Past, present and future of the Speech Transmission Index

Past, present and future of the Speech Transmission Index

Tammo Houtgast
Herman Steeneken

Wolfgang Ahnert
Louis Braida
Rob Drullman
Joost Festen
Kenneth Jacob
Peter Mapp
Steve McManus
Karen Payton
Reinier Plomp
Jan Verhave
Sander van Wijngaarden

Introduction by *Manfred R. Schroeder*
Edited by *Sander J. van Wijngaarden*

Published by:

TNO Human Factors, PO Box 23, 3769 ZG Soesterberg, The Netherlands

ISBN 90-76702-02-0

Tammo Houtgast and Herman Steeneken have worked together at TNO Human Factors for close to four decades. Many of us at TNO, and literally thousands of engineers, scientists, consultants and end-users around the world, enjoy the scientific and technological fruits of their work.

On the eve of their retirement, a symposium was organised in their honour, on the topic that represents their most important joint project: the Speech Transmission Index, or STI. This book was published as a written companion to the symposium.

Despite the fact that Herman and Tammo are responsible for writing the largest part of this book (who else?), we would like to dedicate it to them both.

Herman and Tammo, may your retirements be as enjoyable and fulfilling as your professional careers!

Your friends and colleagues of TNO Human Factors

Table of contents

Tammo's and Herman's Index

An introduction by Manfred R. Schroeder

University of Goettingen, Germany
and AT&T Bell Laboratories (ret.)

Few suggestions for quantifying subjective aspects of human perception have found such widespread acceptance as the Speech Transmission Index, proposed by Tammo Houtgast and Herman Steeneken beginning with their 1970 Dutch paper "Beoordelen van spraakcommunicatiekanalen langs fysische weg" (evaluation of speech transmission channels by physical [quantitative] means), and further elaborated in 1971 by Steeneken and Houtgast "Evaluation of speech transmission channels by using artificial signals" [*Acustica* **25**, 355-367]. This was followed by several ground-laying articles with Reinier Plomp and others from TNO Technische Menskunde ("Human Factors" – literally, and perhaps more appropriately – "technical knowledge of man").

In these papers the term *modulation transfer function*, long known in visual perception in connection with flicker fusion in television, also appeared for the first time in articles in connection with speech intelligibility. The term was so unusual that my own paper on modulation transfer functions [*Acustica* **49**, 179-182] was rejected as "unsuitable" by a reputable IEEE publication on acoustics, whereupon I buried it in my desk drawer only to resurrect it three years later when the editor of Acustica wanted something real quick for a special birthday issue. Ironically, the paper became one of my more frequently cited papers (it shows that the complex modulation transfer function is the Fourier transform of the squared impulse response of a linear passive system).

My interest in modulation transfer was originally awakened in 1973 by a paper by R.K. Cook: "Modulated reverberation – A new method for measurement of absorption and sound power". [*J. Acoust. Soc. Amer.* **54**, 302]. Here was a new approach to measuring reverberation time! (Building on this method, I later suggested measuring the modulation spectra of music on the stage and in the audience area during an actual performance to determine the reverberation of a hall in the presence of a live audience – without recourse to "instant" people.)

It gives me great pleasure to introduce the present volume, a fit tribute to Tammo and Herman and their pioneering work on speech intelligibility and the Speech Transmission Index, now generally known as STI.

Chapter 1. The roots of the STI approach

Tammo Houtgast and Herman J.M. Steeneken

1.1. INTRODUCTION

The STI approach, as it is known now, has a long history. It includes concepts like the speech envelope, the envelope spectrum, modulation reduction and the modulation transfer function, all in relation to speech intelligibility. We like to use this opportunity to briefly review the development of our ideas in this field.

It starts in the sixties of the previous century (our first reference on this topic is from 1969), with the work of French and Steinberg[1], and Kryter[2,3] on the Articulation Index readily available, and always playing a role somewhere in the background. Especially the idea that speech intelligibility is based on a weighted sum of contributions from 'autonomous' frequency bands, with the local speech-to-noise ratio as the important parameter, has long been a guiding principle in our work, as in that of numerous others. I think that our most important contribution is that we have broadened the field of application, to account also for other types of disturbances besides additive noise, including reverberation. The key issue here is the influence of these distortions on the speech envelope. Accepting the significance of the speech envelope for speech intelligibility, the speech-envelope spectrum provides a quantitative basis to estimate the effect of any type of disturbance, or combination of disturbances, on speech intelligibility.

How did it start?

1.2. OUR FIRST EFFORT

As young human-factors engineers, we had to determine the quality of various types of radio communication systems, for a great variety of conditions. We used the classical approach, with talkers and listeners, and we got bored. Being aware of the AI-concept, as a calculation scheme to quantify the effect of noise on speech intelligibility, we devised a test signal, to be use instead of the speech. The leading principle already was that (1) we needed a 'speech-like' test signal (i.e., with somewhat similar spectral and temporal characteristics in order to make the system operate under 'normal' conditions), and that (2) modulation reduction was used to quantify the effect of interfering noise. As indicated by Fig. 1 (taken from ref. 4) the test signal was very simplified, with only a rudimental similarity to a speech signal. We used four amplitude-modulated frequency components (10 Hz modulation frequency), at octave intervals. After transmitting and recording this test signal, a quality index was derived from the observed 10-Hz-modulation reductions. By this procedure, we could practically monitor the transmission quality on line, speeding up the work tremendously. However, evaluation of this approach by comparing it with the results of actual speaker-listener intelligibility scores, learned that there certainly was room for improvements! The principle seemed sound, but the implementation was too simple.

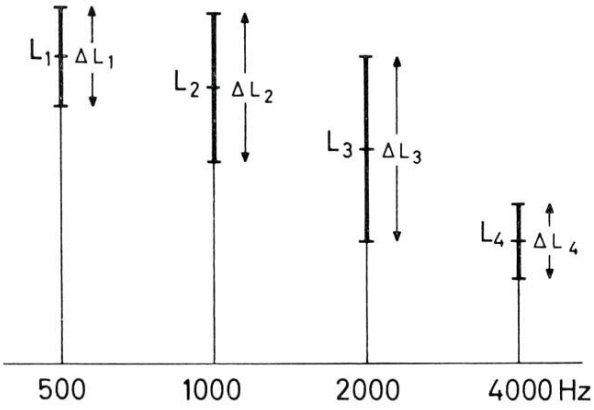


Figure 1. (From ref. 4, 1969) Illustration of the first test signal, with 10-Hz level fluctuations ΔL, used for estimating the quality of radio communication systems. The quality index was derived from the remaining level fluctuations at the output.

1.3. INTRODUCTION OF THE STI

Our first paper[5] in which the term STI was introduced, appeared in Acustica 1971. Interestingly, the manuscript had been submitted to JASA, but was rejected in the review process. When reading it again now, I still consider this a serious mistake. Anyway, we used other publication media, and it was not until 1980 that we considered JASA worthy again of receiving another STI manuscript[10].

The STI approach as described in the 1971 Acustica paper bears much similarity to our original approach, but is more sophisticated. The test signal is speech-shaped noise, rather than a few isolated sine waves, with one fixed ΔL-level-fluctuation rhythm. The quantity of interest is the observed (octave-band specific) reduction in ΔL. Equation (4) taken from that paper (Fig. 2) illustrates the situation at that point in time.

$$STI = \frac{1}{\alpha} \sum_{n=1}^{5} \alpha_n \left(\frac{\Delta L_n'}{\Delta L} \right)^p ,$$

in which: $\alpha = \sum_{n=1}^{5} \alpha_n$.

Parameters: ΔL (dB), initial level difference,
F_r (Hz), alternation rate,
p, power applied in eq. (4),
α_n, octave-band weighting factor.

Figure 2. (From ref. 5, 1971) The first STI formula, as a weighted sum of the contributions from five octave bands, based on the remaining level fluctuations $\Delta L_n'$ at the output.

Note that ΔL is in dB, much different from our later thinking in terms of intensity-modulation reduction *m*. The essence of that paper was that the approach was applied to 50 different transmission channels, including a wide variety of different disturbances, and that the four parameters indicated in Fig. 2 were optimised on the basis of the rank-order correlation between the STI-values and PB-word scores. The disturbances also included a

few conditions with reverberation, and the alternation rate F_r in the test signal is of course a crucial parameter to correctly represent these conditions. The comparison of noise interference conditions and reverberation conditions did lead to an optimal choice of 3 Hz. Interestingly, this is close to the geometrical mean of the relevant modulation-frequency range considered to day, i.e. typically from 0.5 to 16 Hz.

For radio and telephone communication systems, the approach worked quite satisfactory. As we became more and more interested in speech intelligibility in enclosures, with noise, reverberation and echoes as main sources of disturbance, serious shortcomings became obvious too. For other than perfect exponential-decay reverberation, the use of a single alternation rate in the test signal could no longer be maintained. This initiated our thinking in terms of speech-envelope spectrum and modulation transmission function.

1.4. SPEECH ENVELOPE SPECTRUM AND MTF

Of course, it was common knowledge that reverberation affects especially the faster fluctuations in the speech, leaving the slower fluctuations relatively unaffected. To make this more quantitative, we started to study the speech envelope. In line with the critical-band concept in hearing, and the ideas underlying the AI, this should be performed in individual frequency bands, and as a start we applied octave-band filters to the speech before deriving the envelope. A crucial step was the introduction of the envelope spectrum[7], to estimate the amount of fluctuations as a function of fluctuation rate in the envelope. By using traditional $^1/_3$ oct band analysis equipment, with center frequencies from 50 to 5000 Hz, and a 200-fold acceleration of the sampled envelope, the effective range was from 0.25 to 25 Hz. As illustrated in Fig. 3 (from ref. 7), this produced the $^1/_3$ oct spectrum of the modulation level in dB (20log*m*, with *m* being the modulation index).

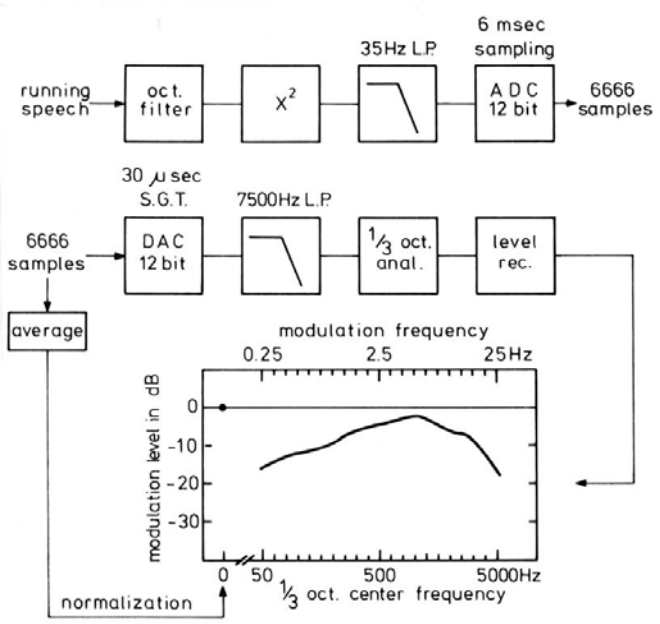


Figure 3. (From ref. 7, 1972) Block diagram of the procedure for estimating the speech-envelope spectrum. Note that by a 200-fold acceleration of the sampled envelope, we could use standard $^1/_3$ oct analysis equipment in the range from 50 to 5000 Hz.

We played along with this a lot[9,12]: what length of a speech sample is required to arrive at a stable envelope spectrum; what about inter-individual differences, or the effect of speaking rate, or systematic octave-band specific differences? A typical example is given in Fig. 4, representing the envelope spectra of five 40-s speech tokens of one talker.

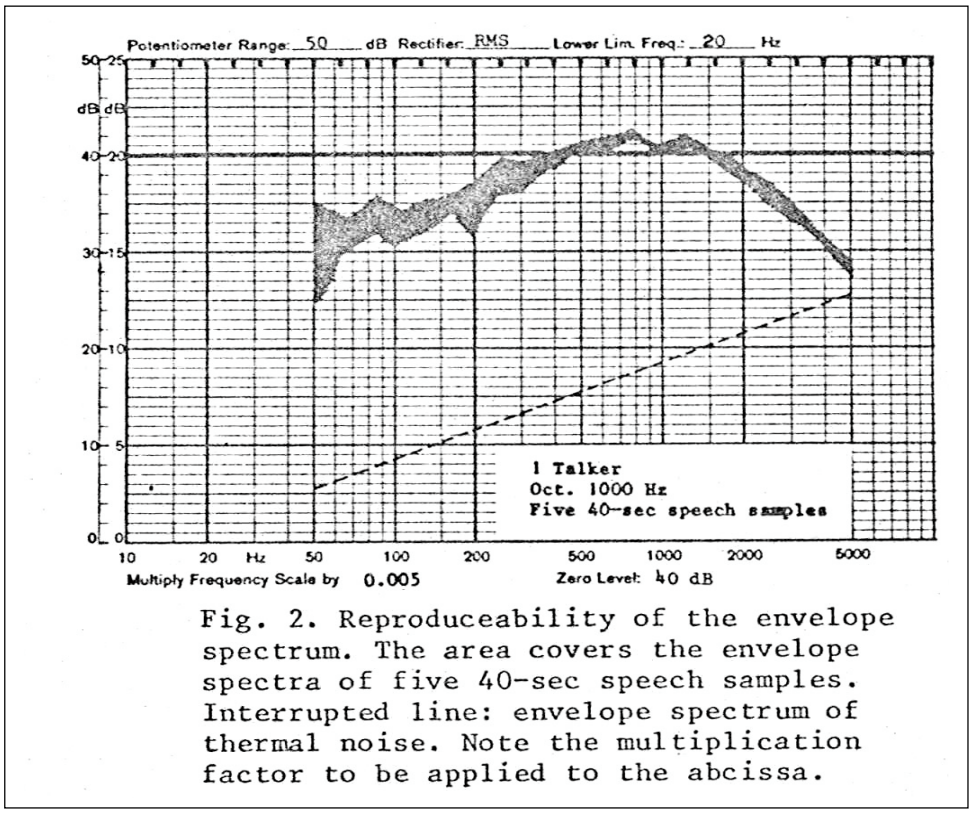


Figure 4. (From ref. 7, 1972) A typical example of some speech-envelope spectra (with original figure caption).

A second crucial step was to consider the effect of various types of disturbances on the envelope spectrum. I consider Fig. 5 (from ref.7) a fine example to illustrate our line of thinking at that point in time. It displays the original envelope spectrum together with the envelope spectra obtained for nine combinations of noise and/or reverberation. Noise produces an over-all reduction of the envelope spectrum, whereas reverberation acts as a low-pass filter on the envelope spectrum. There is a lower limit, indicated by the +3 dB/oct straight line, reflecting the envelope spectrum resulting from the statistical fluctuations of octave-band filtered white noise (thus, again, essentially a white noise spectrum!). The gray area, between the disturbed-speech envelope spectrum and the lower limit, is considered to reflect the amount of 'useful' modulations, and appears to be correlated with the PB-word scores obtained for the various conditions.

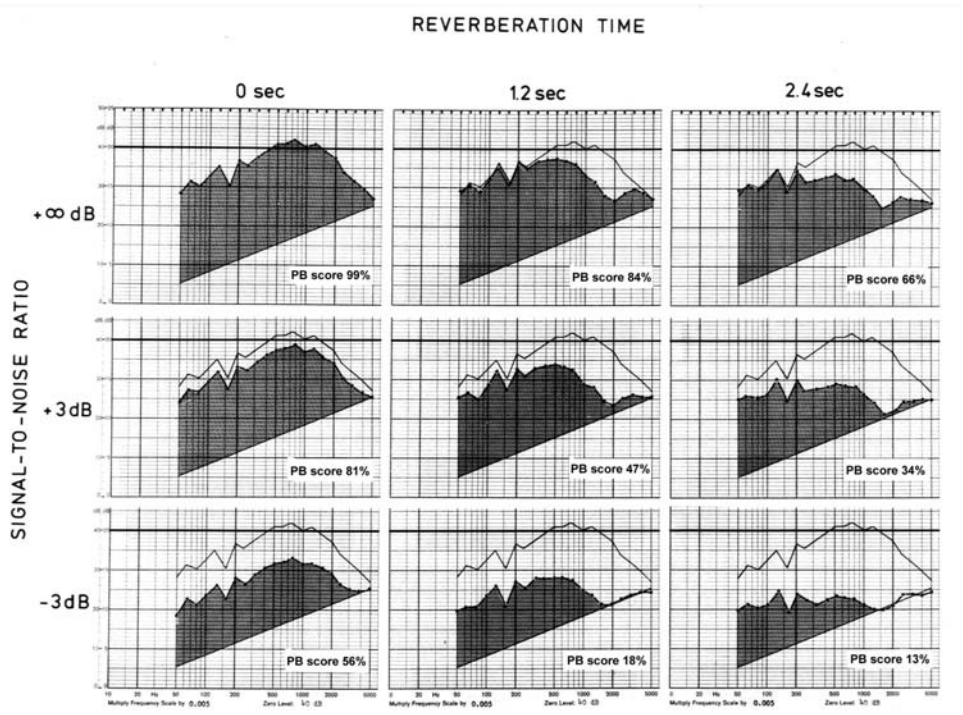


Figure 5. (From ref.7, 1972) Effect of some disturbances on the speech-envelope spectrum. For nine combinations of reverberation and noise, the panels represent the original envelope spectrum (upper curve), the 'disturbed' spectrum, and the lower limit (+3 dB/oct straight line) for noise alone. The areas represent the remaining 'useful' speech modulations, and appear to be related with the PB-word scores.

I would like to add a few notes here. Firstly, the concept of a speech-envelope spectrum was not completely novel at that time. A recent literature search, with the appropriate key words, revealed that the concept was introduced in JASA by Horii, House and Hughes in a 1971-paper[6]. They present a figure illustrating "Spectral analysis of the amplitude envelope of a connected speech sample." It is the amplitude-envelope, rather than the intensity-envelope, and it is broad-band rather than octave-band filtered speech, but nevertheless the concept is there.

Secondly, a few words about the use of $^1/_3$ oct band analysis to represent envelope spectra. Without much thinking, the use of a logarithmic frequency scale was simply copied from the common procedure in the audio-frequency domain. It is interesting to note that the typical shape of the speech-envelope spectrum as we know it now, with its top around 4 Hz, depends on this choice. Would we have used a linear frequency scale, in terms of spectral density, the initial part with its slope of typically +3 dB/oct would become flat, and the shape would be that of a simple low-pass characteristic, with a cut-off around 6 to 8 Hz. Later research on amplitude-modulation detection substantiated the use of a logarithmic frequency scale in the modulation domain[15] (see also section 6).

