topics were investigated:

(1) The original algorithms of the STI (Speech Transmission Index, Steeneken and Houtgast, 1980) and AI (Articulation Index, Fletcher, 1953; Kryter, 1962) assume that the prediction of the intelligibility is based on a weighted contribution for a number of frequency bands. For each band the *effective* signal-to-noise ratio (SNR) is determined. The effect of ambient noise, temporal distortion (reverberation, echoes), non-linear distortion (distortion components), and auditory masking, determine this effective signal-to-noise ratio. STI and AI assume that the frequency weighting is independent of the SNR. However, this was never tested experimentally and therefore requires verification.

(2) Application of the STI during the 1980s taught us that for some specific conditions, errors of the prediction of intelligibility by the STI occurred. This was especially the case for conditions that included gaps in the frequency transfer or in case of a very limited frequency transfer. The latter occurs with the use of small horn loudspeakers, which have a typical frequency response that begins around 1000 Hz (mainly due to a limited length of the horn). It was found that for some frequency regions redundancy between adjacent frequency bands of the speech signal occurred. This led to a revised model of the frequency-weighting algorithm for calculation of STI.

(3) At the time AI and STI were developed, only the prediction of the intelligibility of male speech was considered. Nowadays, female speech is used as frequently as male speech for almost any application. Hence, revision of algorithms to predict intelligibility should cover application for both male and female speakers. The frequency range of female speech generally starts at about 200 Hz rather than at about 100 Hz (for males), therefore the frequency weighting focused on female speech had to be determined separately.

(4) The frequency weightings found in the various experiments are different. Our experimental results reported in 1980 (Steeneken and Houtgast 1980) differ significantly from those reported in 1992 (Steeneken, 1992). In these studies the speech material (phonetically balanced nonsense words in 1980 and equally

balanced nonsense words in 1992) was different. The AI (presently referred to as SII, Speech Intelligibility Index, see ANSI standard S3.05) gives six different sets of frequency weightings resulting in the prediction of six related intelligibility scores. These include PB-words, sentences and rhyme words. We looked for a more universal description of frequency weighting and found a relation with the type of phonemes that were used with the speech material for the various studies. Ordering of the phonemes into four groups provided a more generic approach.

(5) The adverse conditions that include high noise levels and strong reverberations (e.g., in tunnels for traffic, at sports venues or industrial areas) require high output levels for PA-systems that are used. These lead to distortion components introduced by the hearing organ of the listener. This effect requires reconsideration of the contribution of masking by the STI algorithm.

This overview gives the results of our studies that focused on the improvement of the prediction accuracy of STI. The results of these studies were published in various papers: Steeneken (1992), Steeneken and Houtgast (1999, 2002a, 2002b), and Van Wijngaarden and Steeneken (1999).

2. RECONSIDERATION OF FREQUENCY WEIGHTING FUNCTIONS

The original model for STI and AI is based on additive contributions of frequency bands that cover the spectral range of speech signals. These models assume statistical independence between frequency bands. We are aware that energy contents in adjacent frequency bands may be correlated, hence the fluctuations in these bands show a high co-variance, and the information provided by such bands may be redundant. Therefore, an experiment was designed to estimate the contribution of individual frequency bands, and their mutual dependence. For this purpose, the speech spectrum was subdivided into seven octave bands with centre frequencies ranging from 125 Hz to 8 kHz. For 26 different combinations of three or more octave bands the CVC-word score (Consonant-Vowel-Consonant, nonsense words) was determined at three signal-to-noise ratios. It was found that for some specific frequency transfer conditions, considerable errors of the prediction of the CVC-word score by the STI were observed. This is shown in Fig. 1. In this graph, the relation between the CVC-word score and the STI is given for 26 frequency transfer conditions, at three signal-to-noise ratios. The frequency transfer conditions are based on various combinations of the seven octave bands that cover the frequency range of male speech. The combinations include: all possible *contiguous* selections of the seven octave bands, selection of three *non-adjacent* bands that introduces gaps in the frequency transfer, selection of three adjacent bands (*triplets*), and finally a selection which provides a *rippled envelope* of the frequency transfer. This leads to 26 different combinations. Each frequency transfer condition was used at three signal-to-noise ratios (SNR 15, 7.5, and 0 dB), thus obtaining 78 transfer conditions. For each of these conditions the CVC-word score was determined for four male and four female speakers, and eight listeners. The noise spectrum was equal to the average speech-spectrum of the speakers used for the experiments. Hence, the STI value could be calculated as the frequency transfer conditions and the SNR at each frequency band was known.

