TECNIACUSTICA PAMPLONA • 1992

OBJECTIVE EVALUATION OF AUDITORIA
Assessment of speech intelligibility from acoustical measurements

Herman J.M. Steeneken and Tammo Houtgast

Institute for Perception TNO, P.O. Box 23,
3769 ZG Soesterberg, The Netherlands.

1 INTRODUCTION

In room acoustics a condition is generally characterized by means of the reverberation time and the signal-to-noise ratio. For use in auditoria more relevant, speech related, measures can be considered. An obvious method is to determine the speech transmission quality between two positions by means of a subjective intelligibility test with speakers and listeners. For a good description of the speech transmission in an auditorium, many listening positions have to be considered. However, these subjective tests are time consuming, and do not offer diagnostic information which can be used to improve the performance.

Another method is to predict speech intelligibility from physical measurements. Running speech can be considered as a sequence of speech sounds and short silent periods. The speech sounds consist of signals with different frequency spectra and preservation of these differences leads to a high intelligibility. The amount of preservation can be determined for speech signals but also for artificial test signals. This method results in an objective measure which can predict the intelligibility for a certain condition. Based on the results of these physical measurements also other measures can be obtained, such as reverberation time, signal-to-noise ratio, and frequency transfer. In general, objective measures offer some diagnostic information, from which the origin of a reduced transmission can be determined.

In this tutorial a few subjective and one objective intelligibility measure will be described, and application examples will be given.

2 SUBJECTIVE INTELLIGIBILITY ASSESSMENT

Subjective intelligibility tests involve a variety of different aspects: items tested, diagnostic information, minimum number of subjects required for reliable results, training and measuring time. Another aspect is the application: are we comparing and *rank-ordering* conditions or systems, are we evaluating a system for a *specific application* or are we supporting the *development* of a system?

A general classification of these tests can be made based on the items tested and the method of response. The lowest level (segmental evaluation i.e. phonemes) is covered by the rhyme tests and the open response word tests.

A rhyme test is a multiple-choice test where a listener has to select the auditorily presented word from a small group of visually presented possible responses. In general only the initial consonants of the response words are changed such as for the plosives Bam, Dam, Tam, Kam, Pam. A frequently used rhyme tests is the Modified Rhyme Test (MRT) [1].

A more general approach is obtained with an open response, as used with word tests. Word tests are based on short nonsense or meaningful words of the CVC-type (Consonant-Vowel-Consonant). Sometimes only CV-words are used. The test words are presented in isolation or in a carrier phrase. The listener can respond with any CVC combination he or she has heard. Hence all confusions between the phonemes are possible. The test results include the phoneme score, the word score and the confusions between the initial consonants, vowels and final consonants. The confusion matrices present useful information to improve the performance of a system [13,14].

Quality rating is a more general method used to evaluate the user's acceptance of an auditorium or a transmission channel. For quality ratings normal test sentences or a free conversation is used to obtain the listener's impression. The listener is asked to rate his impression on a subjective scale like the five-point scale: bad, poor, fair, good, and excellent. Different types of subjective scales are used, such as intelligibility, quality, acceptability or naturalness.

Fig. 1 gives the score for five intelligibility measures as a function of the signal-to-noise ratio of speech combined with noise [14]. This gives an impression of the effective range of each test. The given relation between intelligibility and the signal-to-noise ratio is only valid for noise with a frequency spectrum equal to the long-term speech spectrum. This is for instance the case with voice babble. A signal-to-noise ratio of 0 dB means that the speech and the noise have equal energy. (The STI-scale in Fig. 1 will be introduced in the next section.)

As can be seen from the figure the nonsense CVC-words discriminate over a wide range, while meaningful test words have a slightly smaller range. The digits and the alphabet give a saturation at a SNR of about -5 dB. This is due to: (a) the limited number of test words and (b) the recognition of these words being mainly controlled by the vowels rather than the consonants. Vowels have an average level approximately 10 dB above the average level of consonants, and are therefore more resistant against noise. On the other hand non-linear distortion as clipping will have a greater impact on the vowels than on the consonants. Therefore the exclusive use of the digits and the alphabet, where the recognition is mainly based on vowels, may lead to misleading results.