Thirdly, I want to stress the rationale of our choice for the *intensity*-envelope. Noise and reverberation concern the addition of essentially uncorrelated signals, and this is a linear operation in the intensity-envelope domain only. For instance, if we start with a sine-wave shaped intensity modulation, it remains a (reduced and perhaps phase shifted) sine-wave shaped intensity-modulation under the influence of noise and reverberation. This is not true for any other domain (amplitude, decibel, or other). Also, this implies that the observed

reduction in the speech-envelope spectrum (as in Fig. 5) can be considered an *attenuation filter* acting upon the original envelope spectrum, *irrespective* of the nature of the input signal. That attenuation filter is defined by the distortions (S/N ratio, degree of reverberation), and applies to the envelope spectrum of *any* input signal. This very characteristic underlies the significance of that attenuation filter, which was later named the Modulation Transfer Function. It also implies that, in principle, the MTF can be measured with any input signal, be it actual speech or a specific sine-wave shaped (intensity-!) modulated test signal.

1.5. FURTHER DEVELOPMENTS

The concept of the MTF, as a filter acting in the speech-envelope domain, was introduced in our Acustica 1973 paper[8]. It concentrated on quantifying the effect of reverberation, single echoes and noise on speech intelligibility. We considered 21 modulation frequencies (from 0.25 to 25 Hz, at $^1/_3$ oct intervals), and the MT at each frequency was derived from the modulation reduction observed for a sine-wave shaped intensity modulated test signal [MT=20log(m_reduction)]. Our prime interest was to derive the 'importance function' along the modulation-frequency scale in relation to speech intelligibility (PB-word scores), somewhat analogous to the importance function in the audio-frequency domain. The relation between the weighted MTF and the PB-word score was very good, as illustrated by Fig. 6. Much to our surprise, the optimal relation was obtained for an essentially flat importance function: apparently, all modulation frequencies in that 0.25 – 25 Hz range contribute equally to speech intelligibility. That flat weighting has been applied since then, be it that the scale has been somewhat reduced (at present 0.5 to 16 Hz).

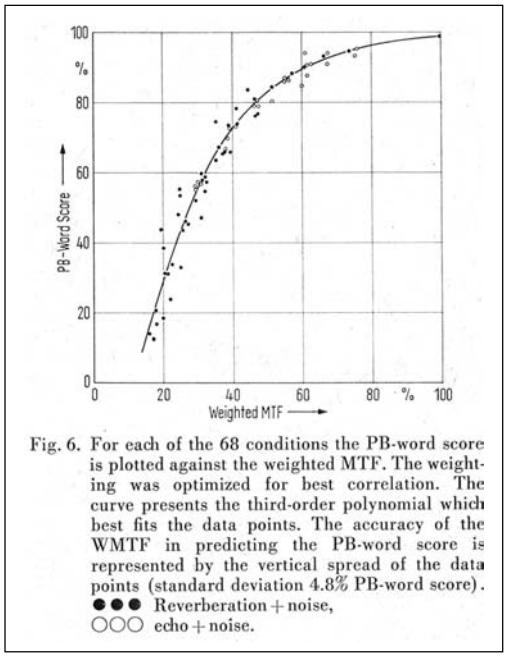


Fig. 6. For each of the 68 conditions the PB-word score is plotted against the weighted MTF. The weighting was optimized for best correlation. The curve presents the third-order polynomial which best fits the data points. The accuracy of the WMTF in predicting the PB-word score is represented by the vertical spread of the data points (standard deviation 4.8% PB-word score).
●●● Reverberation + noise,
○○○ echo + noise.

Figure 6. (From ref. 8, 1973) A typical example of the observed correlation between our index (in this case the Weighted MTF) and the PB-word scores (with original figure caption)

Given this ground work, in the seventies our work proceeded along two lines. One line concentrated on *calculations and predictions*, especially in the field of room acoustics[9,14].

Schroeder's paper[11] contributed much to the general acceptance of the significance of the envelope as the 'information-carrier', and the concept of the MTF acting as a filter on the modulation spectrum. Assuming reverberation with a purely exponential decay defined by one parameter only, the reverberation time T, the modulation reduction can be calculated for any combination of S/N ratio, reverberation, single echoes, also including the contribution of the direct field. Formulae like the one presented in Fig. 7 (effect of T and S/N only, taken from ref. 9) played a key role here, leading to the graph presented in Fig. 8, nicely illustrating the combined effect of T and S/N on the STI.

$$m(F) = \left[1 + \left(2\pi F \frac{T}{13.8}\right)^2\right]^{-1/2} [1 + 10^{(-S/N)/10}]^{-1}$$

Figure 7. (From ref. 9, 1980) Formulae like this one, on the theoretical relation between the modulation reduction m(F) and physical characteristics of the transmission path (in this case the reverberation time T and the S/N ratio), played a key role in the calculations and predictions of speech intelligibility in room acoustics.

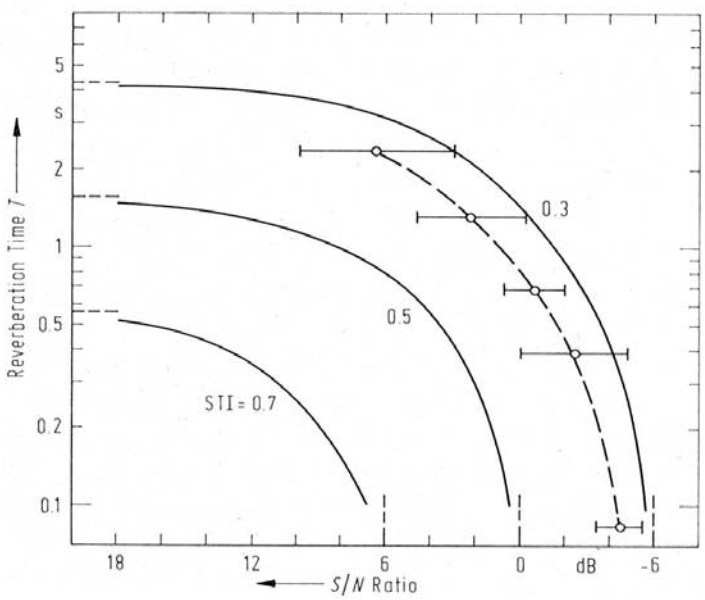


Figure 8. (From ref. 9, 1980) Theoretical equal-STI contours as a function of T and S/N ratio (assuming of course that these are constant over the audio-frequency range). The data points (S/N ratio's for 50% intelligibility of simple sentences at various reverberation times) illustrate the relevance of these contours.

Using basic room acoustics, the STI can be derived from the (design) specifications in terms of a room's volume and reverberation time, its shape (using a ray tracing or mirror-image approach), the ambient noise level, the talker's vocal output and the talker-to-listener distance. This has led to numerous quantitative estimates of, for instance, the effect of the room shape and the distribution of absorption over the walls, the critical talker-to-listener distance and the critical noise level, given a certain amount of reverberation.

The second line concerned the definition and evaluation of *measuring systems*. Depending on the field of application (i.e., telecom systems, room acoustics) we devised several systems, differing in the range of octave band filters and modulation frequencies considered: full-STI, STITEL, RASTI. The JASA 1980 paper[10] provides a detailed description of the STI as a measuring system, along with elaborate data on the relation with PB-word scores for a wide variety of disturbances. A multi-language evaluation of the STI-approach[13], involving laboratories from eleven different countries, greatly supported the broad acceptance of our method. Still, working on improvements appeared to be an ongoing process; even one of the corner stones of the original AI-approach which we adopted from the start on, the concept of autonomous frequency bands contributing to speech intelligibility, had to be reconsidered[16].

1.6. ENVELOPE MODULATION AND THE EAR

Implicitly, thinking in terms of speech-envelope spectrum and modulation-transfer function, we treated the modulation-frequency scale conceptually in much the same way as the traditional audio-frequency scale. Whereas from a physical point of view, the spectral decomposition of the modulations of the speech envelope is a perfectly legal operation, its relevance from a perceptual point of view is not at all obvious. It would require, for instance, that simultaneous modulations at different frequencies are perceptually separated and processed individually.

Given my interest and experience in auditory research, I approached this matter by adapting classical psycho-acoustic principles from the audio-frequency domain: is there frequency selectivity in modulation masking, does Weber's law apply, is there something like a 'critical bandwidth'? A typical result is presented in Fig. 9 (taken from ref. 15), illustrating the modulation-equivalent of pure-tone masking patterns. The results did strongly support our MTF-approach.

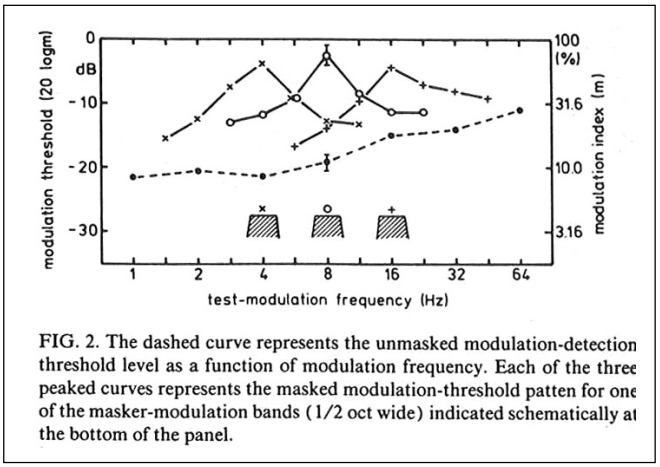


FIG. 2. The dashed curve represents the unmasked modulation-detection threshold level as a function of modulation frequency. Each of the three peaked curves represents the masked modulation-threshold patten for one of the masker-modulation bands (1/2 oct wide) indicated schematically at the bottom of the panel.

Figure 9. (From ref.15, 1989) Psycho-acoustic data on modulation-masking. These data support the conceptual similarity between the modulation-frequency domain and the traditional audio-frequency domain, being an implicit presumption underlying the MTF approach (with original figure caption).

The very existence of perceptual selectivity in the modulation domain, with an effective bandwidth of ½ to 1 oct, provides a rationale for the modulation-frequency analysis (although the grid of $^1/_3$ oct intervals used in the MTF would appear somewhat too fine from

a perceptual point of view). Also, the similarity of the masking-pattern shapes on a log-frequency scale endorses the choice of a logarithmic modulation-frequency scale.

Looking back, I feel that our 'surround', with close interactions between speech research and hearing research in our laboratory, and also with people working in vision where the MTF already was a well known and accepted concept, has strongly contributed to the development of our ideas. For us, it has been an enjoyable and fulfilling experience, and we consider the present status and broad acceptance of the STI as very gratifying.

REFERENCES

1. French, N.R. and Steinberg, J.C. (**1947**). "Factors Governing the Intelligibility of Speech Sounds", J. Acoust. Soc. Am. **19**, 90-119.
2. Kryter, K.D. (**1962**). "Methods for the Calculation and Use of the Articulation Index", J. Acoust. Soc. Am. **34**, 1689-1697.
3. Kryter, K.D. (**1962**). "Validation of the Articulation Index", J. Acoust. Soc. Am. **34**, 1698-1706.
4. Houtgast, T. (**1969**). "The physics of intelligibility" (in Dutch), TNO-Nieuws 24, 296-300.
5. Houtgast, T. and Steeneken, H.J.M. (**1971**). "Evaluation of Speech Transmission Channels by Using Artificial Signals", Acustica **25**, 355-367.
6. Horii, Y., House, A.S. and Hughes, G.W. (**1971**). "A Masking Noise with Speech-Envelope Characteristics for Studying Intelligibility", J. Acoust. Soc. Am. **49**, 1849-1856.
7. Houtgast, T. and Steeneken, H.J.M. (**1972**). "Envelope Spectrum and Intelligibility of Speech in Enclosures", Proceedings IEEE Speech Conference, Newton, MA, 392-395.
8. 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.
9. Houtgast, T., Steeneken, H.J.M. and Plomp, R. (**1980**). "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I. General Room Acoustics", Acustica **46**, 60-72.
10. Steeneken, H.J.M. and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Am. **67**, 318-326).
11. Schroeder, M.R. (**1981**). "Modulation Transfer Functions: Definition and Measurement", Acustica **49**, 179-182.
12. Steeneken, H.J.M. and Houtgast, T. (**1983**). "The temporal envelope spectrum of speech and its significance in room acoustics", Proceedings 11th ICA, Paris, 85-88.
13. Houtgast, T. and Steeneken, H.J.M. (**1984**). "A Multi-Language Evaluation of the RASTI-Method for Estimating Speech Intelligibility in Auditoria", Acustica **54**, 185-199.
14. Houtgast, T. and Steeneken, H.J.M. (**1985**). "A review of the MTF-concept in room acoustics and its use for estimating speech intelligibility in auditoria", J. Acoust. Soc. Am. **77**, 1069-1077.
15. Houtgast, T. (**1989**). "Frequency selectivity in amlitude-modulation detection", J. Acoust. Soc. Am. **85**, 1676-1680.
16. Steeneken, H.J.M. and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility", Speech Communication **28**, 109-123.

Chapter 2. Basics of the STI measuring method

Herman J.M. Steeneken and Tammo Houtgast

PREFACE

In the late sixties, we were asked to perform range measurements for VHF-radio systems. These measurements should make use of (subjective) intelligibility tests. The effort required for this project was enormous. This was due to the number of individual parameters included in the test, but also by the time consuming nature of subjective intelligibility measurements. Therefore, we initiated the use of objective testing in order to *predict* the intelligibility by simple physical measurements. This first step was very much appreciated and resulted into an objective intelligibility measure: the Speech Transmission Index (STI). The measurement of the STI was performed with a simple analogue real-time measuring system (STIDAS-I).

Further developments have led to a robust method that produced an accurate prediction of the intelligibility for many types of transmission channels and in room acoustics: the STI method. This procedure was also realised in a specific measuring device (STIDAS-II, 1978). Twenty-five of these devices, which were based on specific hardware and a PDP 11-03 computer, were in use all over the world.

As a spin-off, a screening device for measurement of the STI in auditoria was developed in 1979. The RASTI method (Room Acoustical Speech Transmission Index) is defined in an former IEC recommendation, IEC 268-16. Several companies built specific hardware for the measurement of RASTI, or incorporated STI-related measures in their own systems.

The accuracy of the STI method has been improved ever since, and has been extended to predict the intelligibility for both male and female speech. The application is not restricted to specific hardware, but has also been implemented in software packages.

The use of the STI-method has grown steadily over the past years. Many standards and recommendations on transmission quality now include the STI procedure (ISO 9921, IEC 60268-16). For transmission quality testing in accordance with these standards, the RASTI system is often (incorrectly) used for assessment of communication systems including deteriorated sound sources, for which the RASTI method was *not* designed. For this purpose, the STI-PA system was recently designed. This system is applicable for public address systems and accounts correctly for the distortions that are related to public address. The test signals are provided on a CD and a specific hand-held analyser performs the analysis.

This overview describes the principles underlying the STI method, and gives a detailed description of the use of the method, the diagnostics, and examples of a number of applications.

2.1. INTRODUCTION

Speech is considered to be the major means of communication between people. In many situations the speech signal we are listening to is degraded, and only a limited transfer

of information is obtained. This may be due to factors related to the speaker, the listener, and the type of speech, but in most situations it is due to limitations imposed by the transmission of the speech signal from the speaker's mouth to the listener's ear. The purpose of the measuring method described in this overview is to quantify these limitations and to identify the physical aspects of a communication channel that are primarily related to the intelligibility of the speech signal passed through such a channel. During transmission, degradation may occur that results in a decrease of the information content[1] of the speech signal such as: limitations of the frequency range, the dynamic range, and distortion components.

All these aspects have been studied in the literature during the past seven decades. This has resulted in design criteria for transmission channels and in the development of speech quality measures, speech intelligibility tests, articulation tests, and a few diagnostic and objective assessment methods. Three methods of assessment can generally be distinguished:

(a) subjective measures making use of speakers and listeners,
(b) predictive measures based on physical parameters,
(c) objective measures obtained by measurements with specific test signals.

Ad (a). Subjective tests make use of various types of speech material. All these tests have their specific advantages and limitations mostly related to the speech items tested. Frequently used speech elements for testing are phonemes, words (digits, alphabet, short words), sentences, and a free conversation in combination with quality rating.

Ad (b). Predictive measures based on physical and perceptual parameters that quantify the effect on the speech signal and the related loss of intelligibility due to for instance: a limited frequency transfer, masking noise, reverberation, echoes, and a non-linear transfer resulting from peak clipping, quantisation, or interruptions.

From the perceptual (listener) point of view, hearing properties, such as frequency resolution, auditory masking, and reception thresholds, also define the intelligibility for a given condition.

One of the first descriptions of a model to predict the effect of a transmission path on the intelligibility of speech was presented by French and Steinberg (1947) and later evaluated by Beranek (1947). This work formed the basis for the so-called Articulation Index (AI), which was described, evaluated and made accessible by Kryter (1962a).

Ad (c). The objective measurement of speech intelligibility has been studied for many years. Specific measuring devices were developed, improvements were made, and the range of applications extended. Therefore, in the next chapter, an overview is given of these developments during the past forty years.

One such objective method to predict the speech transmission quality of an existing communication channel was developed by Houtgast and Steeneken (1971), and Steeneken and Houtgast (1980). This method is based on the application of a specific test signal. The transmission quality is derived from an analysis of the received test signal, and is expressed by an index, the Speech Transmission Index (STI). The STI is based on weighted contribution from a number of frequency bands. For this purpose, the STI uses a fixed bandwidth (octave bands) with a contribution (weighting factor α_k) as indicated in Fig. 1.

[1] Information content: properties of a speech signal that contribute to identification of a speech item (phoneme, word, or sentence).

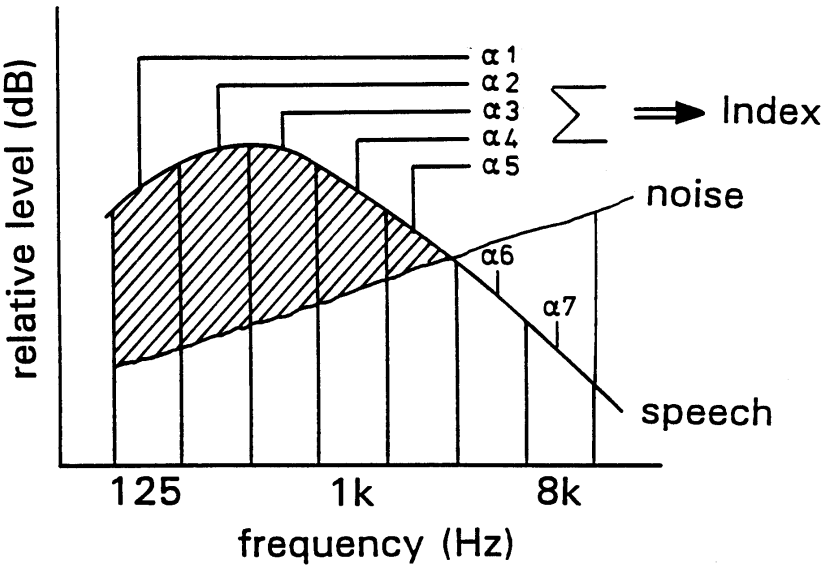


Figure 1. Illustration of the long-term spectrum of a speech signal masked by noise, and the weighted summation to an objective intelligibility prediction.