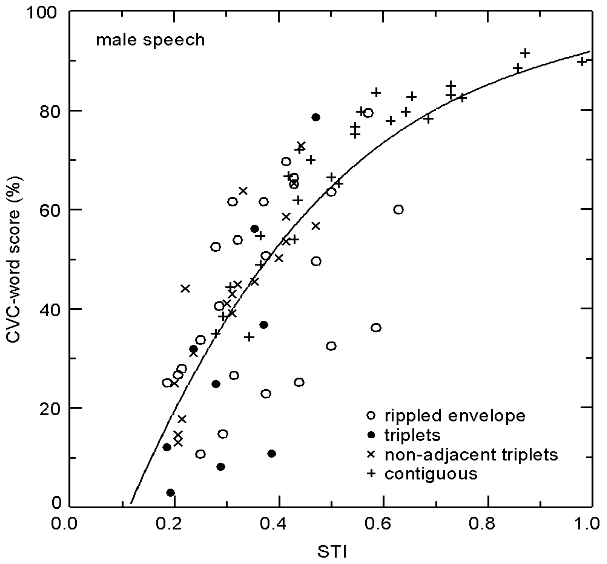

FIG2BAS 20-08-98 15:41:05

Figure 1. Relation between the STI and the CVC-word score for the 78 conditions involving MALE speech. The parameters used for the STI calculation were adopted from the procedure described previously by Steeneken and Houtgast (1980). The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 12.8%.

Evaluation of the experimental results led to a revision of the additive model, in which a redundancy correction between adjacent frequency bands, was introduced. This is given by:

$$\textit{index} = \alpha_1 \bullet \mathrm{TI}_1 - \beta_1 \bullet \sqrt{(\mathrm{TI}_1 \bullet \mathrm{TI}_2)} + \alpha_2 \bullet \mathrm{TI}_2 - \beta_2 \bullet \sqrt{(\mathrm{TI}_2 \bullet \mathrm{TI}_3)} + ... + \alpha_7 \bullet \mathrm{TI}_7 \qquad (1)$$

Where α represents the octave contribution weight, β the redundancy correction, and TI the transmission index based on the SNR in each band. For the three SNR values used in this experiment (15, 7.5, and 0 dB) the corresponding TI values are 1.0, 0.75, and 0.5. In an iterative procedure the weighting factors were optimized for optimal prediction of the CVC-word score by the index. The performance of this prediction can be expressed by the vertical spread of the data points around the regression line, this was for the original additive model without redundancy correction s=12.8% according to Fig. 1. The results of the revised model for male speech are given in Fig. 2, the corresponding vertical spread is s= 4.7%. In order to indicate that a revised STI model was used the index is given as STIr.

A similar experiment was performed for female speech. As the frequency range of female speech does not cover the octave band with center frequency 125 Hz some conditions of the original set of 26 transfer conditions had to be rejected. For female speech 17 different frequency transfer conditions were selected again at three signal-to-noise ratios. The results of this experiment are given in Fig. 3. The vertical spread s= 4.2%.

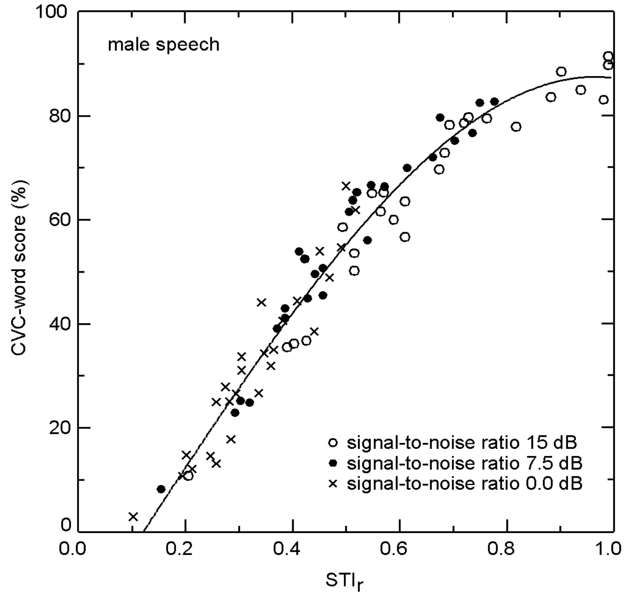

Figure 2. Relation between the STI$_r$ and the CVC-word score for the 78 conditions involving MALE speech. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s=4.7%.

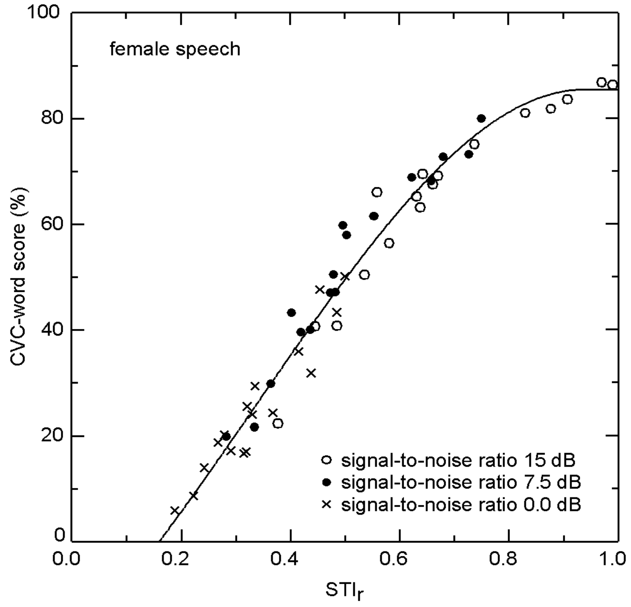

Figure 3. Relation between the STI$_r$ and the CVC-word score for the 51 conditions involving FEMALE speech. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 4.2%.