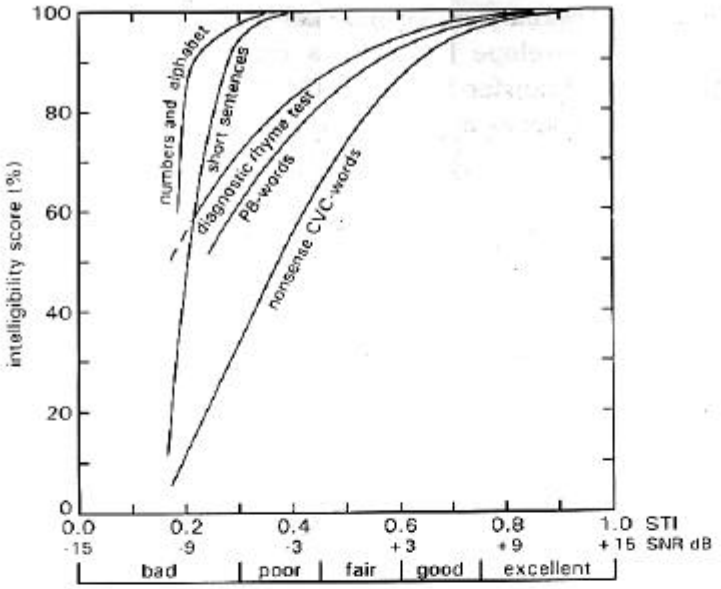

Fig. 1 Qualification of some intelligibility measures and their relation with signal-to-noise ratio (SNR) for noise with a spectrum shaped according to the long-term speech spectrum. (The STI-scale is introduced in section 3.)

A well balanced test, as has been found in our earlier studies, is the CVC-word test based on nonsense words and with the test words embedded in a carrier phrase. The use of a carrier phrase (which is neglected in many studies) is essential for applications in auditoria as it will cause echoes and reverberation similar to running speech [5]. Another aspect of using a carrier phrase is that it stabilizes the vocal effort of the speaker during the pronunciation and minimizes the vocal stress on the test words.

3 OBJECTIVE INTELLIGIBILITY ASSESSMENT

3.1 **Envelope function and envelope spectrum**

Connected discourse can be considered as a sequence of the smallest speech items, the phonemes. Each phoneme is represented by a specific frequency spectrum. For the recognition of a speech token the differences between the spectra of the phonemes must be preserved. These spectral differences can be described by the envelope function within a number of frequency bands. Distortion of the speech signal, such as noise or reverberation, will result in a reduction of the spectral differences between the spectra of the corresponding phonemes. This is also reflected in the envelope function by a reduction of the fluctuations. In Fig. 2 (panel A) the envelope function, for the octave-frequency band 250 Hz, is given.

The shape of the envelope function is unique for a specific sequence of phonemes. A more general description of the fluctuations in the envelope function is given by the *envelope spectrum*. The envelope spectrum results from a 1/3-octave-band analysis of the envelope function (typically of a one-minute speech fragment), and reflects the spectral distribution of the envelope fluctuations relative to the mean intensity: the modulation index as a function of modulation frequency (Fig.

1 panel B). The difference between the original and the resulting envelope spectrum reflects the reduction in the envelope fluctuations caused by the transmission path. This leads to the Modulation Transfer Function (MTF) which represents the reduction factor of the modulation index as a function of modulation frequency.

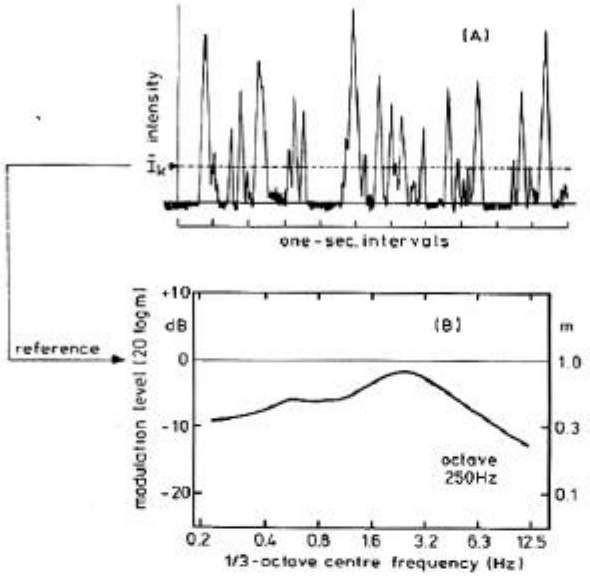

Fig. 2 Envelope function (panel A) of a 10 s speech signal for the octave band with centre frequency 250 Hz. The corresponding envelope spectrum (panel B) is normalized with respect to the mean intensity (I_k).