The STI value is obtained from measurements on the transmission channel in operation, or based on a calculation scheme making use of physical properties of the transmission channel. The STI measurement requires a special test signal from which the effective signal-to-noise ratio in each octave band at the receiving side is determined and used for the calculation of the STI. The specific features of this approach are that the test signal design allows an adequate interpretation of many degradations than just a limited frequency transfer and masking noise, for example non-linear distortion and distortion in the time domain. Hence, almost all types of distortion and their combinations that may occur in an analogue or digital (waveform-based) transmission path are accounted for. However, distortions such as frequency shifts and voiced/unvoiced decision errors that may occur with certain types of vocoders, are not included in this concept. Up to 1993 the measurement of the STI for telecommunication channel evaluation made use of a specific measuring device (Steeneken and Agterhuis, 1982). Over the years, twenty-five of these devices have been built and have been distributed to many laboratories all over the world.

Fifteen years of experience with the development and the application of the STI have shown the need for further improvements, for instance when applied to conditions with a very limited frequency transfer or non-contiguous frequency transfer. Also, effects of speaker variation, the gender of the speaker, and the individual relation with consonant and vowel recognition required further attention.

We were able to improve the STI model and extend the model with respect to male/female speech, the type of speech being assessed, and speaker variations. Steeneken (1992) describes the results of this study.

2.2. OVERVIEW OF OBJECTIVE MEASURING METHODS FOR PREDICTING SPEECH INTELLIGIBILITY

The first description of the use of a "computational method for the prediction of the intelligibility of speech and its implementation in an objective measuring device" was given by Licklider et al. (1959). They described a system that 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. A spectral-weighted contribution of the similarity between temporal envelopes of the speech signals at the input and at the output of a transmission channel is used for the computation of the PCI. A total of 15 minutes of speech was required for this analysis. The paper reports that the results of a comparison between the PCI and human listener evaluation show a monotonic relation for conditions with an increasing effect of one type of distortion. Contributions of different types of distortion show a "sufficient agreement." Schwarzlander (1959) described the electronic design of the system. Licklider proposed an improvement of the PCI by making use of synthetic signals, physically related to average speech, and with a duration of about one second for the total measurement of the PCI.

Five years later Kryter and Ball (1964) described a system called the Speech Communication Index Meter (SCIM), which was based on the AI as described by Kryter (1963). The measurements were mainly concentrated on deriving the signal-to-noise ratio within a frequency range of 100–7000 Hz and a dynamic range of 30 dB. The auditory masking corrections according to the AI concept were also included. An evaluation of the system was performed for several types of transmission conditions, including low-pass filtering, noise, frequency shifts, and clipping.

In 1970 we developed a system based on the use of an artificial test signal which was transmitted over the channel-under-test and which was analysed at the output. The test signal was an amplitude-modulated noise signal with a square-wave envelope. Hence the signal level alternated between two values. The difference between these two levels was 20 dB and the switching rate was 3 Hz (Houtgast and Steeneken, 1971). The noise carrier had a frequency spectrum corresponding to the long-term speech spectrum. This was the first approach in which speech-related phenomena, concerning spectral variations and temporal variations, were included in an artificial test signal. The essential point of this approach was that the resulting level variation at the output of a communication system reflects the signal-to-noise ratio, providing a basis for subsequent calculations according to the AI concept. The method was based on measurements in five octave bands (centre frequencies 250 Hz – 4 kHz). The effect of band-pass limiting, noise, peak clipping, and reverberation on intelligibility was included in the test signal concept and in the evaluation procedure. This resulted in an index ranging from 0 – 1, the so-called Speech Trans-mission Index (STI). A measuring device was developed, based (at that time) on analogue circuits, which could determine the STI within 10 s. It should be noted that this method is different from the STI approach published later, which is described in section 3 of this chapter.

The next step was to use a test signal with various modulation frequencies instead of the fixed (3 Hz) square-wave modulation signal. This modulated test signal was based on the measurement of the fluctuations of the envelope of connected discourse (Houtgast and Steeneken, 1971). The envelope fluctuations were determined for separate frequency bands (octave bands). While the envelope function is unique for a certain combination of successive speech sounds, the frequency spectrum of the envelope fluctuations, called the "envelope spectrum" proved to be a stable and reproducible characteristic of running speech (for speech tokens of at least 10 s, see Fig. 2). This envelope spectrum (with a frequency range from about 0.2 Hz to 12.5 Hz) was measured in $^1/_3$-octave bands and normalised with respect to the mean level (intensity).

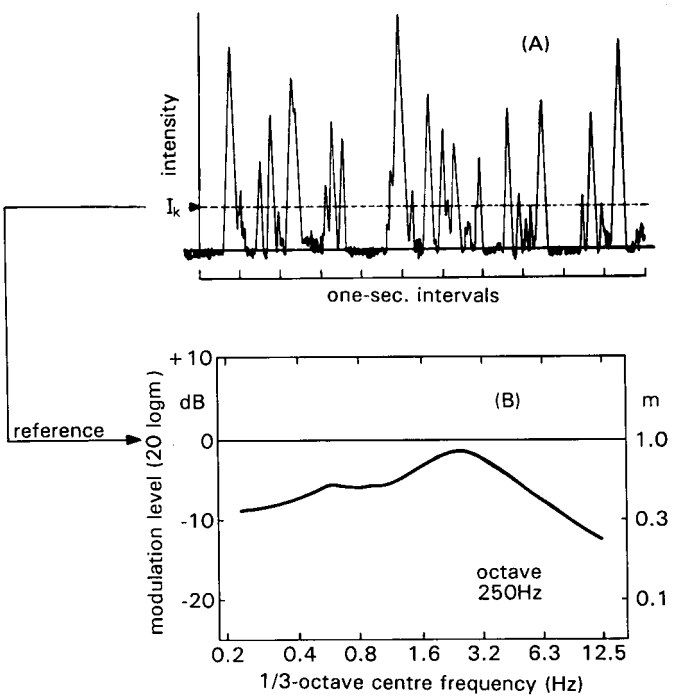


Figure 2. Envelope function (panel A) of a 10s speech signal filtered for the octave band-width centre frequency 250 Hz. The corresponding envelope spectrum (panel B) is normalised with respect to the mean signal intensity (I_k).

The transfer of these fluctuations of speech by a communication channel can be obtained by comparing the envelope spectra of the same speech signal at the input and at the output of the channel under test (Steeneken and Houtgast, 1973). For that purpose a 60-second segment of natural speech can be used as the test signal. The effect of noise on the envelope spectrum of speech is independent of the fluctuation frequency, however, this is not the case for distortions in the time domain. Reverberation will act as a low-pass filter for fluctuations and can be predicted for an exponential decay. Since there is a simple relation between the relative decrease of the fluctuations and the actual signal-to-noise ratio, this relation can be used to measure the effective signal-to-noise ratio as a function of fluctuation frequency.

Next to the use of natural speech as a test signal, Houtgast and Steeneken (1972) also proposed the use of an artificial test signal, where each relevant fluctuation frequency was tested separately. This resulted in the so-called Modulation Transfer Function (MTF). The (octave-band specific) MTF represents the transfer of the (octave-band specific) envelope of a signal between the input and output of a transmission channel.

The method was extensively evaluated for conditions with noise, reverberation, and echoes. The analysis and the generation of the echo conditions, at that time, were performed with a digital (PDP-7) computer, a system with a 1.75 μs cycle time and 8K-words of memory!

Payne and McManamon (1973) introduced the Speech Quality Measure (SQM) for communication channels. This system was based on the AI concept. The authors mentioned limitations for digital encoding, fading, and non-linear distortion. They remarked "when using the system it should be checked to have none of these distortions present". The test signal

was based on 20 tones with frequencies at the mid-point of the 20 frequency bands with "equal contribution to intelligibility" as used for the original AI concept. The paper also proposes the use of mini-computers to perform the analysis and to display the results. No validation was reported.

Steeneken and Houtgast (1980) extended the MTF approach (that had already been validated for channels with noise, echoes, and reverberation) to channels with distortions more specific for communication channels, namely band-pass limiting, noise, non-linear distortion, quantisation errors from digital coders, and reverberation.

Schroeder (1981) developed a mathematical background of the MTF referred to as CMTF. This function is more generic as it also includes the phase transfer. However, this parameter is not used for the STI.

Based on the STI concept, the RASTI method (Room Acoustical Speech Transmission Index) was developed in 1979 (Steeneken and Houtgast, 1979; Houtgast and Steeneken, 1984). This simplified method was especially developed as a screening device for applications in room acoustics and restricted to person-to-person communications. The method was standardised in 1988 by IEC 268-16. Notice that the effect of PA-systems on the frequency transfer and possible not linear distortion was not accounted for.

Quackenbush et al. (1988) gave an overview of "Objective measures of speech quality" especially applied to digital coders. They also evaluated some objective measures, which were mainly based on signal-to-noise ratios.

A major improvement of the STI method, in use since 1980, was achieved in 1992. The additive model on which the AI and STI were based was extended with a so-called redundancy correction. This correction accounts for the correlation of the information content within two adjacent frequency bands of a speech signal. This is essential for systems with a very limited frequency transfer (PA systems) and a discontinuous frequency transfer.

Also, various extensions were added to the STI measuring procedure such as a separate assessment of male and female speech, the type of speech material used for the prediction of the intelligibility, and a model for the prediction of speaker variations. The results of this study are described by Steeneken (1992) and by Steeneken and Houtgast (1999, 2002a, 2002b).

2.3. MEASUREMENT AND CALCULATION OF THE STI

2.3.1. Description of the algorithm

The STI is an objective measure, based on the contribution of a number of frequency bands within the frequency range of speech signals, the contribution being determined by the effective signal-to-noise ratio. This signal-to-noise ratio is called effective because it may be determined by several factors. The most obvious one is background noise, which contributes directly to the signal-to-noise ratio. However, products of distortions in the time domain and non-linearity's are also considered as noise. This is derived by the specific design of the test signal. In Fig. 3, an illustration is given of the estimation of the signal-to-noise ratio within each frequency band. The test signal consists of a noise signal with a frequency spectrum equal to the long-term frequency spectrum of the speech signal. Each octave-band is modulated with a periodic signal in such a way that the *intensity envelope*[2] is modulated

[2] The addition of uncorrelated signals (echoes, reverberation, and masking noises) is based on intensity summation. For instance, the addition of two sinusoidal modulated signals (same modulation frequency) with uncorrelated carriers will consist of a signal with a sinusoidal envelope modulation being the vector summation of the sinusoidal envelope of the two primary signals. This statement is only valid for intensity modulations, and not for amplitude modulations.

sinusoidal. This is indicated in Fig. 3 for the octave band with centre frequency 250 Hz. The modulation index (*m*) in this example is $m = 1$ at the input side and reduced to $m = 0.5$ at the output side.

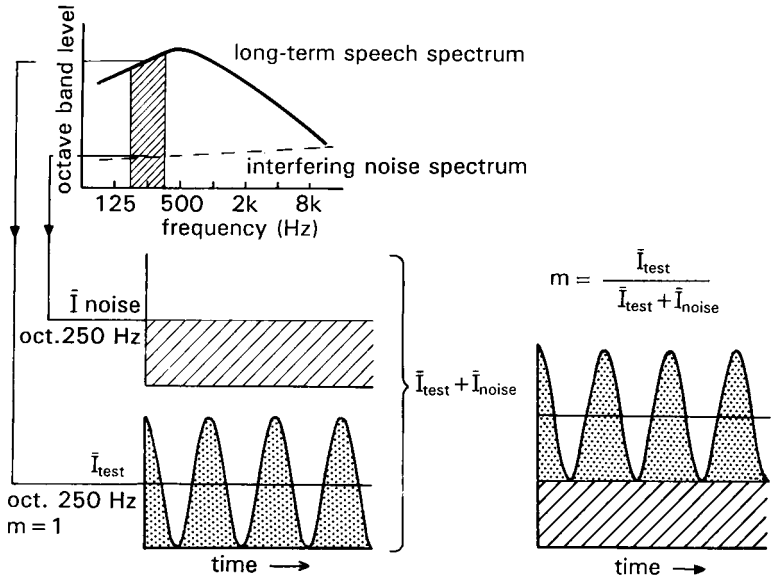


Figure 3. Illustration of the effect of interfering noise on the modulation index *m* of a test signal.

Noise may be added to the test signal and the resulting envelope is obtained by addition of the intensity of both signal envelopes. Hence, the resulting envelope of this example is defined by a steady noise envelope (of a stationary noise signal) and the test signal envelope. The resulting modulation index (m), being the test signal intensity divided by the total intensity (test signal and noise), is directly related to the signal-to-noise ratio (SNR) according to:

$$\mathrm{SNR} = 10 \log \frac{m}{1 - m} \mathrm{dB} \qquad (1)$$

As described in section 2 of this chapter, the envelope function of a fluctuating speech signal contains a range of frequencies, representing the succession of speech events from the shortest speech items (such as plosives) up to words and sentences. Due to distortion in the time domain (reverberation, echoes, and automatic gain control) this fluctuation pattern may be affected, in this way reducing intelligibility. This is modelled in the STI procedure by determining the modulation transfer function for the range of relevant frequencies present in the envelope of natural speech signals. As described before (Steeneken and Houtgast, 1980) a relevant range for these modulation frequencies extends from 0.63 Hz up to 12.5 Hz. Separation in $^1/_3$-octave steps, yields 14 bands. This results in a measuring procedure according to Fig. 4, where the modulation transfer index, m, for each octave band (125 Hz – 8 kHz) and each modulation frequency (0.63 – 12.5 Hz) is determined separately. The figure gives the measuring set-up for one octave band. A noise signal with the required frequency spectrum (normally the long-term speech spectrum) is amplitude modulated by a signal $\sqrt{\{1+\cos(2\pi.f_m.t)\}}$ which results in a sinusoidal intensity modulation $I^*\{1+\cos(2\pi.f_m.t).\}$This

modulation function can be obtained digitally and can be generated by computer. At the receiving side, octave-band filtering and (intensity) envelope detection is applied. From the resulting envelope function a Fourier analysis determines the modulation index reduction, due to the reduction by the transmission channel. This procedure is repeated for each cell of the matrix given in Fig. 4. It should be noted that the block diagram of Fig. 4 represents only one channel corresponding with one octave band. The original set-up consists of a set of separate channels for all octave bands considered.

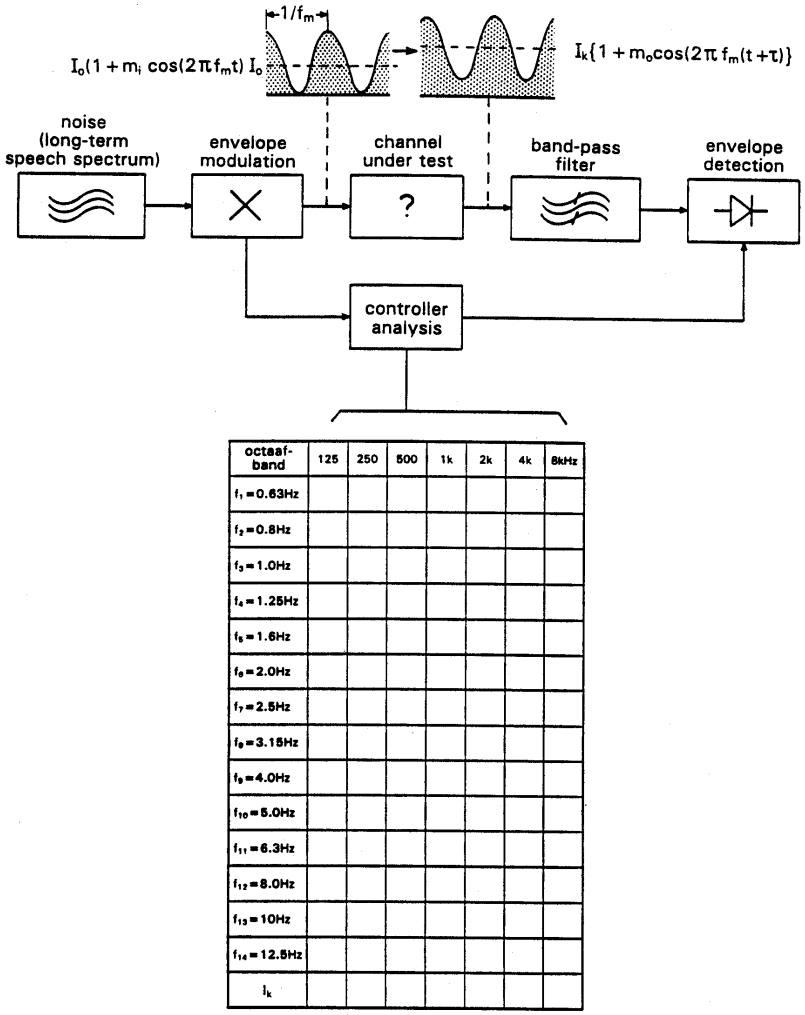


Figure 4. General block diagram of the measuring set-up. The modulation index reduction at the output (m) is determined for all cells of the matrix (7 octave bands and 14 modulation frequencies). Also the octave levels (I_k) are obtained, for calculation of the auditory spread of masking.

With the test signal as described above, distortions such as band-pass limiting, and noise masking, as well as distortion in the time domain can be dealt with. Non-linear distortions, however, have to be modelled additionally. If a speech signal is passed through a system with a non-linear transfer (e.g. peak clipping or quantisation), harmonic distortion components

and inter-modulation components will be produced in other frequency bands. For this reason the test signal should not be modulated with one and the same modulation frequency for all octave bands simultaneously. Otherwise, non-linear distortion components cannot be discriminated from the modulated test signal in the frequency band considered. Therefore, in the case of non-linear distortion, all frequency bands, except the one under test, are modulated with uncorrelated signals so that the envelopes of the distortion components are not correlated with the test signal envelope in the octave band under test. Such distortion components are then considered as noise (they add to the noise in the octave band under test) and reduce the effective signal-to-noise ratio in a similar way as would occur with other interfering signals. The relative levels of the test signal in the octave bands with the uncorrelated (speech-like) envelope were adjusted for optimal prediction of intelligibility in non-linear transfer conditions. The consequence of this procedure is a successive measurement for each of the seven octave bands rather than a simultaneous measurement as can be applied for communication channels with a linear transfer.

Besides the masking introduced by the noise in the transmission channel two other factors have to be taken into account: (1) an additional auditory masking phenomenon[3] (auditory spread of masking) and (2) the absolute hearing threshold. Both effects are modelled as an imaginary masking noise that leads to a decrease of the effective signal-to-noise ratio. Hence, resulting in a reduction of the modulation transfer index m. For this purpose not only the modulation transfer has to be determined but also the signal levels in the frequency bands have to be considered. In Fig. 5, the effect of the masking by frequency band (k-1) upon frequency band k is indicated for a signal level of 60 dB SPL. The masking as a function of the signal level is given in Table I.