In Fig. 4, the frequency weightings and the corresponding redundancy corrections are given for male and female speech. Notice that these curves were obtained independently. The redundancies for very low frequency bands and the highest two bands are high. This can be explained by the formant structure of speech in relation with the bandwidth of the octave bands used in this experiment. The high weighting factor for the contribution of the octave band with center frequency 2000 Hz was also found in various other experiments. In relation to the low redundancy correction around this octave band it can be argued that, for a better resolution around the frequency axis, two half-octave bands should replace the octave band 2000 Hz.

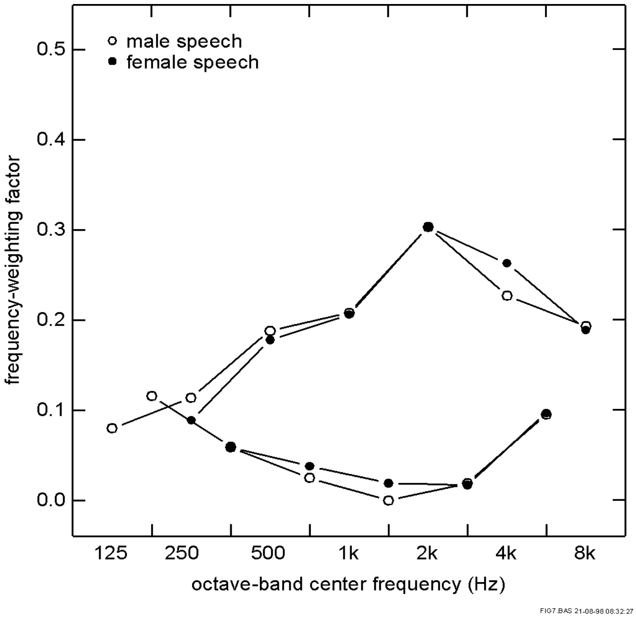

Figure 4. Optimal set of frequency-weighting factors α_k (upper curves) and redundancy-correction factors β_k (lower curves), derived separately for the male and female conditions as given in Figs. 2 and 3.

3. SIGNAL-TO-NOISE RATIO DEPENDENCY OF THE FREQUENCY WEIGHTING

STI and AI models convert signal-to-noise ratio to an index that represents the contribution to intelligibility, and use frequency weighting functions that are independent of the signal-to noise ratio. The validity had been evaluated only indirectly, by using test conditions that include many combinations noise and frequency transfer. Independent verification was not yet performed. The data of the experiments described in section 2 of this chapter allow for such verification. The first step is to separate the conditions for the three signal-to-noise ratios and to determine the optimal frequency weighting for each subgroup independently. The results for male speech are given in Fig. 5. The frequency weighting and redundancy correction (13 parameters) is based on 26 independent observations. This is not very much, but the three independent results (78 observations) show a good similarity.

Fig. 6 gives the STI_r values for the three noise conditions without including a correction for the signal-to-noise ratio, hence all TI values (eq. 1) are set to the maximum value (1.0). In this way three curves are obtained each curve representing a different signal-to-noise ratio. The vertical spread around these curves for signal-to-noise ratios of 15, 7.5, and 0 dB is respectively s = 3.4%, s = 4.0%, and s = 5.9%. Similar results were obtained for female speech with s = 2.4%, s = 4.3%, and s = 4.4%.

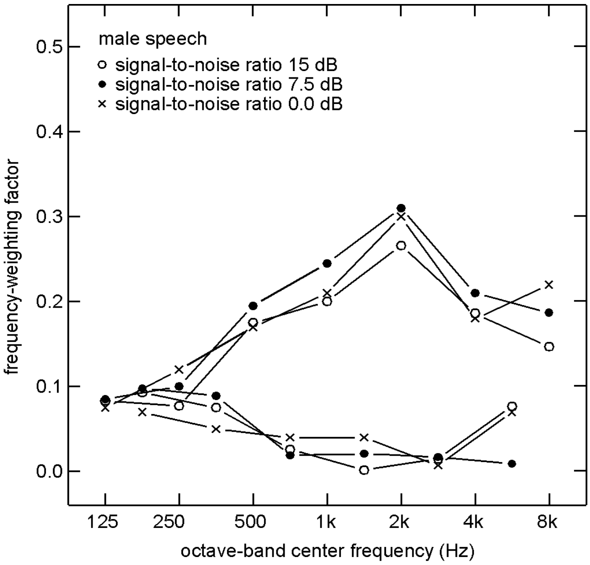

Fig.5. Frequency-weighting factors α_k (upper curves) and redundancy-correction factors β_k (lower curves) for the MALE conditions at three signal-to-noise ratios as given in Fig. 6.