3.2 The Modulation Transfer Function (MTF)

The rationale underlying the application of the MTF concept in room acoustics has been described in various papers [2,6,9]. The MTF quantifies to what extent the modulations in the original signal are reduced, as a function of the modulation frequency. The modulations are defined by the *intensity envelope* of the signal: it is in the intensity domain, that interfering noise or reverberation will affect only the degree of modulations of a sine-wave shaped modulation *without* affecting the sine-wave shape. The scheme in Fig. 3 illustrates how the MTF may be used to quantify the relation between the original speech signal at the input and the output signal (A or B). Since most disturbances may vary considerably as a function of frequency, the analysis is octave-band specific. The example of Fig. 3 considers one octave band only, i.e., the intensity envelopes in the octave band with centre frequency of 500 Hz. Two simple sound transmission systems are illustrated, one with reverberation only (case A; T = 2.5 s) and one with interfering noise (case B; signal-to-noise ratio S/N = 0 dB.

Typically, as may be observed in Fig. 3, in the case of reverberation the MTF has the form of a low-pass filter: the faster fluctuations are relatively more affected by reverberation. In the theoretical case of an ideal exponential reverberation process, the MTF is defined mathematically (see Fig. 3). The typical low-pass character is determined by the product FT (F = modulation frequency, T = reverberation time). In the case of noise interference, the MTF is defined by the S/N ratio and is independent of modulation frequency: the interfering noise results in an increased mean

intensity and thus reduces the (relative) modulation index for all modulation frequencies by the same factor.

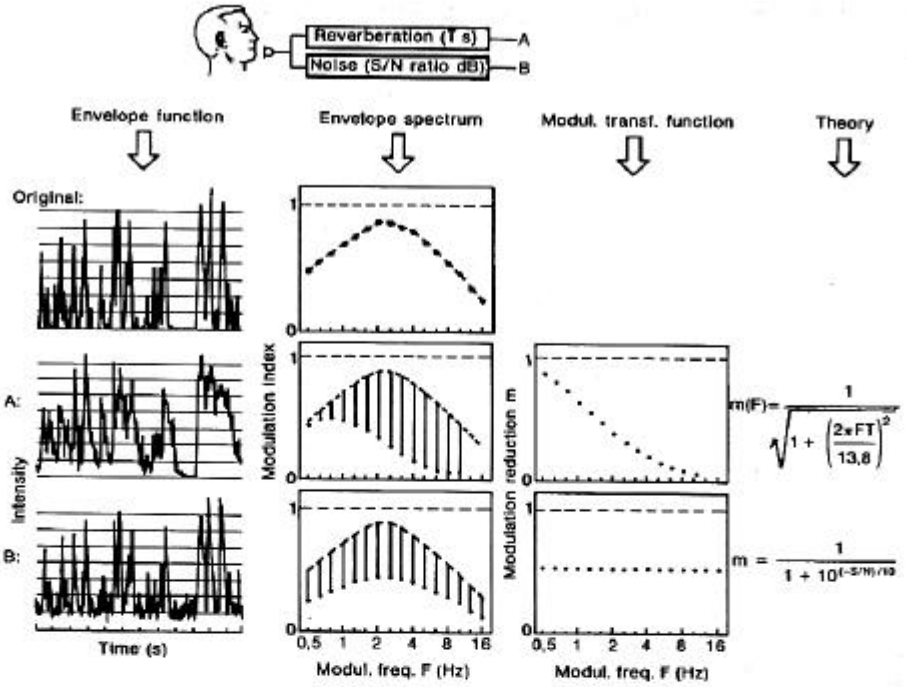

Fig. 3 The reduction of the fluctuations in the (octave-band specific) envelope of an output signal (A or B) relative to the original signal can be expressed by the Modulation Transfer Function. The two conditions considered (reverberation or noise interference), lead to characteristic MTFs, according to the theoretical expressions given at the right-hand side.