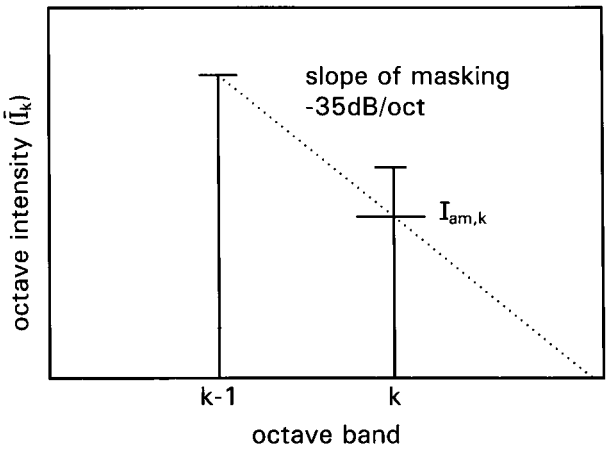


Figure 5. Auditory masking of octave band k-1 upon the next higher octave band k. The slope of the masking effect versus frequency band corresponds to –35 dB/oct. This is equivalent to an auditory masking factor of amf = 0.000316.

The masking effect, as modelled in the STI approach, does not depend on the frequency band considered but does depend on the level. For example, the slope of masking decreases with 35 dB/oct for signal levels between 55 and 65 dB. (In the original STI

[3] Auditory spread of masking is the effect, introduced by the hearing organ, that a strong masker in a lower frequency range may reduce the perception of a tone or narrow-band signal. The amount of masking depends on the level difference between masker and masked signal, on the absolute level of the masker, and on their frequency distance. Zwicker and Feldtkeller (1967) give a detailed description.

approach, the slope of masking was *not* level dependent and fixed on –35 dB/oct. See van Wijngaarden and Steeneken, 1999). The corresponding auditory masking factor (amf) of the intensity of the primary masking signal amounts amf = 0.000316 (intensity attenuation of masking signal upon adjacent next higher octave band). As the masking effect by only one lower frequency band is considered, the intensity of the masking signal becomes:

$$I_{am,k} = I_{k-1} * amf \tag{2}$$

where $I_{am,k}$ represents the intensity level of the auditory masking signal for octave band k, and I_{k-1} represents the signal intensity of octave band (k-1).

In Table I, the slope of the masking as a function of the octave level is given.

Table I. Octave level specific slope of masking

Octave level dB	*46–55*	*56–65*	*66–75*	*76–85*	*86–95*	*>95*
Slope of masking	–40	–35	–25	–20	–15	–10
Auditory masking factor	0.000100	0.000316	0.003162	0.010000	0.031622	0.100000

The effect of the absolute hearing threshold is modelled in the STI approach as the lower limit of the masking noise level within each octave band ($I_{rs,k}$, see Table II). This level is only relevant if I_k refers to the presentation level to the listeners.
The auditory spread of masking and the hearing threshold are accounted for by a reduction in the modulation index. The corrected modulation index becomes:

$$m'_{k,f} = m_{k,f} \frac{I_k}{I_k + I_{am,k} + I_{rs,k}} \tag{3}$$

where $m_{k,f}$ represents the modulation index for octave band k and modulation frequency f, and m' the corrected modulation index.

The effective signal-to-noise ratio for octave band k and modulation frequency f then becomes:

$$SNR_{k,f} = 10 \log \frac{m'_{k,f}}{1 - m'_{k,f}} \quad dB \tag{4}$$

According to the STI concept a signal-to-noise ratio between –15 dB and 15 dB is linearly related to a contribution to intelligibility of between 0 and 1. Therefore, the effective signal-to-noise ratio is converted to transmission index ($TI_{k,f}$), specific for octave band (k) and modulation frequency (f), by the equation:

$$TI_{k,f} = \frac{SNR_{k,f} + shift}{range}, \text{ where } 0<TI_{k,f}<1.0 \tag{5}$$

The shift equals 15 dB and the range equals 30 dB. In this way a relation between the effective signal-to-noise ratio and the TI is obtained as shown in Fig. 6.

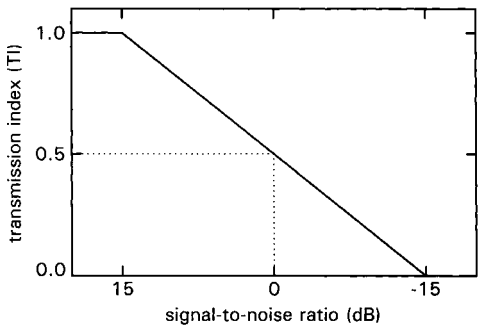


Figure 6. Relation between the effective signal-to-noise ratio and the transmission index for a shift of 15 dB and a range of 30 dB.

All 14 transmission indices related to modulation frequencies between 0.63 and 12.5 Hz[4] , are obtained for each octave band. The mean of these indices results in the modulation transfer index (MTI_k) and is specific for the contribution of octave band k. The MTI_k is given by:

$$MTI_k = \frac{1}{14}\sum_{f=1}^{14} TI_{k,f} \tag{6}$$

Finally, according to the revised formula (IEC 60268-16 2nd edition, 1998), the STI_r is obtained by a weighted summation of the modulation transfer indices for all seven octave bands and the corresponding redundancy correction. This is given by:

$$STI_r = \alpha_1 \cdot MTI_1 - \beta_1 \cdot \sqrt{(MTI_1 \cdot MTI_2)} + \alpha_2 \cdot MTI_2 - \beta_2 \cdot \sqrt{(MTI_2 \cdot MTI_3)} + \ldots + \alpha_7 \cdot MTI_7 \tag{7}$$

where

$$\sum_{k=1}^{7} \alpha_k - \sum_{k=1}^{6} \beta_k = 1 \tag{8}$$

The factor α_k represents the octave-weighting factor and β_k the so-called redundancy correction factor. This redundancy correction is related to the contribution of adjacent frequency bands. Steeneken and Houtgast (1999, 2002a, 2002b) describe the optimal weighting factors and redundancy factors for male and female speech and different groups of phonemes.

In Table II, the α and β values are given for male and female speech, as well as the level of the reception threshold (Eq. 3), which is given in decibel units. The reception threshold represents the absolute hearing threshold, increased with a correction for the dynamics of the speech signal. A flow diagram of the calculation procedure of the STI is given in Fig. 7.

[4] This range provides an optimal fit for conditions with temporal distortions in relation to conditions with noise distortion.

Table II. STIr octave-band specific male and female weighting factors and the absolute reception threshold in decibel.

Octave band (Hz)		*125*	*250*	*500*	*1k*	*2k*	*4k*	*8k*
Males	α	0,085	0,127	0,230	0,233	0,309	0,224	0,173
	β	0,085	0,078	0,065	0,011	0,047	0,095	–
Females	α	–	0,117	0,223	0,216	0,328	0,250	0,194
	β	–	0,099	0,066	0,062	0,025	0,076	–
Absolute reception threshold (dB)	$L_{rs,k}$	46	27	12	6,5	7,5	8	12

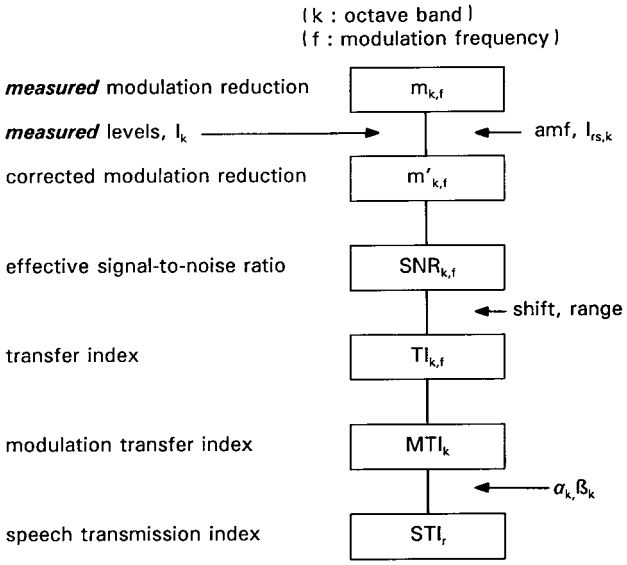


Figure 7. Flow diagram of the STI calculation scheme.

Some simplifications of the procedure described above were made in order to decrease the measuring time, but these simplifications restrict the range of applicability. The measurement of a complete matrix of 98 *m*-values according to Fig. 4 and a measuring time for each *m*-value of 10 s results in a total measuring time of 15 minutes.

A reduction of the 14 modulation frequencies to only three modulation frequencies results in a total measuring time of less than 4 minutes, but as a consequence no complete modulation transfer is obtained. This means that distortions in the time domain are not accounted for correctly. Therefore, this method is normally used only for communication channels with no degradation due to echoes or reverberation such as with person-to-person communication.

Also, the number of octave bands considered may be reduced. This is the case with the RASTI method, where only the contributions of the modulation transfer for the octave bands with centre frequencies 500 Hz and 2 kHz are considered. This can be used as a screening approach for direct person-to-person applications.

Another simplification can be applied to the test signal if the uncorrelated (speech-like) modulations, required for the correct interpretation of non-linear distortions, are omitted. This opens the possibility of applying a simultaneous modulation and parallel processing of

all frequency bands, thus decreasing the measuring time. This procedure is used in the STITEL and STIPA method that requires a measuring time of about 15 s.

It should be noted that the STI method can be applied to transmission channels with the type of distortions listed before. Due to the specific compilation of the test signals and the type of analysis some types of distortions are not accounted for. These are[5]:

- frequency shifts (such as obtained with single side-band transmission),
- frequency multiplication (such as obtained with analogue tape recorders which run at an incorrect tape-speed), and
- vocoders (systems which introduce errors related to voiced-unvoiced speech fragments and pitch errors).

2.4. OVERVIEW OF METHODS, TEST SIGNALS, AND CALCULATION CONSTANTS

The 'full' STI method includes measurements within seven octave bands and 14 modulation frequencies within each octave band. However, certain applications do no require such a robust measuring scheme. For those measurements specific simplifications of the measuring method can be applied in order to increase the measuring efficiency. The various simplifications of the measuring procedure have led to different measuring schemes that are adapted for specific groups of applications. Respective versions are:

STI-14: A universal measuring scheme, which is applicable to all types of communication systems (except vocoders), includes a successive measurement of the full matrix as given in Fig. 4. This method is called STI-14 and refers to the original. For this method test signals for seven octave bands and 14 modulation frequencies are transmitted and analysed successively.

STI-3: As the STI-14 method is time consuming a limitation in the modulation frequency domain is applied in order to decrease the measuring time. This version, based on three modulation frequencies, has limited applicability with respect to conditions with distortions in the time domain (the resolution is decreased). The measuring method is referred to as STI-3.

STITEL: The STITEL (Speech Transmission Index for TELecommunication channels) is a stripped version of the STI and has no robust coverage for transmission channels with distortion in the time domain and for non-linear systems.

STIPA: The STIPA (Speech Transmission Index for Public Address systems) is a stripped version of the STI-14 and has a robust coverage for distortions in the time domain and limitations in the frequency domain. A limited coverage of non-linear distortions is obtained.

RASTI: The RASTI system (Room Acoustical[6] Speech Transmission Index) is based on the MTF for only two octave bands, no coverage for band-pass limiting and non-contiguous noise spectra is obtained. This method is developed for person-to-person communications in a room acoustical environment and does account for distortion in the time domain.

---

[5] Note by the editor: also see chapter 4 of this book

[6] Sometimes referred to as RApid Speech Transmission Index.

An overview of these methods is given in Table III. The field of application is also indicated. For some programs the applicability is condition dependent (e.g. the type of non-linear distortion or the type of reverberation). This means that a test with the STI-14 or STI-3 has to be performed in order to verify the applicability.

Table III. Overview of the measuring procedures, the applications, and the corresponding test signals.

Application	*Band-pass limiting*	*Non Linear Distortion*	*Reverberation Echoes*	*Test signal types*	*Measuring time*
STI-14 (7 octaves, 14 f_{mod})	yes	yes	yes	male, female	15 min
STI-3 (7 octaves, 3 f_{mod})	yes	yes	condition dependent	male, female	4 min
STITEL (7 octaves, 7 oct. related f_{mod})	yes	condition dependent	condition dependent	male, female, original, phoneme groups	15 s
STIPA 7 octaves, 14 oct. related f_{mod})	yes	condition dependent	yes	male, female	15 s
RASTI (2 octaves, 4-5 f_{mod})	no	no	yes	original	15 s

The frequency weighting and redundancy correction factors are identical for the STI-14, STI-3, STITEL and STIPA method but different for male and female speech. For the RASTI only two octave bands are used (500 Hz and 2 kHz).

2.5. INTERPRETATION OF THE STI VALUE: RELATION WITH SUBJECTIVE MEASURES

The use of the STI-method for more than 30 years, the international application, and the validation in other studies (Houtgast and Steeneken, 1984; Anderson and Kalb, 1987; Barnett, 1999; Mapp, 2001; van Wijngaarden and Steeneken, 1999) has led to a robust qualification of the STI value in terms of speech intelligibility. The validation of the method with different intelligibility tests resulted into a robust relation with a variety of subjective measures. In Fig. 8, this relation for the original STI concept and various intelligibility measures is given. It should be noted that the earlier experiments were designed to establish the optimal relation for CVC words of the type "phonetically balanced" for Dutch nonsense words. In later studies CVC words with a uniform phoneme distribution were used. This introduced a slightly different relation between STI and CVC-word score. All the data in this manual refer to CVC-word lists with such uniform (equally balanced) phoneme distribution and nonsense words.

The improvement of the STI method by the introduction of the redundancy corrections resulted in essentially the same relation between the CVC-word score and the STI. However,

the STI values obtained according to the new method are referred to as STI_r. The improvement becomes apparent mainly when transmission channels with severe band-pass limitation, non-contiguous frequency transfer or masking noise with a discontinuous spectrum are tested. In Fig. 8, the relation between the STI_r, the CVC-word score, and sentence intelligibility (short simple sentences) is given for male speech. Additionally the relation between the STI_r and the CVC-word score for female speech is given in Fig. 10.

The relation between the STI_r, the CVC-word score and phoneme group scores can also be derived from the expressions given in Table IV.

$$\text{predicted score} = \{A * e^{(B*STI)} + C\}*100 \ \ (\%) \tag{9}$$

Table IV. Relation between the STI_r, the CVC-word score, and phoneme-group scores for male and female speech.

Word or phoneme type	Male			Female		
	A	B	C	A	B	C
CVC words	-1.5301	-2.0	1.15	-1.7584	-1.5	1.37
Fricatives	-0.9000	-4.2	0.90	-0.9466	-4.1	0.90
Plosives	-1.1531	-4.1	1.01	-1.1256	-6.0	0.95
Vowel-like consonants	-1.4602	-4.2	1.05	-1.3216	-4.0	1.09
Vowels	-0.9976	-2.9	1.03	-1.2057	-3.1	1.04

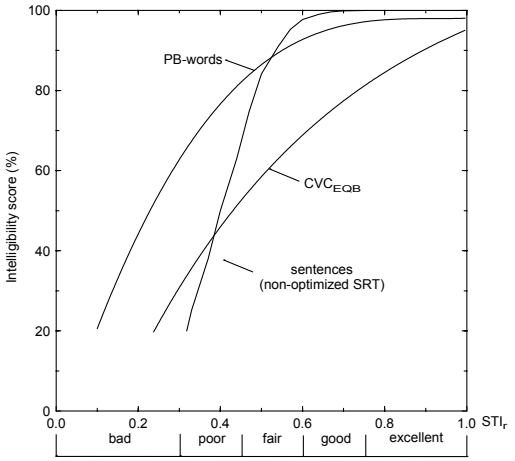


Figure 8. Qualification of the STI_r (Steeneken and Houtgast, 2002b) and relation with various subjective intelligibility measures for MALE speech.

As indicated before, the STI method can also be used to predict the intelligibility scores for certain phoneme groups. For this purpose specific test signals (corresponding to the mean phoneme-group spectrum) and frequency weightings and redundancy correction factors are used. It should be noted that this method couldn't be used for all types of channels. Specifically channels with a "memory" (such as reverberation or automatic gain control) are affected by the level of embedded signals that may interact with the various (phoneme-group-specific) test-signal levels. The relation between the phoneme-group specific STI (referred to as STI_s) and the phoneme-group score is given in Figs 9 and 10 (for

male and female speech, respectively). The equations for calculating the various phoneme-group scores are given in Table IV.

Besides a direct estimation of the CVC-word score this score can also be predicted by combining phoneme-group scores obtained from the STI_s values for the fricatives, plosives, vowel-like consonants, and vowels. This is performed in two steps: (1) calculation of the initial consonant and final consonant score (a weighted combination of the plosive, fricative, and vowel-like consonants scores), and (2) calculation of the CVC-word score from the product of the initial consonant, vowel, and final consonant probabilities.

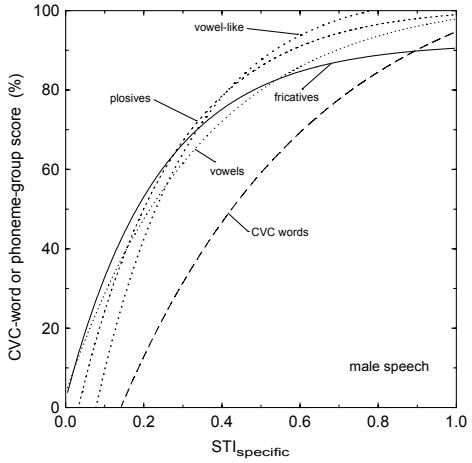


Figure 9. Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for MALE speech. The relation for the CVC-word score is also given.

The advantage of predicting the word score by a (weighted) combination of the predicted phoneme-group scores is that it is not restricted to the example with the equally balanced CVC words, but that it can also be used to predict the word score of PB-words or any other combination. The restriction is, however, that the word score is indeed defined by independent phoneme scores.

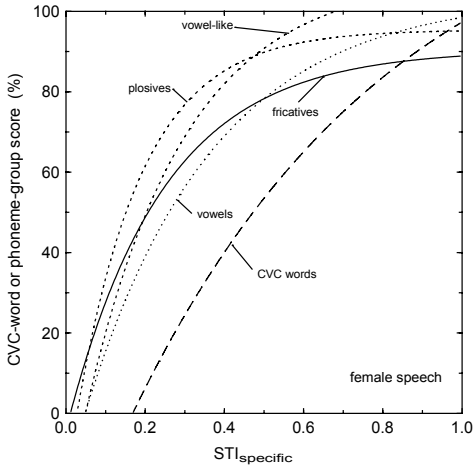


Figure 10. Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for FEMALE speech. The relation for the CVC-word score is also given.

The relative test signal spectra for phoneme groups and the embedded CVC test words are given for males and females in Figs. 11 and 12.

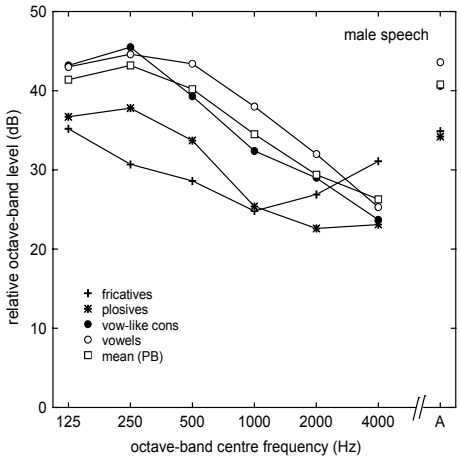


Figure 11. Relative test signal spectra for the four phoneme groups and for phonetically balanced speech (connected discourse). The dBA values represent the relative level of each group for connected discourse of MALES.