The effect of a reduced signal-to-noise ratio in the STI model is accounted for by correction of the TI. According to:

$$TI_k = \frac{SNR_k + shift}{range}, \text{ where } 0<TI_k<1.0 \tag{5}$$

Where, TI_k represents the transmission index for octave band k and SNR_k (the effective signal-to-noise ratio). Range and shift are two parameters to convert an SNR value from –15 dB to 15 dB into the TI range 0–1. These range and shift parameter values are hidden in the horizontal shift between the three curves of Fig. 6.

According to the STI model, using range = 30 dB and shift = 15 dB, a correction of 0.75 and 0.50 is obtained for converting the curves for signal-to-noise ratios of 7.5 dB and 0 dB. We derived the same conversions (similar to eq. 2) from Fig. 6 for the male speech and from a similar graph for female speech. These correction values are given in Table I. These results show that the used TI correction is largely independent to the intelligibility range and fairly well predicted by the range and shift parameters of equation (2). We verified the effect of the increased TI_k values in comparison with the original values (0.8 versus 0.75, and 0.51 versus 0.50) and did not obtain a significant improvement of the prediction accuracy.

The assumption of independence of the frequency weighting and redundancy correction according to STI and AI seems to be correct.

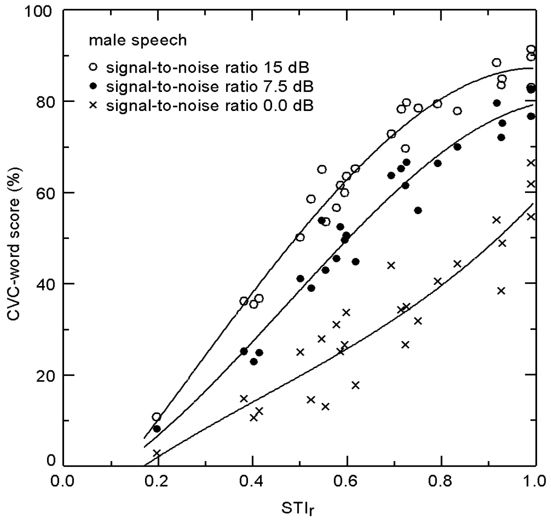

Figure 6. Relation between the STI$_r$ (with TI$_k$=1.0 for all three signal-to-noise ratios) and the CVC-word score for the conditions involving MALE speech.

Table I. Reduction of the transmission index TI corresponding to a reduction of the signal-to-noise ratio from 15 dB to 7.5 dB, and from 15 dB to 0 dB. The values are derived individually for the male speech (Fig. 6) and for the female speech, at three levels of the CVC-word score. The reduction according to the original STI concept (Eq. 2) is also given.

CVC word	SNR 15–7.5 dB			SNR 15–0 dB		
score	male	female	STI	male	female	STI
70%	0.80	0.80	0.75	-	-	0.50
50%	0.80	0.80	0.75	0.51	0.62	0.50
30%	0.81	0.85	0.75	0.51	0.58	0.50

4. LEVEL DEPENDENT MASKING

In many cases, the intelligibility of speech in noise may be assumed independent of the presented sound level to the listeners; the speech-to-noise ratio primarily determines the intelligibility. However, at high presentation levels, speech intelligibility is found to decrease. Subjective Speech Reception Threshold (SRT) measurements were performed at various speech and noise levels, and with various noise spectra. Decreases in intelligibility between noise levels of 75 and 105 dBA were found that correspond to 1 to 3 dB difference in signal-to-noise ratio, depending on the noise spectrum. This decrease is not predicted by the original STI. By introducing level-dependent auditory masking in the STI-calculations, a decrease in intelligibility can be predicted that corresponds well to the SRT results.

Rather than the fixed upward slope of masking of the original STI model (-35 dB/octave) a level dependent masking after Carter and Kryter (1962) was used according to Table II.

Table II Level dependent masking after Carter and Kryter (1962).

Octave level (dB)	*46–55*	*56–65*	*66–75*	*76–85*	*86–95*	*>95*
Slope of masking	–40	–35	–25	–20	–15	–10

The effect of the level dependent slope of masking was validated by comparison of the original model that includes a fixed slope of masking of -35 dB/octave and the new level dependent masking. This was performed at various signal levels between 75 and 105 dBA and for four types of noise (speech noise, traffic tunnel, low frequency boost, and for white noise). In Fig. 7 the deviation from the target value for the original model and the modified model are given. With this improvement an accurate prediction of the intelligibility, specifically for PA-systems that produce very high levels, can be given.