It is important to note that the (octave-band specific) MTF of a sound transmission system is *independent* of the input signal considered. It quantifies the modulation transfer for any input signal: speech, music or an artificial signal.

The MTF of a sound transmission system can be determined in various ways, the principle always being that the modulation reduction factor is derived from a comparison of the intensity modulations at the output and at the input of the system. For this purpose either (1) speech signals, (2) the impulse response of a system, or (3) specific test signals can be used.

ad. 1. The use of speech signals has the advantage that the MTF of an enclosure can be determined during a life performance. However the method is less accurate than the use of artificial test signals [11].

ad. 2. The impulse response can be used to determine the effect of reverberation or echoes on the MTF. The impulse response method is not suited for including the effect of background noise and non linear distortion (PA systems).

ad. 3. The use of an artificial test signal allows the determination of the modulation reduction factor for each modulation frequency successively. In principle, the test signal is produced at the position of the speaker's mouth. It consists of a noise carrier with 100% intensity modulation. The remaining modulation index at a listener's

location directly reflects the modulation transfer for that particular modulation frequency. The noise carrier is octave-band filtered, and the measurements are performed for different centre frequencies (typically from 125 Hz up to 8 kHz). For a representative determination of the signal-to-noise ratio, the mean intensity of the test signal should be related to that of the speech as normally produced by a speaker at that position. As a rule, for each octave band considered, the L_{eq} of the test signal is to be adjusted to the L_{eq} of ongoing speech typical for the condition being tested. An example of this measuring scheme is given in Fig. 4.

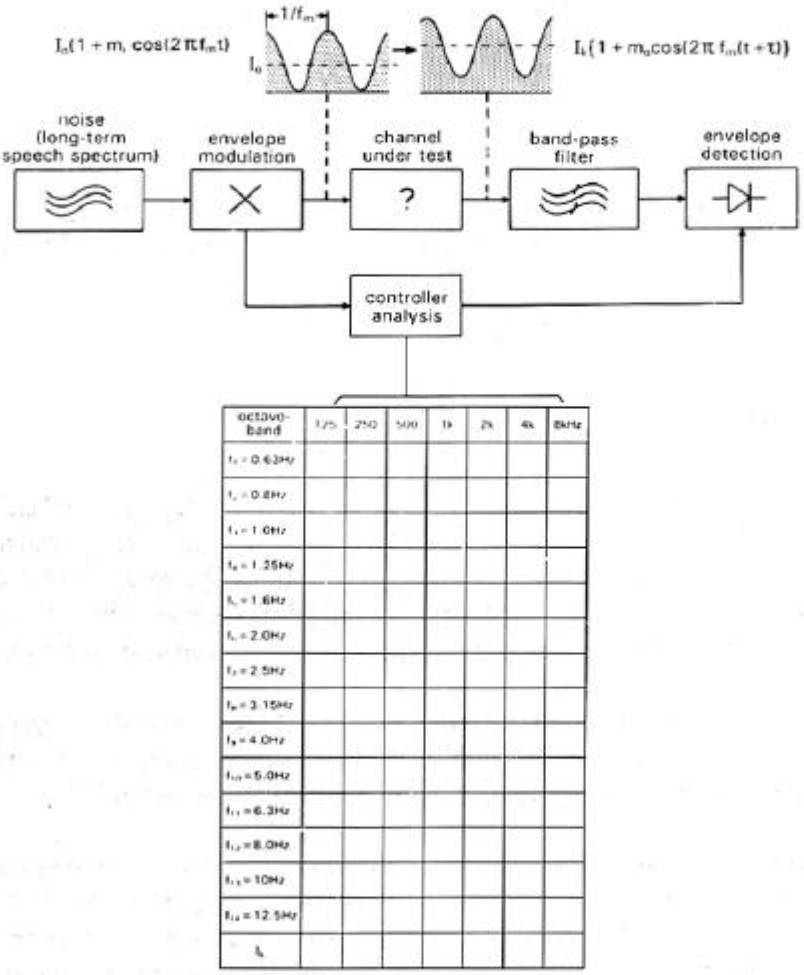

Fig. 4 General block diagram of the measuring set-up. The modulation index reduction at the output (m_o/m_i) is determined for all cells of the matrix (7 octave bands and 14 modulation frequencies). Also the octave levels are obtained, for calculation of the auditory spread of masking.