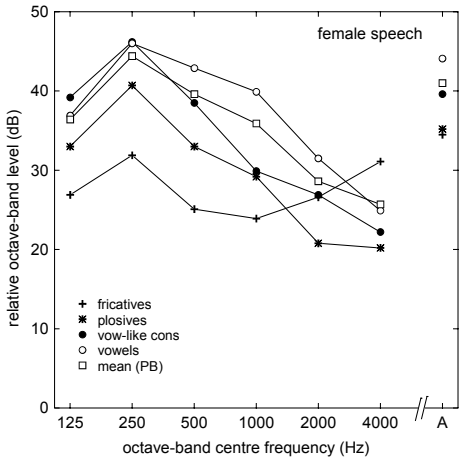


Figure 12. Relative test signal spectra for the four phoneme groups and for phonetically balanced speech (connected discourse). The dBA values represent the relative level of each group for connected discourse of FEMALES.

2.6. DIAGNOSTIC FEATURES, SOME EXAMPLES

The STI-method allows for two types of diagnostic analysis (1) based on the analysis of the test signal, and (2) based on the type and level of the test signal.

(1) As shown in Fig. 4, the modulation index reduction (effective signal-to-noise ratio) is obtained for seven octave bands and 14 modulation frequencies. The contribution of each octave band to the STI-value represents information on the frequency response of the system and on the spectrum of a masking signal. Generally, a low modulation index in combination with a low octave level indicates a poor frequency response. However, a low modulation

index in combination with a high octave level represents a high impact of a masking signal. In Table V, an example is given of a communication system with a limited frequency transfer. The table represents a typical output of the STI-calculation using the STITEL program. Both the STI_r (according to the concept including a redundancy correction) and the STI (according to the concept given by Steeneken and Houtgast, 1980) are given. Also the CVC-word score, based on the STI_r value and on the test signal type (male, female) are presented. The signal level at the input of the analogue-to-digital system is given and expressed in dBμV (if no additional calibration correction is applied). This example refers to the frequency transfer of a normal telephone channel. By comparison of the spectrum of the input signal and the output signal the frequency transfer can be obtained. This method is only valid if no additional noise is added between input and output. As mentioned above this can be detected by the reduction of the modulation transfer (all the TIs are close to '1').

Table V. Example of the STI value, levels, and octave-band specific information for a transmission channel with a limited frequency transfer obtained with the STITEL method.

STI_r	=	0.89			(Male speech, corresponding CVC-word score 89%)				
STI	=	0.86							
Level	=	110.0 dB			Level correction = 0.0 dB				
Level (A)	=	107.8 dBA			AD/DA range : 6.0 V(pp) equals 16 bit				
Octave centre freq.		125	250	500	1000	2000	4000	8000	Hz
Octave level		70.5	101.8	107.4	103.6	98.4	82.1	52.1	dB
Mod. Index (m)		0.97	1.00	1.03	1.02	1.00	1.01	0.01	
m-correction		1.00	1.00	1.00	1.00	1.00	0.99	0.76	
Transm. Index (TI)		1.00	1.00	1.00	1.00	1.00	1.00	0.00	
Relative Freq-resp.		-40.22	-8.86	0.39	2.63	3.45	-6.85	-30.88	dB
Modulation Frequency		1.12	11.33	0.71	2.83	6.97	1.78	4.53	Hz

Table VI. Example of the STI value, levels, and octave-band specific information for a transmission channel with a limited frequency transfer and a white noise masking signal (signal-to-noise ratio 0 dBA).

STI_r	=	0.48			(Male speech, corresponding CVC-word score 56%)				
STI	=	0.48							
Level	=	108.5 dB			Level correction = 0.0 dB				
Level (A)	=	107.3 dBA			AD/DA range : 6.0 V(pp) equals 16 bit				
Octave centre freq.		125	250	500	1000	2000	4000	8000	Hz
Octave level		67.7	99.1	104.8	102.1	101.0	96.3	53.6	dB
Mod. Index (m)		0.92	0.92	0.95	0.72	0.27	0.02	0.02	
m-correction		1.00	1.00	1.00	1.00	1.00	1.00	0.15	
Transm. Index (TI)		0.86	0.85	0.92	0.64	0.36	0.00	0.00	
Relative Freq-resp.		-42.53	-11.04	-1.72	1.66	6.52	7.82	-28.88	dB
Modulation Frequency		1.12	11.33	0.71	2.83	6.97	1.78	4.53	Hz

An example of a combination of band-pass limiting and additive noise is given in Table VI. As the noise signal used for this example is white noise (increase of 3 dB per octave band) the modulation indices for the higher octaves are lower than those for the low frequency bands (the TIs decrease from 0.92 to 0.02).

The modulation transfer function (MTF) offers information concerning the type of distortion in the time domain. If only a stationary noise is added to the speech or test signal, the decrease of the modulation transfer will be modulation-frequency independent. This is illustrated in Fig. 13. The reduction of the modulation transfer (m) is also given as a function of the signal-to-noise ratio.

In the case of distortion in the time domain (automatic gain control, echoes, and reverberation) a modulation-frequency specific reduction will be obtained. Reverberation acts as a low-pass filter on the fluctuations of the envelope. This is shown by the MTF given in

Fig. 14. In this graph also the theoretical relation between the modulation reduction (m) and the reverberation time (T) is given according to Houtgast and Steeneken (1973).

For echoes a rippled modulation transfer is obtained. For a fixed echo delay time (τ) the modulated envelope of the reflected signal (relative level δ) will, as a function of the modulation frequency, vary in phase with respect to the primary signal. This result in a rippled modulation transfer function (MTF). In Fig. 15, an example of such a MTF is given. The theoretical relation is also given.

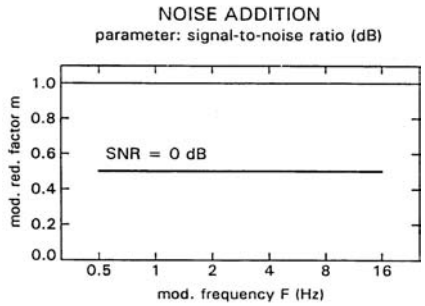


Figure 13. Example of the modulation transfer function for conditions with noise.

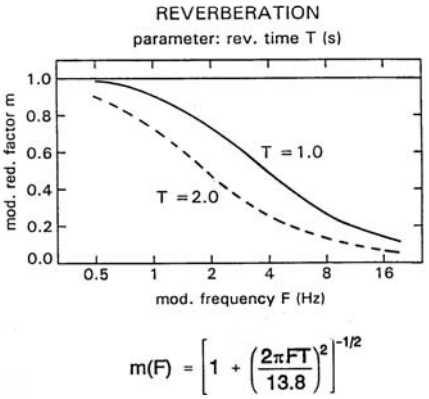


Figure 14. Example of the modulation transfer function for conditions with reverberation.

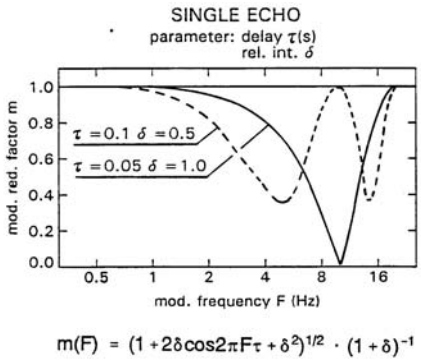


Figure 15. Example of the modulation transfer function for conditions with echoes.

Automatic gain control systems mainly reduce slow level variations, which may be described as a high-pass filter applied to the envelope fluctuations. Here, the relation between the modulation reduction and the attack and release time of the AGC is complex and cannot be represented by a simple formula. It is obvious that for systems with a distortion in the time domain the full MTF should be determined, hence based on the complete matrix of Fig. 4.

(2) Another method to obtain diagnostic information is to vary the test signal level or the type of test signal. Variation of the test-signal level will discriminate between signal-level dependent and signal-level independent distortions. For example, the effect of masking noise will increase at lower test signal levels while the effect of reverberation and echoes is not signal-level dependent. This feature is often used in the evaluations in room acoustics and becomes even more powerful if it is used in combination with the frequency dependent analysis as mentioned above.

As described in section 3 of this chapter, a specific test signal is applied for non-linear communication channels. While performing an analysis in one of the seven octave bands, uncorrelated fluctuations are present in the other six octave bands. These may introduce distortion components within the octave band under test and hence reduce the modulation transfer.

Comparison of the modulation transfer (or the STI) for a given channel measured with two types of test signals, one with the representative uncorrelated fluctuations present and one measured without these fluctuations, will show the effect of the deterioration by the non linear frequency transfer.

2.7. SPEECH AND TEST SIGNAL LEVEL ADJUSTMENT

For reproducible experiments concerning the effect of noise on speech transmission quality, it is important to specify the speech levels, the noise levels and the corresponding signal-to-noise ratios.

Various studies (Brady, 1965; Kryter, 1970; Berry 1971; Steeneken and Houtgast, 1978, 1986) have defined speech level measures. It was also shown that a signal-to-noise ratio variation of only 1–2 dB might have the same effect on the results as typical speaker and inter-listener variations. We therefore specified a method for measuring speech levels and noise levels, which offers such a resolution. The measure should be robust for the various speech types (male/female, connected discourse/isolated words), recording conditions (background noise, frequency transfer), and should also be applicable to noise signals. We have developed such a measure (Steeneken and Houtgast, 1978, 1986) mainly for adjusting the signal level of the STI test signal to the speech level for similar conditions. The measuring method was made generally available by development of a, platform independent, digital signal-processing algorithm.

2.7.1. Speech level measuring method

A high correlation was found between the speech level and the speech intelligibility for level measures based on frequency-weighted speech signals with a reduced contribution of frequency components below approx. 250 Hz (Kryter, 1970; Steeneken and Houtgast, 1978, 1986). The standardised frequency-weighting function according to the A-filter was used for this purpose (standardised for acoustical measurements).

After filtering, the running (intensity) envelope is determined by squaring and low-pass filtering (47 Hz) the waveform. From this envelope function the envelope distribution histogram is obtained, and the RMS value can be computed from this histogram. The advantage is that the RMS value can also be obtained for values above a certain level after

sampling. In order to compare the level of short speech tokens (simple words altered with long silent periods) and the level of connected discourse, a level threshold for suppression of the silent periods is required. Hence, this threshold is applied to the envelope function of the speech signal rather than to the waveform, and therefore does not affect each zero crossing of the speech signal. The threshold level is defined to be 14 dB below the resulting RMS level (iterative procedure). This definition is signal-related and does not strongly depend on other effects such as background noise level (down to signal-to-noise ratios of 4 dB), shape of the envelope distribution, etc. The same principle can be applied to stationary noises but in that case the threshold function is not effective.

The relation between various level measures obtained from two types of speech signals (connected discourse, and CVC words in a short carrier phrase) is given in Fig. 16.

The level measures are: the 1% peak level (1% overflow criterion), the mean of the peak deflections of a sound level meter set to "fast" (dBA fast), the RMS values obtained with a squaring detector from the envelope function (RMS, true rms), the RMS values obtained with direct sampling and by squaring the waveform samples (RMS_{dir}, true rms), and the equivalent peak level (EPL) according to Brady (1968). The last method is not applicable to noise signals. The RMS-A_{thr} is obtained with the speech level measuring program SLM and is also used by the former (EU sponsored) Esprit-SAM group. For some of these measures the use of the A-weighting or a threshold is applied, this is indicated by a suffix respectively (A) or (thr).

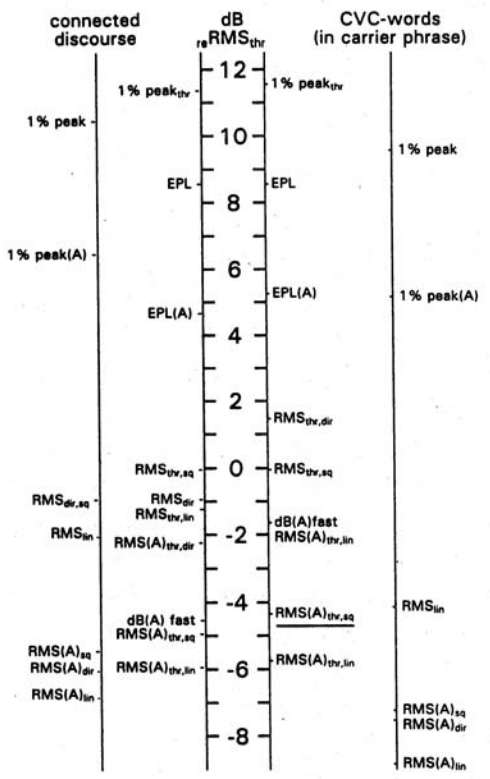


Figure 16. Relative speech levels for different speech-level measures applied to connected discourse and embedded CVC words. The values are relative to the RMS_{thr} value. For STI measurements (related to CVC-word score prediction) the test-signal level must be adjusted –2 dB to the (underlined) RMS-A_{thr} value.

2.8. APPLICATION EXAMPLES

The STI method can be used for speech communication systems: (1) radio links, intercoms or digital (waveform based) speech coders, (2) electro-acoustic transducers (microphone and telephone), and (3) for room acoustics. Although the STI measuring method for all three types of speech communication systems is the same, some system-specific simplifications of the measuring method are allowed. This leads to a faster result. Usually linear communication channels and electro-acoustic transducers (used close to the mouth or ear) can be assessed with the STITEL measuring program. For room acoustical applications the full STI measurement (STI-14, or STIPA) should be used. However for some specific applications (such as public address systems in open environment) we may also use STITEL. This has to be decided by making a reference measurement with STI-14 and check that echoes or reverberation does not affect the MTF.

The method of connecting the test signal to the system under test is different for the various systems. For communication systems an electrical input and output can be used. However, for applications with microphones or in room acoustics an artificial mouth has to be used in order to obtain an acoustically coupled test signal. For testing telephones and headsets an artificial ear is used. It is obvious that also a test of a complete communication system is possible (including microphone, communication system, headset and acoustically added background noise).

In the next sections, some examples of these applications are given.

2.8.1. Communication channels

The first example concerns a diver underwater telephone system. The evaluation method of such a system with the STI approach is similar to the method used for radio links or other transceivers. The effect of various parameters upon the transmission quality can be studied. The following parameters are of interest: the STI value as a function of the range between transmitter and receiver, propagation conditions, and the input level of the modulator (especially if there is no automatic gain control).

The underwater telephone system presented in this example consists of a base station and a diver station. At the base station side the acoustical transmitter receiver (a hydrophone) was placed in the water of a lake at a depth of 3 m. At various distances (4 m, 100 m, and 125 m) the diver set was put into the water at a depth of 3 m. Such an underwater telephone system is based on an amplitude modulated carrier with a carrier frequency between 8 and 40 kHz. This is similar to a radio communication link but with a relatively low carrier frequency. The STI test signal was electrically connected with the transmitter (base station). The test signal input level was variable. At the diver station side an electrical output (headphone) connection was used. In Fig. 17, the STI_r (obtained with the STITEL method) is given for the three distances between transmitter and receiver and as a function of the input level. For this type of application the maximum range is obtained at a STI_r of 0.35, which is related to a sentence intelligibility of just 100% (for very simple sentences). For the 4 m and 100 m distance this STI_r value is obtained at various input levels. However at a distance of 125 m, which is a condition without a direct view between the two hydrophones, a very low STI_r is obtained.

In this example fixed conditions (distance between transmitter and receiver, and fixed input level of the modulator) were used. However, for some applications a continuously increasing range (e.g. a transmitter in a vehicle moving from or to the receiver) may be more appropriate. For this purpose a continuous analysis is made at the receiving side while at the transmitter side the test signal can be supplied from tape.

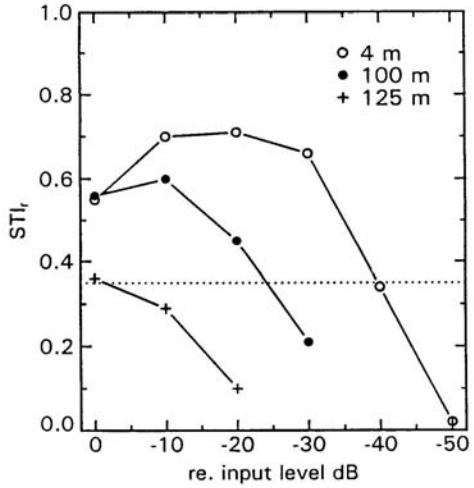


Figure 17. STI_r as a function of the audio input level of the transmitter, for three distances between an underwater telephone base station and diver station at a carrier frequency of 40 kHz.

A second example of a communication channel concerns digital waveform coders. With waveform coders, parameters such as bitrate and bit errors are to be considered. We compared two CVSD systems (Continuous Variable Slope Delta Modulation) at a bitrate of 8 kb/s and 16 kb/s. In the connection between the coder and decoder of the systems, random bit errors were introduced. The bit error rate could be varied in steps of 1%. In Fig. 18, the STI_r for both systems as a function of the bit error rate is given.

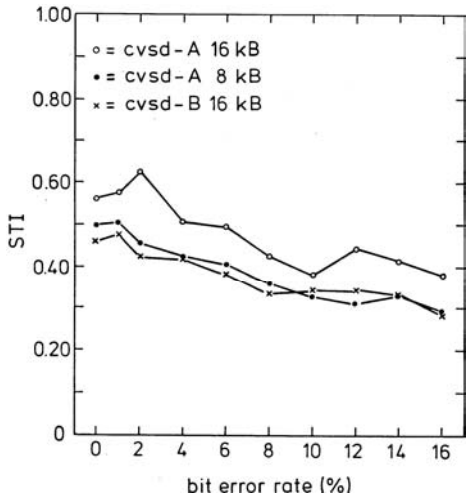


Figure 18. STI_r for two CVSD systems, two bit rates as a function of the bit error rate.

The measurements were performed with the STI-3 method, making use of three modulation frequencies within each octave band and suitable for non linear distortion. The results show that system A offers a better performance than system B. It is also shown that

system A gives the same intelligibility at 8 kb/s as system B at 16 kb/s. The results also show the robustness of these CVSD systems with respect to bit errors.

2.8.2. Electro-acoustic transducers

Microphones and telephones (headsets) are often used in noisy environments. Therefore, the assessment of these transducers should be performed in such an environment or by simulation. A second point of consideration for the use of a microphone is its position close to the mouth.

For the assessment of a microphone, an acoustical coupling is required. We developed an artificial mouth consisting of a (horn loudspeaker) driver unit, artificial head and connection tube between driver and outlet (mouth). The frequency transfer between the driver and the mouth is, due to the resonances in the tube, not flat. Therefore, the tube was filled with sound absorption material. This resulted in a frequency transfer that is flat within 10 dB. With the addition of a $^1/_3$ octave equaliser a flat response between 100Hz and 10 kHz was obtained. The system was built into a box with the shape of a torso (see Fig. 19). At the moment of design no systems with suitable specifications were commercially available.

Figure 19. Artificial mouth used for the assessment of microphones.