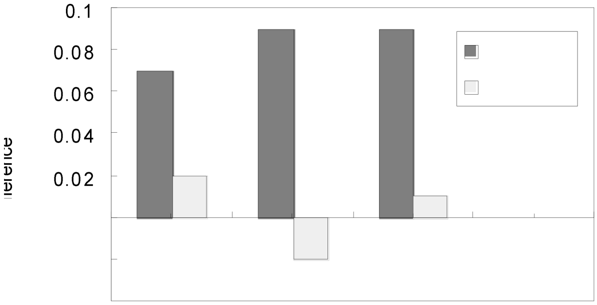

Figure 7. Differences in STI between 75 and 105 dBA (standard and modified model) based on male SRT results (2 talkers, 4 listeners).

5. PHONEME GROUP SPECIFIC WEIGHTING FUNCTIONS

As shown in section 2 and 3, frequency weighting functions do not vary significantly for signal-to-noise ratio or gender, other studies have shown that using different types of speech material, (i.e., nonsense words, phonetically balanced words, and connected discourse), resulted in quite different frequency weighting functions. In Fig. 8 three of these functions are compared, two are based on nonsense words and one on connected discourse "easy speech". The differences may be related to the distribution of specific phonemes in the test material.

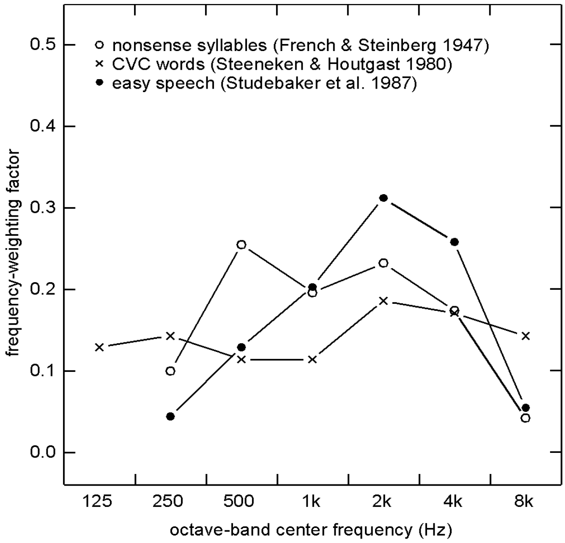

Figure 8. Frequency weighting functions for the AI and STI after French and Steinberg (1947), Steeneken and Houtgast (1980), and Studebaker et al. (1987).

In order to obtain a generic description of frequency weighting, relevant groups of phonemes were identified. Selection was based on confusions between phonemes, a high confusions rate at identical transfer conditions indicates similarity between the phonemes, a low confusion rate indicates dissimilarity. This is illustrated in a confusion matrix (Table III, consonants only) for the 26 transmission conditions with a different frequency transfer as described in section 2.

Table III. Cumulative confusion matrix for initial consonants, male speech, and 26 conditions of combinations of band-pass limiting. The initial consonants symbols are according the SAMPA notation (1987).

Response	p	t	k	b	D	f	s	v	z	x	m	n	l	R	W	j	h
Stimulus																	
P	1068	22	37	62	8	12	4	4	0	9	4	0	0	0	2	0	3
T	38	1099	51	0	29	6	3	3	1	3	0	0	0	0	0	0	2
K	52	58	1105	1	3	4	0	3	0	9	0	0	0	0	0	0	1
B	112	1	2	1002	41	0	0	0	0	0	11	7	2	0	50	3	3
D	8	113	16	49	1031	0	0	0	0	1	0	7	4	0	5	1	0
F	44	6	2	1	0	915	10	193	1	53	0	0	0	0	0	2	5
S	22	29	9	0	4	52	1037	13	41	14	0	0	0	0	0	1	1
V	6	3	1	4	1	337	11	739	35	34	0	0	0	2	43	11	8
Z	2	5	0	1	4	6	161	27	934	3	0	0	0	7	24	44	18
X	9	2	4	0	0	26	0	11	0	1083	0	0	0	1	0	0	12
M	1	0	0	5	0	0	0	0	0	0	1068	113	25	1	6	2	15
N	0	0	0	0	0	0	0	0	0	0	111	1081	33	0	2	7	1
L	11	0	0	0	0	0	0	0	0	0	12	59	1112	12	7	25	4
R	1	1	0	2	0	0	0	2	0	15	0	2	9	1161	3	1	39
W	6	0	0	3	7	1	0	13	2	0	30	7	5	25	1065	27	17
J	0	0	0	0	0	0	0	2	5	0	2	11	13	6	21	1163	12
H	9	0	1	8	0	4	0	4	0	6	7	1	3	12	16	20	1145

Three groups of phonemes show little confusion between each group and much confusion within each group (plosives, fricatives, and vowel-like consonants). For vowels such a clustering was not obtained therefore we classified vowels as one additional group. Thus, four groups were derived. For each group the optimal frequency weighting and redundancy correction was obtained with the same data set as used for the CVC-word assessments. From the CVC-word responses the phoneme group scores were obtained for this purpose. The TI values were obtained by measurement of the effective signal-to-noise ratio, as the TI values are different for each phoneme group and within each group for each octave band. This is due to the different long-term spectra of each group. The resulting weighting functions, for male and female speech, are given in Figs. 9-12.