Based on an adequate test signal, the performance of a sound transmission system can be quantified by a family of MTF curves, comprising 7 × 14 = 98 m-values. The question remains of how to transform such a set of data into one single index representing the effect of that transmission system on speech intelligibility: the Speech Transmission Index (STI). The criterion for the relevance of such a transformation is, of course, that for a wide variety of transmission systems with different types of

disturbances, the relationship between the STI-values and the effect on speech intelligibility is unique, i.e., not system specific.

3.3 The Speech Transmission Index (STI)

The algorithm for conversion of a set of m-values into a STI-value, and the experimental verification on the basis of numerous intelligibility tests, is fully described elsewhere [2,9]. An essential step in this transformation is a conversion of each of the 98 m values into an *apparent* signal-to-noise ratio $(S/N)_{app}$, irrespective of the actual type of disturbance causing the m value. It is interpreted as if it had been caused by interfering noise exclusively, $(S/N)_{app}$ being the signal-to-noise ratio which should have resulted in that m value. The conversion is defined mathematically by

$$(S/N)_{app} = 10 \log \frac{m}{1-m} \text{ dB} \qquad (1)$$

being the inverse of the expression given in Fig. 3. A weighted average of the 98 apparent signal-to-noise ratios thus obtained results in the STI, after applying appropriate normalisation such that:

$$(S/N)_{eff} = 15 \text{ if } (S/N)_{app} \geq 15 \text{ dB} ,$$

$$(S/N)_{eff} = -15 \text{ if } (S/N)_{app} \leq -15 \text{ dB} .$$

By this calculation scheme each family of MTF curves can be transformed unambiguously into a STI value, by which the performance of that sound transmission system is quantified. Also, given the theoretical relations between $m(F)$ and the reverberation time T or the S/N ratio, the calculation scheme may be used for theoretical studies on the effect of reverberation and ambient noise in general.

It has been shown that the STI calculation scheme can be used to predict the performance of an auditorium in the design stage, especially when modelling the sound field along the lines of geometrical acoustics, i.e., by ray-tracing [8].

There exists a large body of experimental data on the relation between the STI and intelligibility scores obtained with speaker-listener panels [9,14]. Typical relations are given in Fig. 1. These relations are only illustrative since, besides the performance of the transmission system, intelligibility scores are affected by other factors also, such as the degree of training and skill of the speaker-listener panel and specific aspects of the speech material employed in the test (e.g., the use of a carrier phrase).

The qualification intervals (bad...excellent) specified along the abscissa in Fig. 1 are based on a large-scale study [5,9], involving various intelligibility tests and different languages.

In the middle range, each qualification interval corresponds to an interval of 0.15 along the STI scale. This implies that differences of that magnitude are important: for two conditions with a STI difference of 0.15, the difference in speech intelligibility is significant and clearly noticeable. Accordingly, for an actual STI-measuring device one requires that the accuracy interval (e.g., the standard deviation for

repeated measurements) is considerably smaller than 0.15. This may serve as a guideline for the implementation of the MTF concept in room acoustics along the lines presented in this contribution.

A screening device according to the concept described above has been developed. This device, RASTI (RApid STI) is described in IEC recommendation R269.

3.4 Estimation of the early decay time (EDT) and the signal-to-noise ratio from the MTF

The distortions affecting the intelligibility of a speech signal in an enclosure can be divided into two groups:

— signal-independent disturbing sounds, such as background noise introduced by air conditioners, traffic or by the public;
— signal-dependent disturbing sounds, such as reverberation and echoes.

Both types of disturbing signals have their specific effect on the modulation transfer function. The degradation introduced by background noise and the degradation introduced by reverberation can therefore be estimated individually from the MTF (see also [12]). In principle it is possible to obtain the delay time and the relative strength of an echo as well. However, for such an estimation the MTF has to be described with a high resolution in the modulation-frequency domain.