The level at 1 m distance in front of the mouth is typically 60 dBA. However to simulate a raised voice level (Lombard effect), the system can produce an undistorted signal with a level up to 75 dBA at 1 m distance. The radiation pattern is similar to that of humans. The system can also be used in room acoustics as an artificial speaker with a representative radiation.

Some artificial heads (including an artificial mouth and ears) are commercially available. It should be verified that the following specifications are fulfilled:

(1) the frequency response must cover the frequency range of the STI test signals (85 Hz – 11.2 kHz),
(2) the maximum level at 1 m distance in front of the mouth must exceed 60 dBA, preferably 75 dBA,
(3) the radiation pattern (also close to the mouth) must be representative for humans.

The artificial mouth, shown in Fig. 19, is normally used in a high noise room where a diffuse sound field can be produced. The microphone to be tested is placed at the required position in front of the artificial mouth. The STI is measured by connecting the test signal to the artificial mouth and by analysing the microphone output. The measurements are normally performed at various microphone positions and various levels of the background noise.

In Fig. 20, the STI as a function of the noise level for two microphones is given.

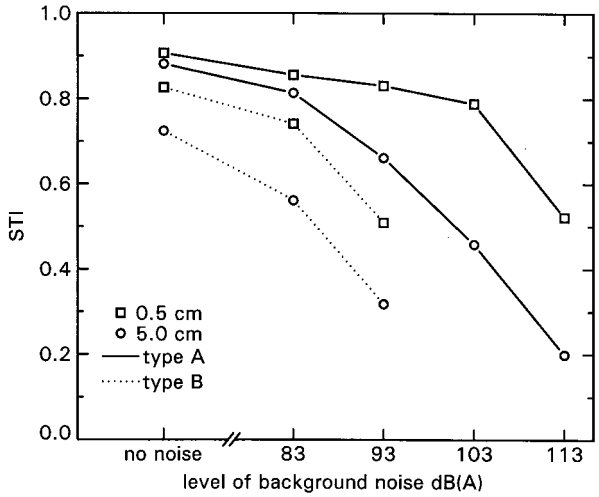


Figure 20. STI_r for two microphones, at two positions in front of the mouth and as a function of the background noise level (for noise of a diesel engine).

For the assessment of telephones an artificial ear is required. Especially for the assessment of headphones mounted in earmuffs the head size, hair and wearing spectacles may influence the sound attenuation and the intelligibility. Therefore, normally a number of five subjects is used with a miniature electret microphone mounted near the ear canal. This is illustrated in Fig. 21B. The mounting and wiring of the microphone assembly is such that it does not interfere with the proper use of a telephone handset or a headset.

For measurements in combination with background noise a high noise room with an adjustable noise level is used. The subject, with the (miniature) sense microphone mounted close to the ear canal entrance, is positioned within this room. Special care must be taken that the subject is not exposed to sound levels above 85 dBA with unprotected ears. In order to obtain calibrated levels, the gain of the recording chain (microphone, microphone pre-amplifier and recording system) must be included in the STI measuring procedure (this can be done by adjusting the correction factor in the configuration file of the STI calculation program).

In general the presentation level of the speech (test) signal with a telephone is 60 – 75 dBA. Background noise levels may vary between 50 – 60 dBA (office) to 105 dBA (inside a fighter cockpit) or even up to 115 dBA (inside an armoured car or helicopter). In Fig. 22, the STI_r is given for two types of telephone systems as a function of the background noise level (STITEL method).

Figure 21. Subject positioned in a high-noise room and the mounting of the electret microphone near the entrance of the ear canal.

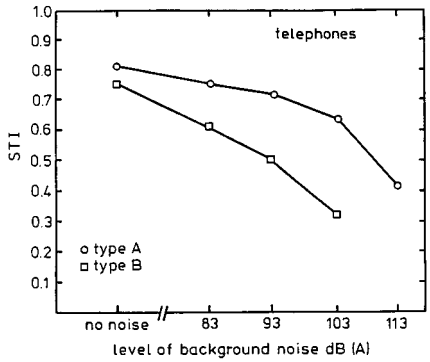


Figure 22. STI_r for two types of headset as a function of the background noise level. The presentation level of the test signal was 75 dBA.

2.8.3. Room acoustics and public address systems

Measurements in auditoria or with public address systems are normally performed with the STI-14 or STIPA method. This includes the measurement of the MTF for 14 modulation frequencies. If a smooth MTF is obtained, one can decide to decrease the resolution by skipping modulation frequencies. For some applications it is not necessary to measure the

MTF for all the seven frequency bands. The resolution in the frequency domain may be reduced to two octave bands (with a centre frequency of 500 Hz and 2000 Hz). This is only valid when no limitation in frequency transfer is effective (no PA-systems) and the background noise is of minor importance or can be described by samples within these two frequency bands. The RASTI method is an example of an application with these limitations.

An example of the use of the RASTI method is given in Fig. 23. For a number of positions in an auditorium the STI was measured and the results were plotted in a lay-out of the room. Adjacent measuring points with a similar STI value were connected. This results in iso-STI contours. The contours are usually made at intervals of 0.05 STI. A high gradient of the STI indicates a poor distribution of the intelligibility in the room.

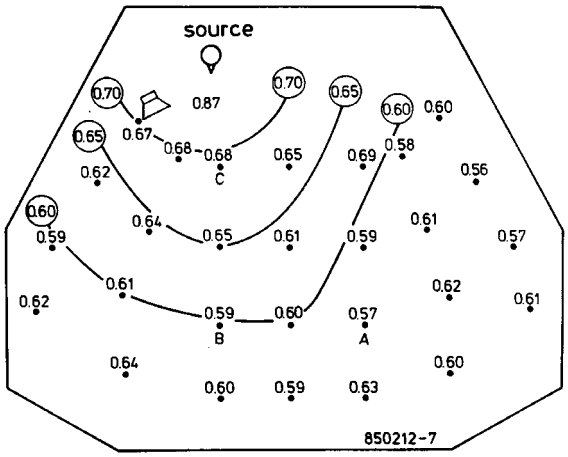


Figure 23. Iso-STI contours for an auditorium with no background noise.

If a PA system is used, this system may increase the direct speech level, but it increases also the level of reverberating speech sound. This depends on the directivity and positioning of the loudspeaker(s) and the presence of sound absorbing material (e.g., the public). Hence in some cases the use of a PA system may be beneficial, in other cases it may reduce the speech intelligibility. An example of this effect is given in Fig. 24. For three positions in the auditorium of Fig. 23 (A, B,C) the full STI is measured as a function of the noise level and for the condition with and without the PA system. Position B shows an increase and position C shows a decrease of the STI due to the PA system.

For position A and B the STI as a function of the noise level is increased. The horizontal shift of the two curves (with and without PA system) shows the effective gain (the same STI at higher noise levels). It is obvious that this gain is minimal for position C. This method can be used to optimise PA systems.

The MTF and the reverberation time in an enclosure are related (theoretically) according to the formula given in section 6 (Fig. 14). Hence, based on the measured MTF, the reverberation time T can be estimated. This is demonstrated in Fig. 25. In this graph the MTFs measured for several conditions in the same auditorium are given. Two parameters were varied: (a) the use of the PA system and (b) additional sound absorbing material spread on the floor. The MTF given in the graph is measured within the octave band with centre frequency 2000 Hz. It is shown that for this example the use of the PA system does not affect the MTF and hence does not change the reverberation time. The use of additional absorbing material however, has a significant effect on the MTF.

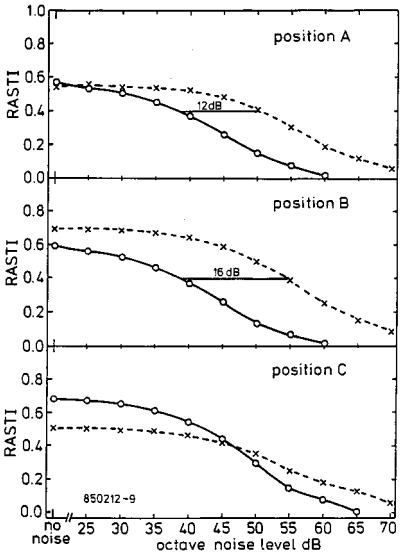


Figure 24. STI as a function of the background noise level for three positions in the auditorium of Fig. 23, and with the PA system switched on and off.

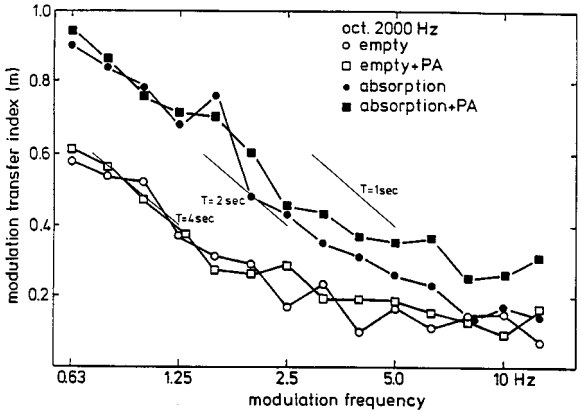


Figure 25. MTF for one position in an enclosure and based on the octave band with centre frequency 2000 Hz. Four conditions are observed being the combination of a PA system switched on and off and the use of additional sound absorbing material.

In Fig. 25, a small part of the theoretical MTFs corresponding to reverberation times of 1, 2, and 4 s respectively are drawn. These curves are calculated according to the formula given in Fig. 14. This formula is based on a simple exponential decay curve (no coupled enclosures). The reverberation time is estimated by fitting the measured MTF with the theoretical MTFs. In the example the reverberation time T = 4 s for the two conditions without additional absorbing material and approximately T = 2.2 s for the condition with the additional absorbing material. It should be noted that the MTF approach is closely related to the perception of fluctuations. The predicted reverberation time is related to the early decay time rather than to the conventional reverberation time.

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.

ANSI (**1969**). ANSI S3.5-1969, American national standard methods for the calculation of the articulation index, American National Standards Institute, New York.

Barnett, P. W. and Knight, R.D. (1995). "The Common Intelligibility Scale", Proc. I.O.A. Vol **17**, part 7.

Barnett, P. W. (**1999**). "Overview of speech intelligibility" Proc. I.O.A Vol **21** Part 5.

Berry, R.W. (**1971**). "Speech volume measurements on telephone circuits," Proc. IEE 118(2), 335-338.

Bos, C.S.G.M., and Steeneken, H.J.M. (**1991**). "Phoneme confusions in distorted speech: a diagnostic study," Report IZF 1991 I-4, TNO Institute for Perception, Soesterberg, The Netherlands.

Brady, P.T. (**1965**). "A statistical basis for objective measurement of speech levels", Bell System Tech. J. **44**, 1453-1486.

Brady, P.T. (**1968**). "Equivalent Peak Level: A threshold-independent speech-level measure," J. Acoust. Soc. Am. **44**, 695-699.

Dunn, H.K., and White, S.D. (**1940**). "Statistical measurements on conversational speech", J. Acoust. Soc. Am. **11**, 278-288.

Egan, J.P. (**1944**). "Articulation testing methods," OSRD report No. 3802.

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

Fletcher, H., and Steinberg, J.C. (**1929**). Bell Sys Tech. J. **8**, 806.

Fletcher, H., and Galt, R.H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89-151.

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

French, N.R., and Steinberg, J.C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90-119.

Hecker, M.H.L., Bismarck G. von, and Williams, C.E. (**1986**). "Automatic evaluation of time-varying communications systems," IEEE Trans. on Audio and Electroacoustics AU-16, 100-106.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (**1965**). "Articu-lation testing methods: Consonantal differentiation with a closed-response set," J. Acoust Soc. Am. **37**, 158-166.

Houtgast, T., and Steeneken, H.J.M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355-367.

Houtgast, T., and Steeneken, H.J.M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66-73.

Houtgast, T., Steeneken, H.J.M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60-72.

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.

Houtgast, T., and Steeneken, H.J.M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069-1077.

Houtgast, T., and Verhave, J. (**1991**). "A physical approach to speech quality assessment: correlation patterns in the speech spectrogram," Proc. Eurospeech '91, Genova, 285-288.

IEC-report (**1988**). "The objective rating of speech intelligibility in auditoria by the 'RASTI' method," Publication IEC 268-16.

IEEE (**1969**). "Speech quality measurements," IEEE Transactions on Audio and Electroacoustics, September, 227-246.

Jacob, K., Steeneken, H.J.M., Verhave, J., and McManus, S., (**2001**). "Development of an accurate, handheld, simple-to-use meter for the prediction of speech intelligibility". Institute of Acoustics, Proc. Reproduced Sound 17, Stratford-upon-Avon.

Kryter, K.D. (**1960**). "Speech band-width compression through spectrum selection," J. Acoust. Soc. Am. **32**, 547-556.

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

Kryter, K.D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698-1702.

Kryter, K.D., and Ball, J.H. (**1964**). "SCIM -- A meter for measuring the performance of speech communication systems," Techn. Doc. report No. ESD-TDR-64-674.

Kryter, K.D. (**1970**). The effects of noise on man (Academic Press).

Licklider, J.C.R. (**1959**). "Three auditory theories," in Psychology: A Study of Science, Vol. 1, edited by S. Koch (McGraw-Hill, New York), pp 41-144.

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.

Mapp, P. (**2001**) "Improving the intelligibility of aircraft PA-systems" Proc Institute of Acoustics, reproduced souns 17, Stratford-upon-Avon.

Miller, G.A., and Nicely, P.E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338-352.

Payne, J.A., and McManamon, P.M. (**1973**). "An objective speech quality measurement of a communication channel," OT report 73-14, Department of Commerce, Office of Telecommunications.

Pavlovic, C.V., and Studebaker, G.A. (**1984**). "An evaluation of some assumptions underlying the articulation index," J. Acoust. Soc. Am. **75**, 1606-1612.

Pavlovic, C.V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413-422.

Plomp, R., Steeneken, H.J.M., and Houtgast, T. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. II. Mirror image computer model applied to rectangular rooms," Acustica **46**, 73-81.

Pollack, I. (**1948**). "Effect of high pass and low pass filtering on the intelligibility of speech in noise," J. Acoust. Soc. Am. **20**, 259-266.

Quackenbush, S.R., Barnwell, T.P., and Clements, M.A. (**1988**). Objective Measures of Speech Quality (Prentice Hall, New Jersey).

Raaij, J.L. van, and Steeneken, H.J.M. (**1991**). "Digital simulation of speech transmission channels," Report IZF 1991-A7, TNO Institute for Perception, Soesterberg, The Netherlands.

Rietschote, H.F. van, Houtgast, T., and Steeneken, H.J.M. (**1981**). "Predicting speech intelligibility in rooms from the modulation transfer function. IV. A ray-tracing computer model," Acustica **49**, 245-252.

Schroeder, M.R., (**1981**) "Modulation Transfer functions: Definition and Measurement", Acustica **49**, pp.179-182.

Schwartzlander, H. (**1959**). "Intelligibility evaluation of voice communications," Electronics **29**, 88-91.

Steeneken, H.J.M., and Houtgast, T. (**1973**). "Intelligibility in telecommunication derived from physical measurements," Proc. Symp. Intelligibilité de la Parole, Liège, 73-80.

Steeneken, H.J.M., and Houtgast, T. (**1978**) "Comparison of some methods for measuring speech levels," Report IZF 1978-22, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Houtgast, T. (**1979**). "Measuring ISO-intelligibility contours in auditoria," Proc. 3rd Symp of FASE on building Acoustics, Dubrovnik, 85-88.

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., and Agterhuis, E. (**1982**). "Description of STIDAS II D, Part 1, General system and program description," Report IZF 1982-29, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Houtgast, T. (**1982**). "Some applications of the Speech Transmission Index (STI) in auditoria," Acustica **51**, 229-234.

Steeneken, H.J.M., and Houtgast, T. (**1983**). "The temporal envelope spectrum of speech and its significance in room acoustics," Proc. 11th International Congress on Acoustics, Paris, Vol. 7, 85-88.

Steeneken, H.J.M., and Houtgast, T. (**1986**). "Comparison of some methods for measuring speech levels," Report IZF 1986-20, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M. (**1987a**). "Diagnostic information of subjective intelligibility tests," Proc. IEEE ICASSP, Dallas, 131-134.

Steeneken, H.J.M. (**1987b**). "Comparison among three subjective and one objective intelligibility test," Report IZF 1987-8, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Houtgast, T. (**1991**). "On the mutual dependency of octave-band specific contributions to speech intelligibility," Proc Eurospeech '91, Genova, 1133-1136.

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.

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,** 399-411.

Steeneken, H.J.M., and Houtgast, T. (**2002b**). "Validation of the revised STIr method". Speech Communication, 2002, vol. **38**, 413-425.

Studebaker, G.A., Pavlovic, C.V., and Sherbecoe, R.L. (**1987**). "A frequency-importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130-1138.

Studebaker, G.A., and Sherbecoe, R.L. (**1991**). "Frequency-importance and transfer functions for recorded CID W-22 word lists," J. Speech Hear. Res. **34**, 427-438.

Voiers, W.D. (**1977a**). "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.

Voiers, W.D. (**1977b**). "Diagnostic acceptability measure for speech communication systems," Proc. IEEE ICASSP, Hartford CT, 204-207.

Wattel, E., Plomp, R., Rietschote, H.F. van, and Steeneken, H.J.M. (**1981**). "Predicting speech intelligibility in rooms from the modulation transfer function. III. Mirror image computer model applied to pyramidal rooms," Acustica **48**, 320-324.

Van Wijngaarden, S.J., Steeneken, H.J.M. (**1999**) Objective prediction of speech intelligibility at high ambient noise levels using the Speech Transmission Index Proc Eurospeech99, Budapest, 2639-2642.

Zwicker, E, and Feltkeller, (**1967**). Das Ohr als Nachrichtenempfanger, (Hirzel Verlag, Stuttgart), 187-200.

Chapter 3. Improvements of the STI method: frequency weighting, gender, level dependent masking, and phoneme specific prediction

Herman J.M. Steeneken

3.1. 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).

3.2. 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.

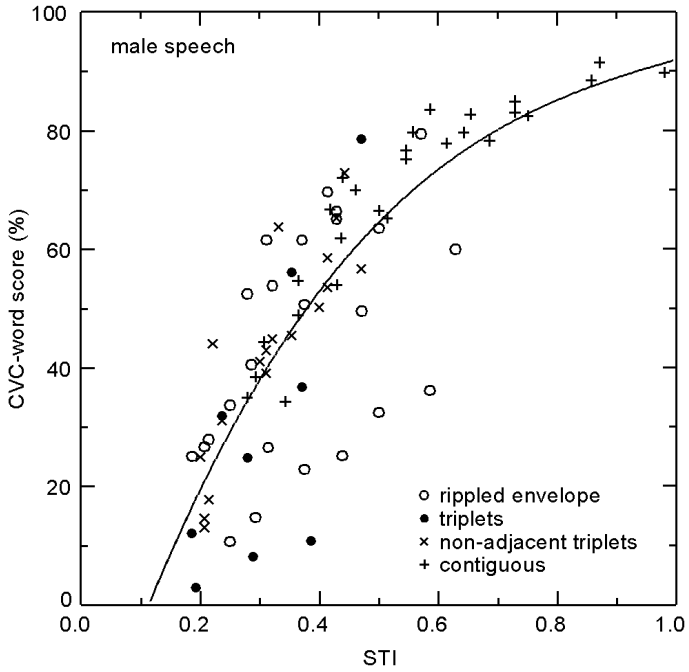


Figure 1. Relation between the STI and the CVC-word score for the 78 conditions involving MALE speech. The parameters used for the STI calculation were adopted from the procedure described previously by Steeneken and Houtgast (1980). The standard deviation, representing the vertical spread around the 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:

$$index = \alpha_1 \cdot TI_1 - \beta_1 \cdot \sqrt{(TI_1 \cdot TI_2)} + \alpha_2 \cdot TI_2 - \beta_2 \cdot \sqrt{(TI_2 \cdot TI_3)} + ... + \alpha_7 \cdot TI_7 \qquad (1)$$

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 STI_r.