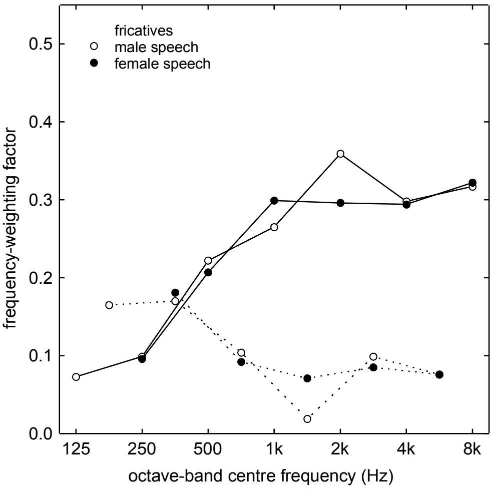

Figure 9. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the FRICATIVES and for the male and female speech.

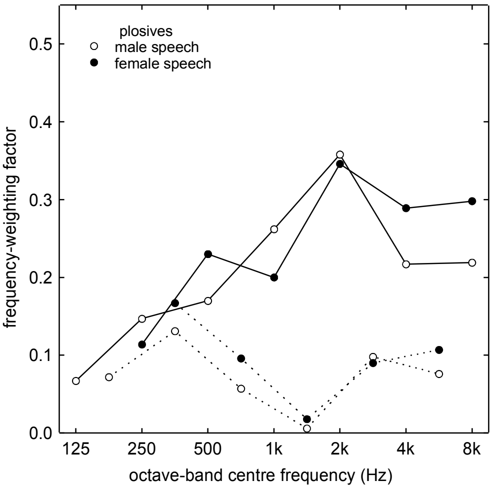

Figure 10. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the PLOSIVES and for the male and female speech.

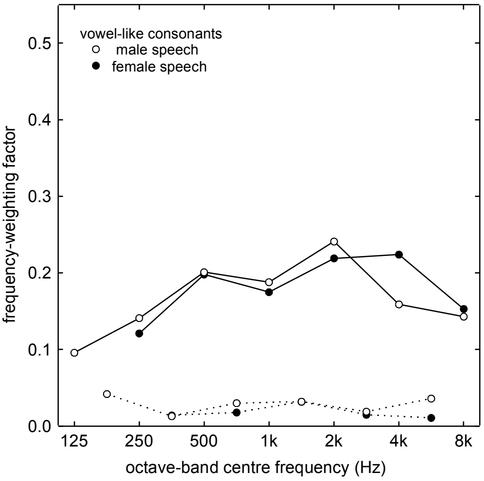

Figure 11. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the VOWEL-LIKE consonants and for the male and female speech.

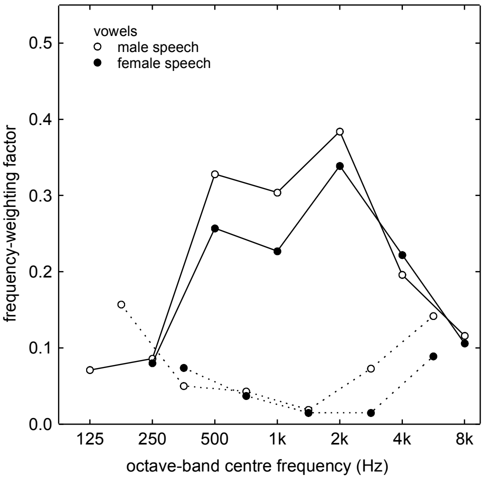

Figure 12. Frequency-weighting factors for the octave-band contribution αk (solid line) and redundancy correction ßk (dashed line) for the VOWELS and for the male and female speech.

Based on the frequency weighting and redundancy correction the specific STI values for each phoneme group were determined. The best fitting exponential curves between the phoneme-group score and the phoneme group specific STI-values are given in Fig. 13 for male speech and in Fig. 14 for female speech. Also the best fitting curves for the CVC-word scores are given in the same graphs. The respective standard deviations around the curves are Males fricatives s = 3.9%, plosives s = 5.6%, vowel-like consonants s = 4.0%, vowels s = 3.6%, CVC-words s = 4.6%; For female speech fricatives s = 4.3%, plosives s = 5.8%, vowel-like consonants s = 4.2%, vowels s = 2.9%, CVC-words s = 4.5%.