4 APPLICATION EXAMPLES

iso-STI Contours

The normal procedure for determining iso-STI contours is to measure STI values at a large number of positions evenly distributed throughout the audience area. In this way the STI can be mapped and used to construct iso-STI contours.

Depending on the gradient between successive STI values and on the resolution of the measuring grid, iso-STI contours can be drawn for 0.05, 0.1, or 0.2 STI intervals. In Fig. 5 iso-STI contours, based on 29 measuring positions, are given for a lecture hall. In this example, for the empty hall and no background noise, the STI varies from 0.70 to 0.58 which implies an intelligibility rating between good and fair. Normally, the acoustics consultant starts with a measuring session in the absence of an audience, which may result in a non-representative absorption and noise level. The absence of a representative background noise can be compensated for by the application of an artificial noise source during the measurements or by correcting the STI for an imaginary background noise. In the latter case we have to correct the MTFs for a certain noise level.

Fig. 6 shows iso-STI contours obtained from the same measuring data as given in Fig. 5, but corrected for an imaginary background noise with a level of 40 dB in each octave band.

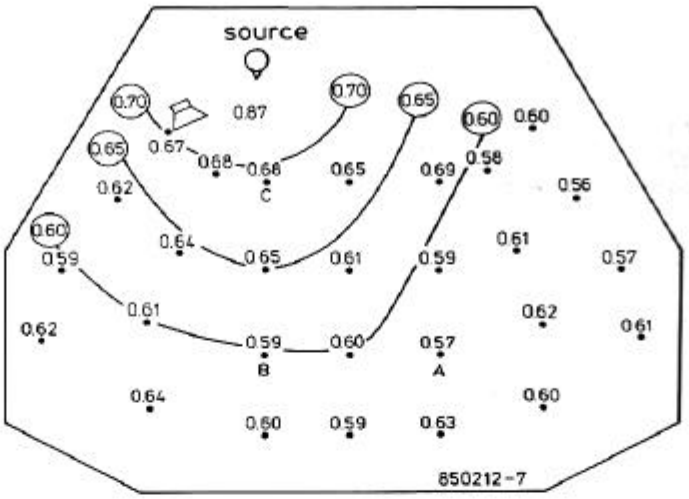

Fig. 5 iso-RASTI contours for an auditorium, without public, background noise and PA-system. The original (29) data points where the contours estimated from, are given as well.

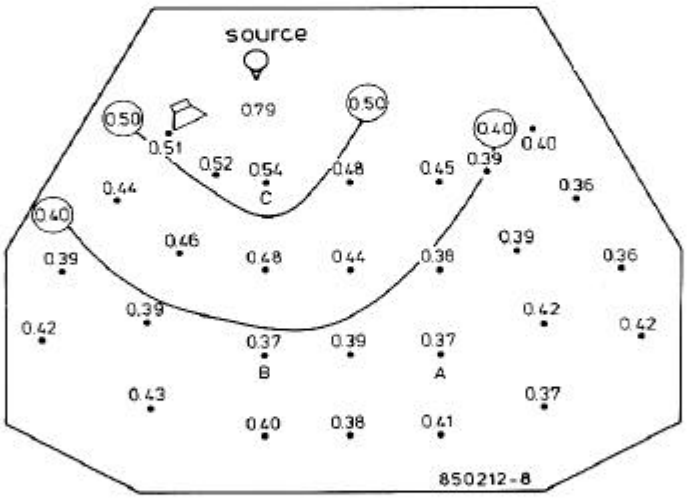

Fig. 6 iso-STI contours for the same data points as given in Fig. 5, but corrected for an imaginary background noise with an octave-band level of 40 dB.

Under the condition presented in Fig. 5, with only small differences of the STI, it is not very relevant to detect areas in the audience with a poor intelligibility. However, for an imaginary background-noise the iso-STI contours change dramatically. As given in Fig. 1 the STI values range from 0.54-0.36 which means that areas with a poor intelligibility can be detected, caused by a low level of the direct sound far from the speaker or far from any reflecting surface.

For three positions in the auditorium (marked A, B, C in Figs 5 and 6) the STI as a function of the background-noise level is given in Fig. 7 A, B, C (solid lines).