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%.

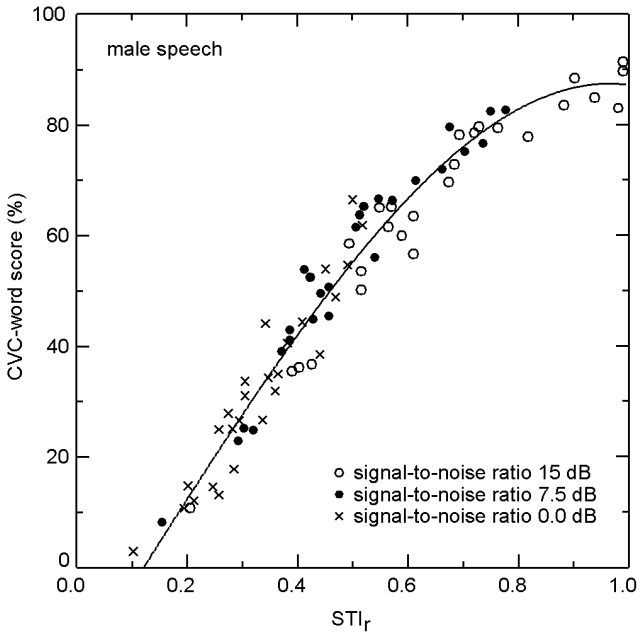


Figure 2. Relation between the STI_r and the CVC-word score for the 78 conditions involving MALE speech. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 4.7%.

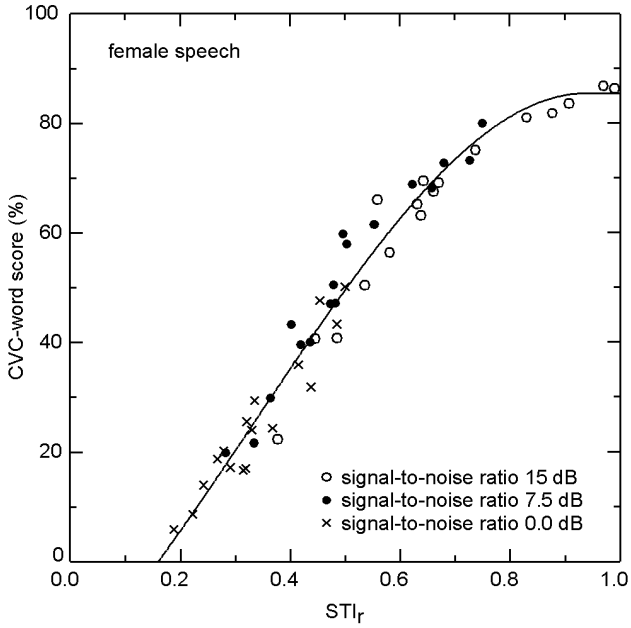


Figure 3. Relation between the STI_r and the CVC-word score for the 51 conditions involving FEMALE speech. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 4.2%.

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.

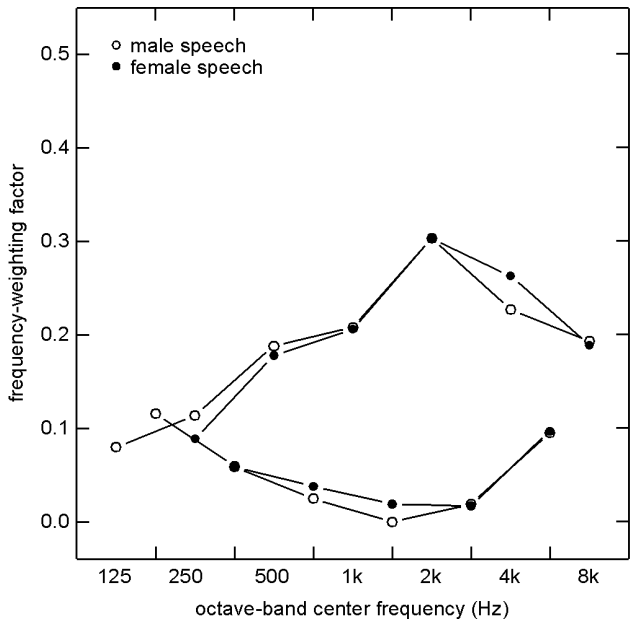


Figure 4. Optimal set of frequency-weighting factors α_k (upper curves) and redundancy-correction factors $ß_k$ (lower curves), derived separately for the male and female conditions as given in Figs. 2 and 3.

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%.

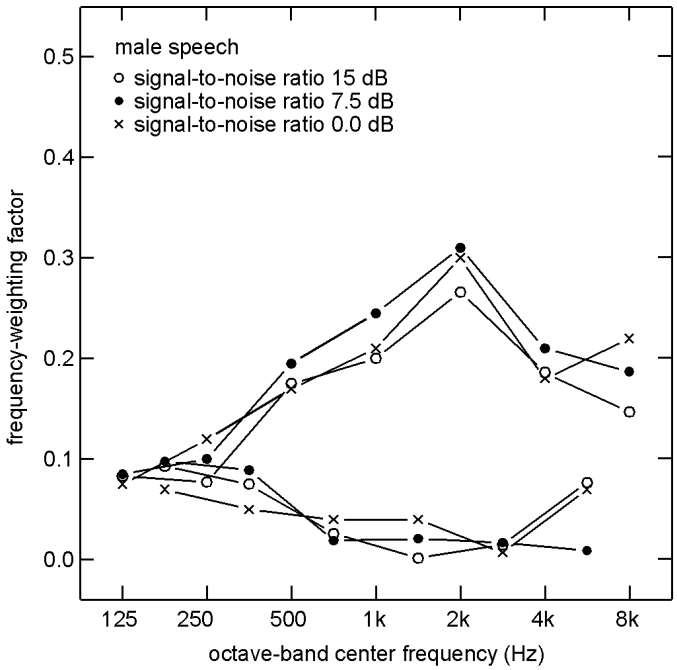


Figure 5. Frequency-weighting factors α_k (upper curves) and redundancy-correction factors $ß_k$ (lower curves) for the MALE conditions at three signal-to-noise ratios as given in Fig. 6.

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

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

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.

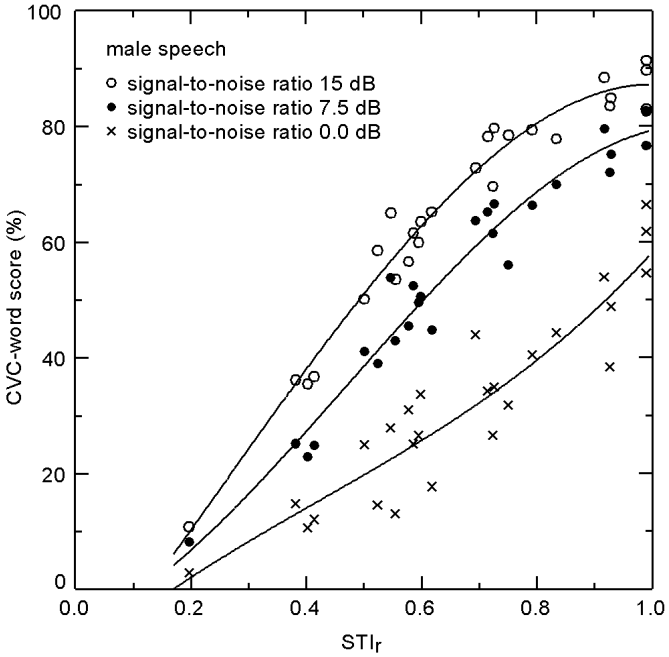


Figure 6. Relation between the STI_r (with TI_k=1.0 for all three signal-to-noise ratios) and the CVC-word score for the conditions involving MALE speech.

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 score	SNR 15–7.5 dB			SNR 15–0 dB		
	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

3.4. 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.

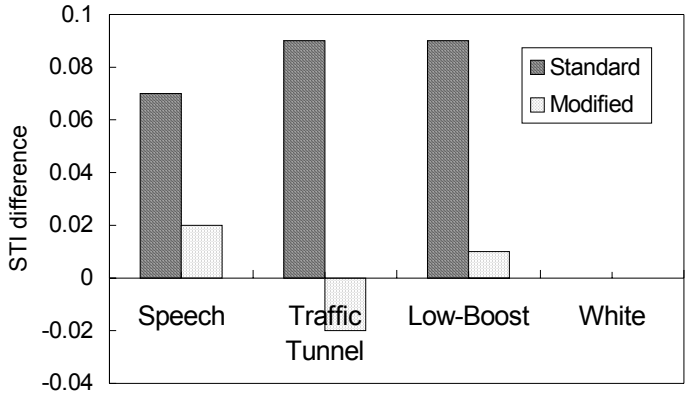


Figure 7. Differences in STI between 75 and 105 dBA (standard and modified model) based on male SRT results (2 talkers, 4 listeners).

3.5. 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.

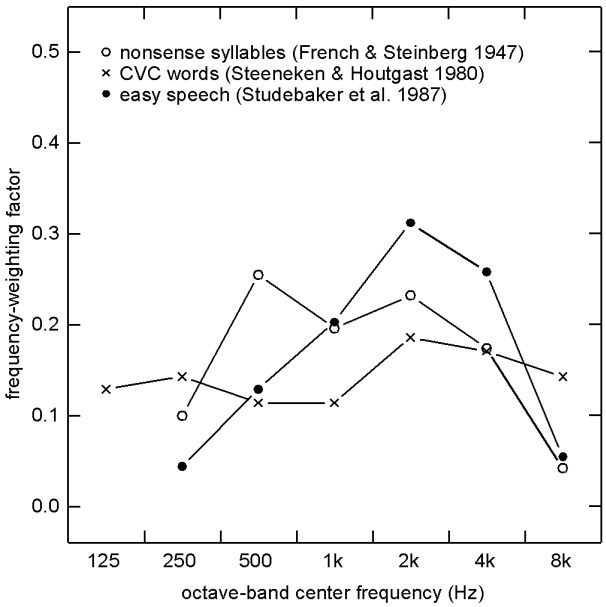


Figure 8. Frequency weighting functions for the AI and STI after French and Steinberg (1947), Steeneken and Houtgast (1980), and Studebaker et al. (1987).

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 groups, 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.

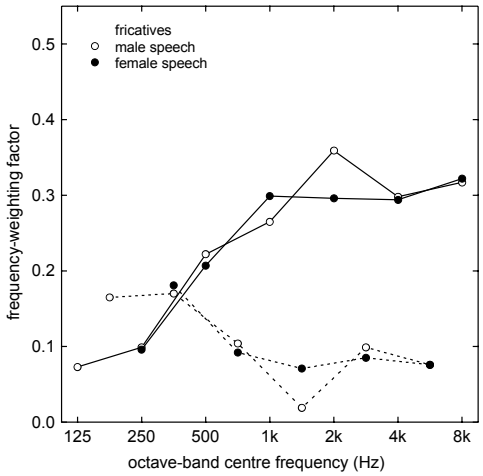


Figure 9. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the FRICATIVES and for the male and female speech.

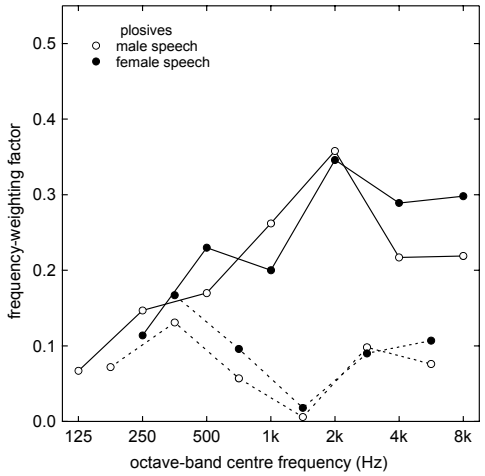


Figure 10. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the PLOSIVES and for the male and female speech.

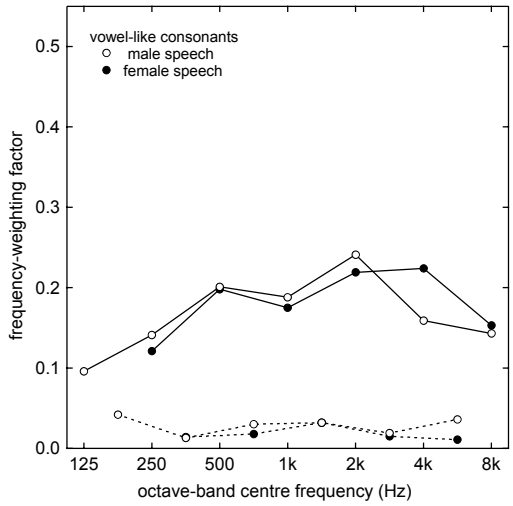


Figure 11. Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction $ß_k$ (dashed line) for the VOWEL-LIKE consonants and for the male and female speech.

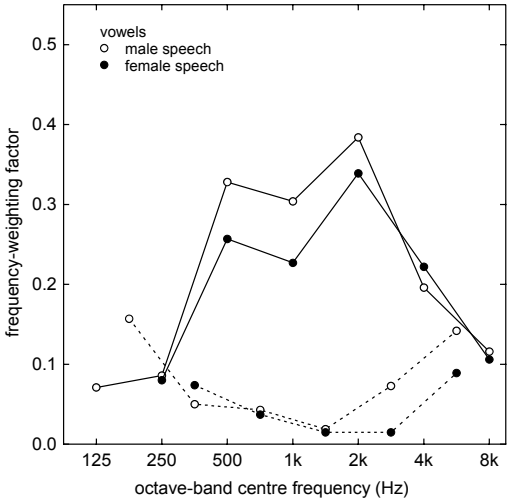


Figure 12. Frequency-weighting factors for the octave-band contribution αk (solid line) and redundancy correction $ß_k$ (dashed line) for the VOWELS and for the male and female speech.

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%.

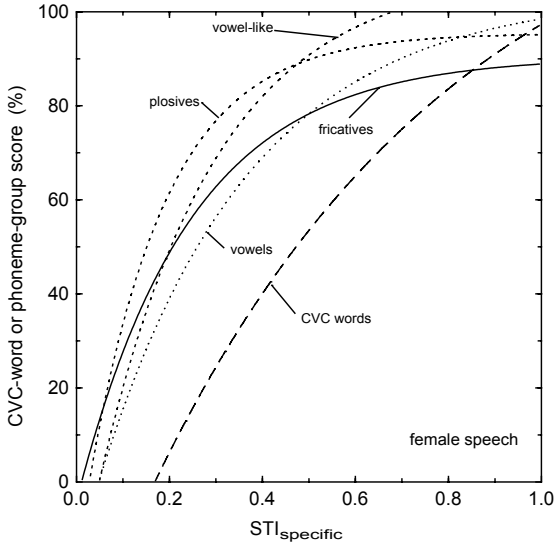


Figure 13. Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for MALE speech. The relation for the CVC-word score is also given.

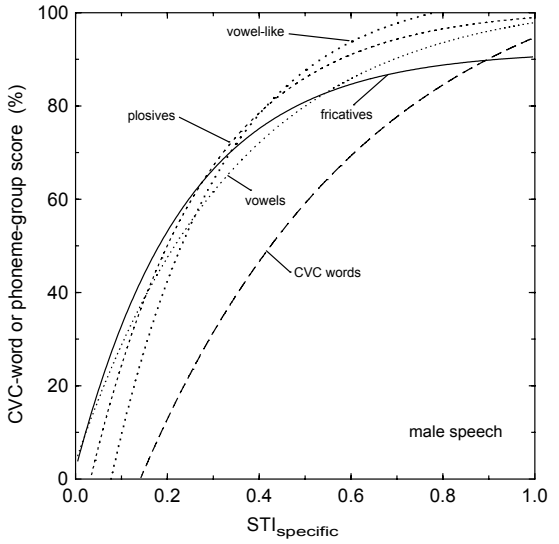


Figure 14. Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for FEMALE speech. The relation with the CVC-word score is also given.

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%).

3.6. 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 Figs. 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%.

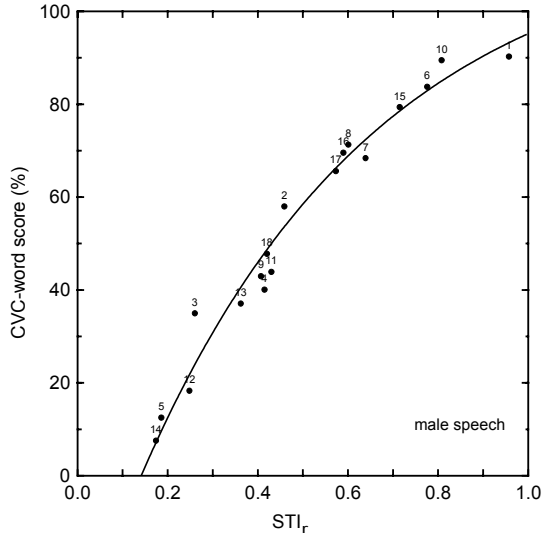


Figure 15. Relation between the STI_r and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 4.4%.

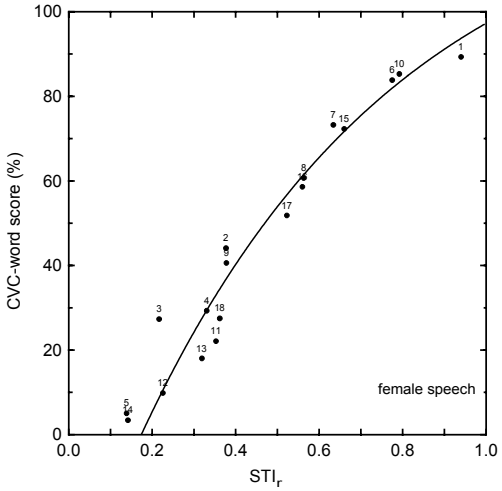


Figure 16. Relation between the STI_r and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for FEMALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 6.6%.

Similar tests were performed with test conditions that include nonlinear distortion such as peak-clipping, center clipping and quantisation 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.