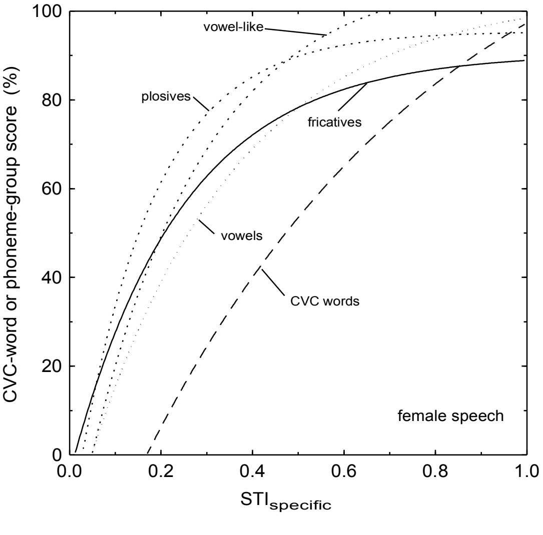

Figure 13 Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STIs for MALE speech. The relation for the CVC-word score is also given.

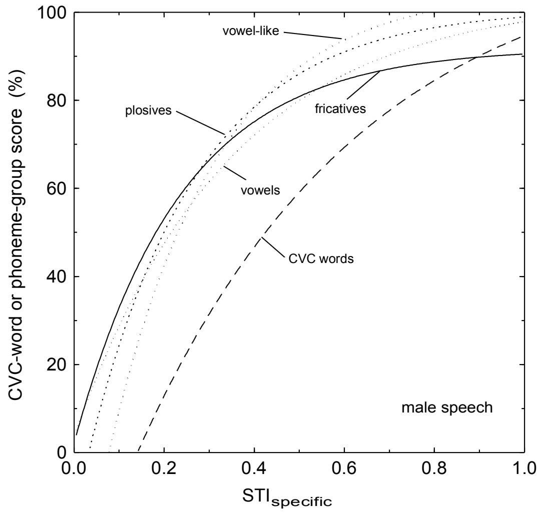

Figure 14. Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STIs for FEMALE speech. The relation with the CVC-word score is also given.

The phoneme group specific scores can be used to predict any other word score if the distribution of the phonemes of that score is known. We practised this by calculating the CVC-word score based on the individual phoneme scores and found for both male and female speech a very high correlation with the direct determined CVC-word score (males s = 4.1%, females s = 3.6%).

6. VALIDATION

The revised model for the Speech Transmission Index (STI_r, Steeneken and Houtgast, 1999), was validated with an independent set of 68 test conditions. For this set the STI values were obtained by measurement according to the STI measuring procedure also the CVC-word score was determined for four male and four female speakers and eight listeners. For a subset of 18 conditions, including only additive noise and band-pass limiting, it was verified that the STI_r provides a good prediction of the CVC-word score. The additional 50 conditions included non-linear distortion, echoes, automatic gain control, and wave-form coding. For conditions with these types of distortion specific parameters of the test signal are of interest. The parameters of the STI model were tuned in an earlier study for an optimal fit between the traditional STI and the CVC-word score, for a similar set of transmission conditions (Steeneken and Houtgast, 1980). It was found that these parameter settings also apply to the present revised model. The prediction accuracy for both male and female speech is 4-6% when expressed in CVC-word scores. This corresponds to a signal-to-noise ratio of about 1-2 dB.

In Fig. 15 and 16 the relation between the STI_r and the CVC-word score are given for 18 independent validation conditions. These conditions consist of combinations on various types of frequency transfer and four types of masking noise at various signal-to-noise ratios. This results in conditions with a wide variation of the contribution of each individual octave band. The data points in Figs. 15 and 16 are not plotted around the best fitting curve for these data but around the curves obtained with the development (see Figs 13, 14) in order to validate with independent data. The standard deviation representing the vertical spread around this predefined curve for male and female speech is respectively s = 4.4% and s = 6.6%.

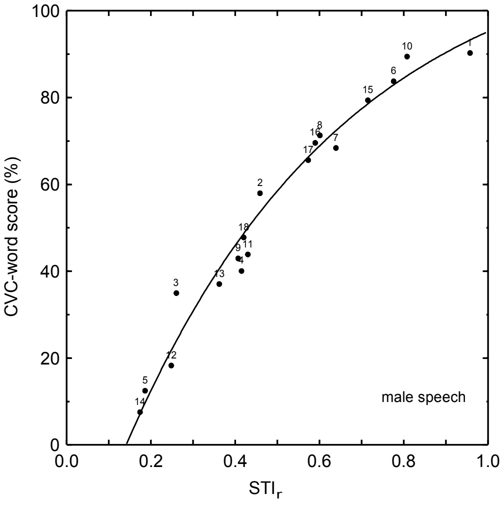

Figure 15. Relation between the STI$_r$ and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 4.4%.

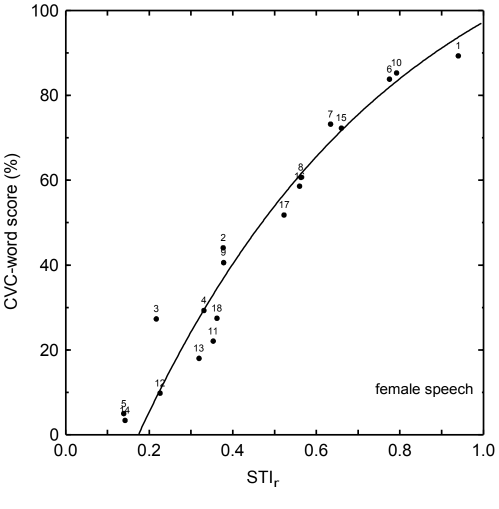

Figure 16. Relation between the STI$_r$ and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for FEMALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 6.6%.