These graphs indicate a low signal level at the positions A and B. Besides acoustical measures, such as a reflecting surface behind the speaker, a public-address system (PA) can be applied to increase the level of the (direct) sound.

The Evaluation of a Public-Address System with the RASTI

The application of a PA-system in an auditorium increases the direct-sound level at the listener's position and hence the signal's resistance against background noise. The signal level at the listener's position is defined by the system gain and by the position and directivity of the microphone and loudspeakers, and also by the acoustics of the room.

For a poorly designed system, however, with the loudspeakers not optimally directed to the (absorbing) audience, the reverberation field increases, which may result in a decrease of the intelligibility at low noise levels. An example of such a situation is given in Fig. 7 C (dashed curve). For this condition a PA-system in the auditorium as given in Fig. 5 was applied. Only one loudspeaker at the marked position was used. The loudspeaker was placed above the audience, directed to position B.

The STI was measured at positions A, B, and C, and the results, as a function of an imaginary background-noise level, are given in Fig. 7 (dotted lines). For position B the STI value is increased even for the conditions without background noise, which implies a better ratio between the direct and the indirect sound.

We can estimate the contribution of the PA-system by the increase of the resistance against background noise for a given, critical STI-value. As shown in Fig. 7 B this "effective" gain is 16 dB for a STI value of 0.4. For position A this effective gain (of the same PA-system) is 11 dB and at position C it is 0 dB. With this method of validation, an optimal adjustment of the loudspeaker positioning and direction can be found.

In order to exclude the contribution of the transmitting room and the microphone and the re-transmission of the indirect sound by the system, the STI test signal can be connected electrically to the PA-system.

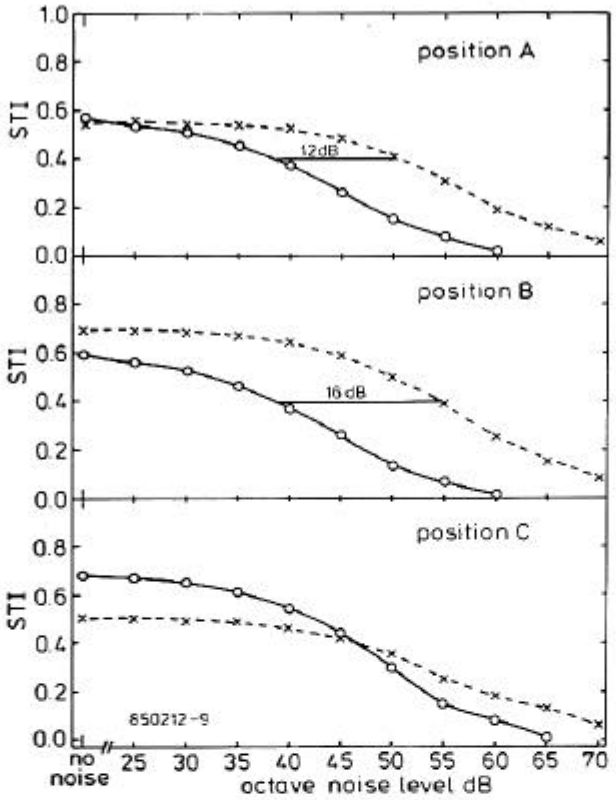

Fig. 7 STI value as a function of an imaginary background-noise level for three positions (marked A, B, C in Fig. 5) and without (o) and with (x) a public-address system. The position of the loudspeaker is also marked in Fig. 5.

Microphone performance in a noise environment

Gradient microphones are developed for use in a high noise environment. The specifications, given by the manufacturers, normally describe the effect of the noise reduction in general terms and are not related to intelligibility, microphone position or type of background noise. In Fig. 8 the transmission quality, expressed by the STI, for two types of microphones is given as a function of the environmental noise level. For these measurements an artificial head was used to obtain the test signal acoustically. The microphone was placed on this artificial head at a representative distance from the mouth. The test signal level was adjusted according to the nominal speech level. This signal level can be increased and the spectrum can be tilted in order to simulate the increase of the vocal effort of a talker in noise (Lombard effect). The head was placed in a diffuse sound field with an adjustable level.