Similar tests were performed with test conditions that include nonlinear distortion such as peak-clipping, center clipping and quantization noise (wave-form coders). The results for male speech are given in Fig. 17. Four data points are far beyond the optimal curve, these represent the center clipping conditions. The standard deviation representing the vertical spread around the predefined curve is s = 6.5% (excluding the four center clipping conditions). Similar results were obtained for female speech (s = 7.8%). The large overestimation of the intelligibility by STI of the conditions with center clipping is still a point of interest.

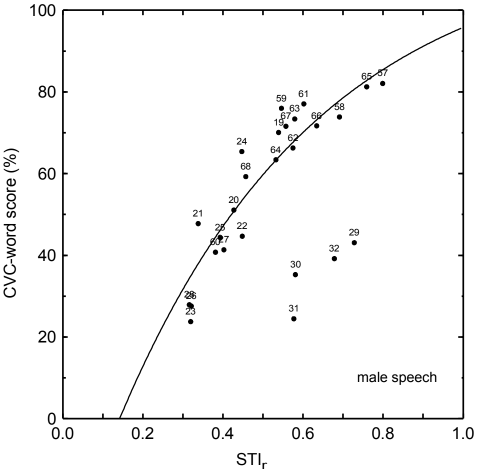

Figure 17. Relation between the STI_r and the CVC-word score for the 26 communication channel conditions including nonlinear distortion for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial for the conditions excluding center clipping (29-32) is s = 6.6%.

An advantage of the STI method is its validity for distortions in the time domain. This is achieved by considering the modulation transfer. We validated the present revised STI_r method also for conditions with this temporal distortion. Both automatic gain control and single echoes were used in combination with band pass limiting and noise. This provided 24 different transmission conditions. The relations between STI_r and the CVC-word-score for these conditions are given in Fig. 18. The vertical spread around the previous defined relation is s = 6.9% (female speech s = 8.2%).

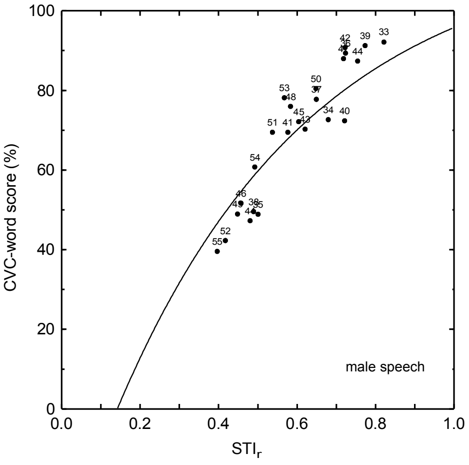

Figure 18. Relation between the STI$_r$ and the CVC-word score for the 24 communication channel conditions including distortion in the time domain for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 6.9%.

7. CONCLUSION AND FUTURE DEVELOPMENTS

The revision and validation of the STI method as presented in earlier studies resulted in the following improvements:

- the effect of a discontinuous frequency transfer is accounted for correctly,
- an extension was made for female speech,
- for diagnostics purpose the method was extended to predict phoneme-group scores for vowels, plosives, fricatives and vowel-like consonants,
- the relation between various subjective intelligibility measures and STI$_r$ are very similar to the relation found in the 1980 study. Hence, previously adopted criteria for various applications are still valid.
- adaptation of the STI$_r$ method for high signal and noise levels is included.

Present research is focused on the replacement of the artificial test signal by a standard speech signal. The use of speech as test signal applied in room acoustics was already presented by Steeneken and Houtgast (1983). An extension will be made for the applications with non linear systems and coders.

REFERENCES

Fletcher, H., (1953). Speech and Hearing in Communication (D. van Nostrand, New York).

Carter, N.L. & Kryter, K.D. (1962). Masking of pure tones and speech. Journal of Auditory Research, 2, 66-98.

Kryter, K.D., (1962). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. 34, 1689-1697.

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). "On measuring and predicting speech intelligibility" Doctoral thesis University of Amsterdam

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.

Steeneken, H.J.M., and Houtgast, T., (2002a). "Phoneme-group specific octave-band weights in predicting speech intelligibility". Speech Communication, 2002, vol.38.

Steeneken, H.J.M., and Houtgast, T., (2002b). "Validation of the revised STIr method". Speech Communication, 2002, vol.38.

Wijngaarden, S.J. van, Steeneken, H.J.M. (1999). "Objective prediction of speech intelligibility at high ambient noise levels using the speech transmission index" In Eurospeech99 - Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, Vol. 6, pg. 2639-2642.