From the figure we can see that the distance from the mouth is an important parameter. It is also obvious that the two noise-cancelling microphones have a different performance in combination with the noise as used in this experiment.

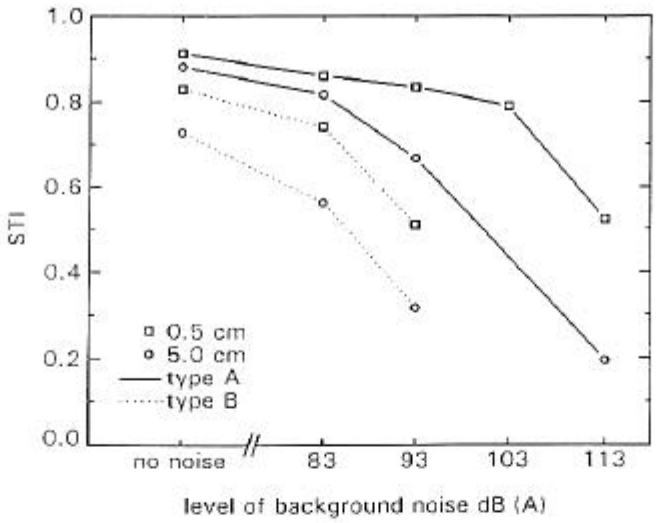

Fig. 8 STI as a function of the noise level for two different microphones and two speaking distances.

5 RÉSUME

Both subjective and objective intelligibility measures were discussed and examples were given. In general the subjective methods, used for room acoustics, required much effort. Care should be taken to adapt the test words for adequate use in reverberant environments. Objective measures, i.e. the STI-method, require a simple straight forward measurement and offer diagnostic information concerning the type of degradation introduced by the transmission path.

6 REFERENCES

[1] Fairbanks, G. (**1958**). "Test of phonetic differentiation: The Rhyme Test", *J. Acoust. Soc. Am.* 30, 596-600.

[2] Houtgast, T. & Steeneken, H.J.M. (**1973**). "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility", *Acustica* 28, 66-73.

[3] Houtgast, T., Steeneken, H.J.M. & Plomp, R. (**1980**). "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I. General Room Acoustics", *Acustica* 46, 60-72.

[4] Houtgast, T. & Steeneken, H.J.M. (**1983**). "Experimental Verification of the STI", *Proceedings Fourth International Congress on Noise as a Public Health Problem* (Turin), 477-487.

[5] Houtgast, T. & Steeneken, H.J.M. (**1984**). "A Multi-Language Evaluation of the RASTI-Method for Estimating Speech Intelligibility in Auditoria", *Acustica* 54 (1984), 185-199.1

[6] Houtgast, T. & 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. Amer.* 77, 1060-1077.

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

[8] Rietschote, H.F. van & Houtgast, T. (**1983**). "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. V. The Merits of the Ray-tracing Model Versus General Room Acoustics", *Acustica* 53, 72-78.

[9] Steeneken, H.J.M. & Houtgast, T. (**1980**). "A Physical Method for Measuring Speech-Transmission Quality", *J. Acoust. Soc. Amer.* 67, 318-326.

[10] Steeneken, H.J.M. & Houtgast, T. (**1982**). "Some Applications of the Speech Transmission Index (STI) in Auditoria", *Acustica* 51, 229-234.

[11] Steeneken, H.J.M. & Houtgast, T. (**1983**). "The Temporal Envelope Spectrum of Speech and its Significance in Room Acoustics", in: *Proceedings 11th International Congress on Acoustics*, Paris, Vol. 7, 85-88.

[12] Steeneken, H.J.M. & Houtgast, T. (**1985**). "RASTI: A Tool for Evaluating Auditoria", *B&K Technical Review* 3, 13-30.

[13] Steeneken, H.J.M. (**1992**). "Quality evaluation if speech processing systems," Chapter 5 in *Digital Speech Coding: Speech coding, Synthesis and Recognition*, edited by Nejat Ince, (Kluwer Norwell USA), 127-160.

[14] Steeneken, H.J.M. (**1992**). "On measuring and predicting speech intelligibility" Doctoral thesis Univ. of Amsterdam.