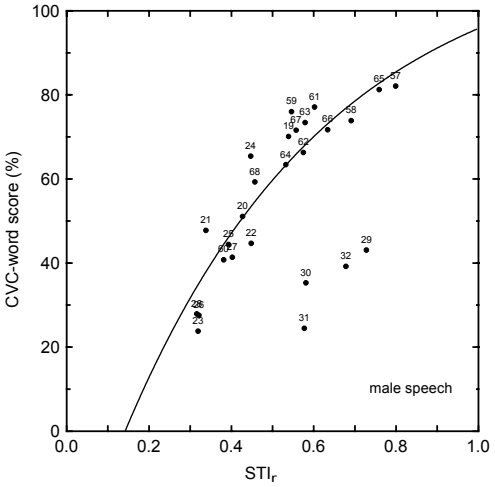


Figure 17. Relation between the STI_r and the CVC-word score for the 26 communication channel conditions including nonlinear distortion for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial for the conditions excluding center clipping (29-32) is s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

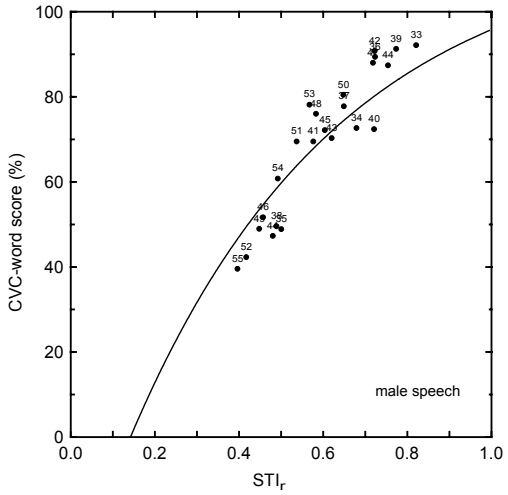


Figure 18. Relation between the STI_r and the CVC-word score for the 24 communication channel conditions including distortion in the time domain for MALE speech. The standard deviation, representing the vertical spread around the predefined polynomial is s = 6.9%.

3.7. 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 digital voice 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,** 399-411.

Steeneken, H.J.M., and Houtgast, T. (**2002b**). "Validation of the revised STIr method". Speech Communication, 2002, vol. **38**, 413-425.

van Wijngaarden, S.J., 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.

Chapter 4. Limitations of the STI method

Rob Drullman

ABSTRACT

The STI concept has shown to be very successful in predicting speech intelligibility in a large number of practical situations. However, the STI and the underlying concept of the modulation transfer function do not have an unrestricted application. In the present chapter, a number of limitations will be discussed and examples will be given of specific situations and signal manipulations which lead to significant under- or overestimation of the STI. This chapter will focus on the effects of direct manipulation of the temporal envelope.

4.1. INTRODUCTION

Speech can be considered as a summation of frequency bands with amplitude-modulated signals. Each frequency band consists if a fine structure (carrier) and a time-varying envelope. This envelope describes the temporal fluctuations reflecting the sequence of the different speech sounds and may be regarded to represent the information-bearing characteristics of speech. The importance of the temporal envelope in evaluating the quality of speech transmission has been elaborated in the concept of the modulation transfer function (MTF; Steeneken and Houtgast 1980), resulting in the STI. According to this concept, the detrimental effects of noise and reverberation on speech are adequately measured in terms of the reduction in modulation depth they produce in each band. The STI method has proven to give valid results for a great number of distortions in the time domain as well as for several non-linear distortions.

However, there are situations and possible (narrow-band) signal manipulations in which one must take good care before using the STI. A number of limitations have been known for some time, and they are reported in earlier work (Steeneken, 1992) and in international standards (cf. IEC 60268-16, 1998). In the sections below, I will summarise some of these limitations and give examples of the impact on the STI value. In some cases intelligibility hardly suffers from a distortion, but the STI shows a significant reduction. In other cases the STI shows only minimal changes, whereas intelligibility drops down considerably.

As for non-linear processing, the general applicability of the STI- and MTF concepts have shown to be rather poor in techniques of dynamic amplitude compression (and expansion), signal processing schemes proposed in hearing aids design. In a way, these direct manipulations of the temporal envelopes of a series of frequency bands appear to have quite different effects than distorting a speech signal by means of noise, for instance. I will discuss some examples of experiments (Drullman, 1995; Noordhoek and Drullman, 1997) in which 'deterministic' modulation reduction is compared to noise-induced reduction of the modulations. From these experiments, a simple revision in the calculation of the STI is proposed.

4.2. DISTORTIONS OF THE MODULATION TRANSFER

4.2.1. Limitations with standard measurements

In IEC-standard 60268-16 (1998) the analysis and test signal for measuring the STI and STITEL (a condensed version of the general STI) are given in detail. The same standard gives a number of limitations of the methods, related to the analysis per se and to the test signals employed. A few limitations – mentioned in the standard and elsewhere, cf. Steeneken (1992) – will be briefly discussed below, with examples of STITEL calculations in the case one would nevertheless use the STI method. The manipulations were done on a speech signals (short sentences) to assess their effects on intelligibility and on the STITEL test signal to compute the STI.

Frequency multiplication. This type of distortion may occur when playing a digital signal at the wrong sampling rate or playing from analogue tape at incorrect speed. A change in the sampling rate (up or down) of less than 1% results in a drop in STI from 1.00 to 0.74. Not a major change in a practical sense (going from 'excellent' to 'good/excellent'), but a noticeable effect in the STI. However, when listening to a speech signal, such a minute change in sampling rate is hardly heard and has no effect on intelligibility.

Center clipping. This type of distortion may occur when low-level parts of a signal are not transmitted faithfully (or silenced). This could happen in (partly) broken amplifiers or loudspeakers. To show the effect, two conditions were made, with the clip levels at 10% and 20% below the maximum signal level (i.e., −20 and −14 dB *re* peak). STI drops to 0.69 for the former and to 0.61 for the latter. Although a substantial reduction, this would still be qualified as 'good' and 'fair', respectively. 10% centre clipping results in severely distorted but intelligible speech, 20% clipping is completely unintelligible. So, the STI underestimates the effects of these rather severe examples of centre clipping.

Drop outs. Signal drop out at regular intervals can result from fading patterns in wireless transmissions. In digital systems, this could give rather sharp distortions by which the signals disappears and reappears quite suddenly. In examples with regular drop outs every 100 and 250 ms (by modulating the amplitude by a square wave between 0 and 1) STI values of 0.85 and 0.94 were obtained. Hence, intelligibility is predicted as 'excellent'. Unfortunately, when listening to speech distorted this way, intelligibility is very poor, in the latter case virtually zero.

These are just some examples of distortions where one will have to pay attention to what is measured. Sometimes the prediction of intelligibility by the STI is too high, sometimes it is too low. The effects can be found in the literature, but not all details are known to people in the field working with the STI. Solutions exist for artifacts that are encountered, by smart interpretation of the output of the different analysis steps in the calculation of the STI.

4.2.2. Limitations with direct envelope manipulations

Some years ago, I published a series of experiments in which the envelopes of narrow frequency bands were manipulated directly (Drullman, 1995). Instead of reducing the modulation depth by means of adding noise, either the envelope peaks were lowered or the troughs were raised. This kind of deterministic modulation reduction is not especially new, since it is a variant of signal processing methods known as amplitude compression. These techniques are mostly applied in order to improve speech intelligibility for the hearing impaired, who suffer from a limited dynamic range. The STI will predict a reduction of intelligibility (for listeners with normal hearing) when fast multi-channel amplitude

compression is applied. Fast compression reacts on the instantaneous amplitude envelopes of a range of frequency bands. It is different from automatic gain control, which is a relatively slowly working adaptation of the signal level, with time constants typically in the range of 250–500 ms.

The question is whether the same amount of modulation reduction brought about by adding noise will result in a similar reduction of intelligibility. There are some arguments to say that it will not. For instance, adding noise to the speech signal causes the weaker elements (consonants) to be masked, which is not the case with compression, where this information is preserved. But the STI for direct manipulation and the addition of noise would be the same. In the examples to be discussed below, it will be shown that reduction of the MTF does not always lead to reduced intelligibility. One can create conditions which have equal MTFs (and thus equal STIs), but yield quite different intelligibility scores, or vice versa.

In my experiments from 1995, one further question was whether troughs in the envelope are equally important as peaks. It is assumed that most information is conveyed in the peaks of the speech signal. This can be inferred from the idea that additional noise acts as a sort of 'fence' where only the peaks of the speech can rise above. The higher the noise level, the less speech peaks can be perceived, until eventually the entire speech signal is masked. A processing scheme was used that operated directly on the temporal envelope of each of 24 ¼-oct bands, removing modulations from the peaks or from the troughs. Examples are given in Fig. 1.

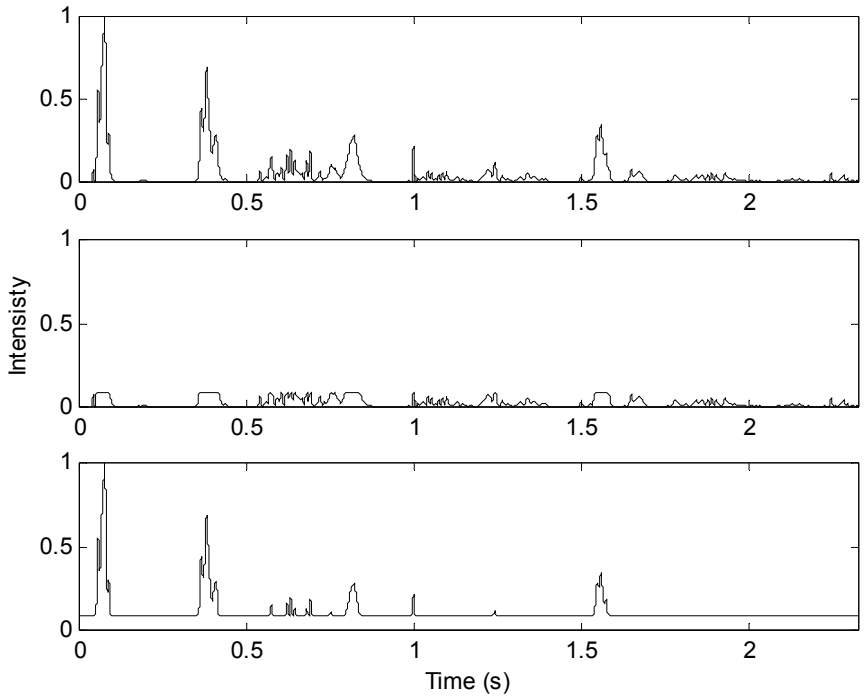


Figure 1. Examples of a narrow-band temporal envelope of a short sentence. Top: original; mid: peaks removed; bottom: troughs removed.

Different experimental conditions were created, as a function of the relative level above or below which peaks or troughs were removed from the envelope. Speech intelligibility of short everyday sentences was measured with normal hearing subjects. Scores were counted

only if the entire sentence was reproduced without a single error. For comparison, a reference set was measured for sentences in speech noise, at different signal-to-noise ratios. Results of the intelligibility measurements are shown in Table I. Scores are displayed as a function of the level above or below which peaks or troughs were removed. This level can be considered to be equivalent to the level of noise (re sentence level) in the reference.

Table I. Intelligibility scores for sentences in different experimental conditions of direct envelope manipulation and regular additive noise.

Equivalent noise level	Intelligibility score		Equivalent noise level	Intelligibility score
	Reference (noise)	Troughs removed		Peaks removed
0 dB	92.5 %	98.0 %	0 dB	99.2 %
5 dB	55.0 %	90.0 %	−5 dB	99.6 %
10 dB	0.0 %	60.4 %	−10 dB	97.5 %

As can be seen in Table I, intelligibility drops considerably in the reference condition as the noise level increases. Scores are in good agreement with predictions by the STI method. Intelligibility is much less affected by the other processing methods. To get an estimate of the modulation reductions, the MTF was measured for each of five octave bands with centre frequencies at 0.25, 0.5, 1, 2, and 4 kHz. MTF measurements were performed on the long-term envelope spectra (processed in ¼-oct bands versus unprocessed) of a 71-s speech fragment. For each octave band the mean modulation reduction (m) in the range 0.5-20 Hz was computed. The average MTF was computed using the weightings from Steeneken and Houtgast (1980). The results are shown in Table II.

Table II. Average MTFs for the experimental conditions of Table I.

Equivalent noise level	Average MTF		Equivalent noise level	Average MTF
	Reference (noise)	Troughs removed		Peaks removed
0 dB	0.52	0.49	0 dB	0.33
5 dB	0.26	0.23	−5 dB	0.24
10 dB	0.10	0.11	−10 dB	0.19

Table II shows a close correspondence between the respective MTFs in the second and third column. However, this correspondence is missing in the intelligibility scores in Table 1. For example, the MTF values of 0.11 and 0.10 for the conditions with troughs removed and the speech+noise reference, correspond to intelligibility scores of 60% and 0%, respectively. A low MTF of 0.19 for the peaks removed condition even corresponds to almost perfect (97%) intelligibility. In general, for the above type of signal manipulations, no one-to-one relation between the MTF (and hence the STI) and the intelligibility scores can be established. Therefore, use of the STI cannot simply be extended to any manipulation of the temporal speech envelope. The same conclusion is drawn by studies investigating multichannel dynamic compression or expansion (cf. Hohmann and Kollmeier, 1995; Van Buuren et al., 1998).

4.3. DISCUSSION AND CONCLUSION

So, why would equal MTFs (and STIs) not lead to equal intelligibility? In the case of deterministic modulation reduction versus modulation reduction by additive noise, there may be two explanations. First, noise introduces nonrelevant modulations (in narrow bands), which leave the listener with a 'sorting problem': he is unable to separate the relevant speech modulations from the spurious noise modulations. Second, noise affects the speech fine structure.

Following the experiments I performed in 1995, Ingrid Noordhoek did some experiments with 'pure' uniform modulation reduction, by proportionally raising the troughs and lowering the peaks of the intensity envelope (Noordhoek and Drullman, 1997). An example is given in Fig. 2. The figure clearly shows the difference in the envelopes after deterministic modulation reduction (mid panel) and after adding noise (bottom panel). In either case the modulation reduction factor is the same, m=0.5.

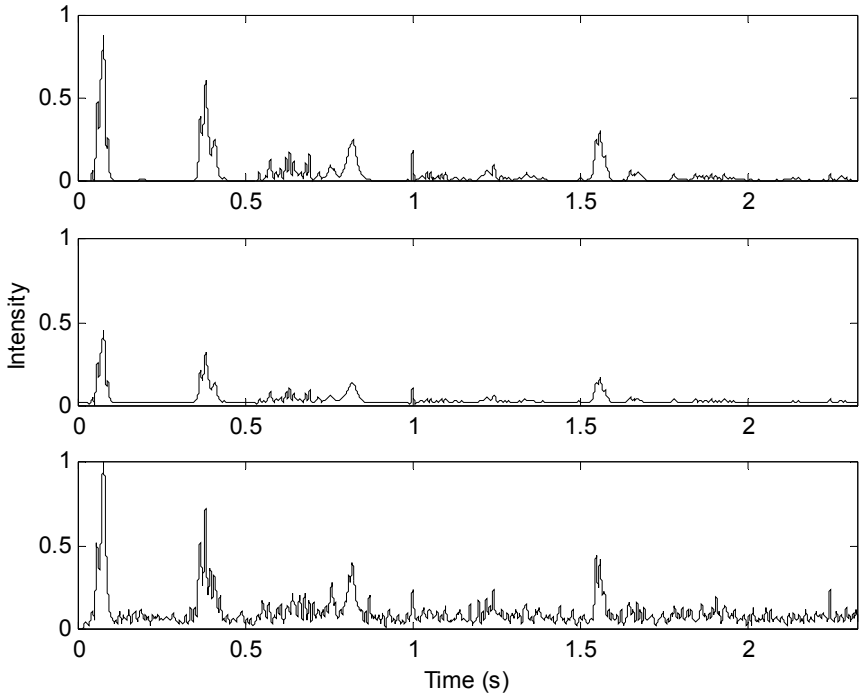


Figure 2. Example of uniform deterministic modulation reduction in a narrow band temporal envelope of a short sentence. Top: original; mid: 50% deterministic modulation reduction (m=0.5); bottom: additive noise at 0 dB speech-to-noise ratio.

For a number of deterministic modulation-reduction factors (m_{det}), the speech reception thresholds (SRT) for sentences in noise was measured with normal-hearing subjects (Plomp and Mimpen, 1979). The SRT defines the speech-to-noise ratio at which 50% of short everyday sentences can be reproduced correctly. Intelligibility is measured against a steady-state speech noise, and the level of the sentences is changed according to a simple up-down adaptive procedure. Results showed SRTs varying from +6.4 dB to −4.3 dB for m_{det} ranging from 0.2 to 1.0, the latter condition being the reference for unprocessed speech in noise.

When measuring the SRT in noise, modulations in the temporal envelope are reduced by the combined effects of the imposed deterministic modulation reduction and the addition

of noise. The modulation reduction brought about by the noise (m_{noise}) depends on the speech-to-noise ratio at a given SRT. The combined modulation reduction at the threshold (m_{thr}) is thus defined as

$$m_{thr} = m_{det} \cdot m_{noise} \tag{1}$$

When deterministic and noise-induced modulation reduction are equally detrimental to speech intelligibility, m_{thr} has a constant value for different values of m_{det} . But the results from the SRT measurements showed that the effects are not the same. The effectiveness of deterministic modulation reduction is a constant fraction of the effectiveness of noise-induced modulation reduction. For the same intelligibility, this means

$$1 - m_{noise} = C \cdot (1 - m_{det}), \qquad 0<C<1 \tag{2}$$

The limit threshold for unprocessed speech (m_{det} = 1) yielded an SRT of −4.3 dB, corresponding to m_{noise} = 0.27. Without going into details here (see Noordhoek and Drullman, 1997), the threshold modulation reduction can be expressed as

$$m_{thr} = 0.27 \cdot m_{det} \cdot (1 - C \cdot (1 - m_{det}))^{-1} \quad , \tag{3}$$

where C = 0.81 is found to be the best fit of the data. Threshold modulation reduction (50% intelligibility) as a function of deterministic modulation reduction is displayed in Fig. 3.

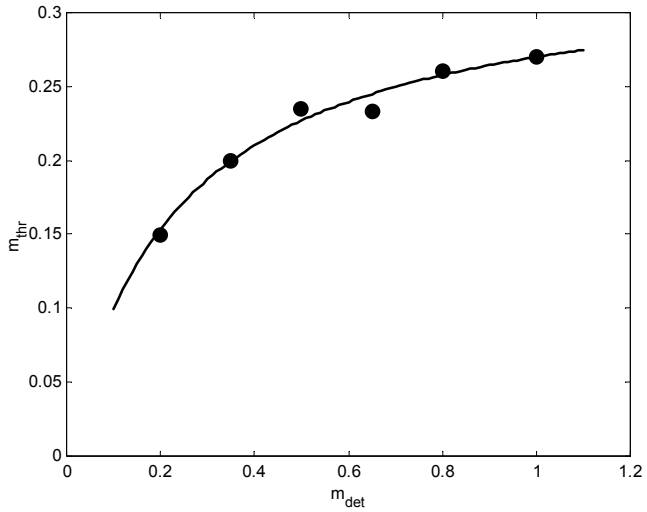


Figure 3. Threshold modulation reduction (m_{thr}) as a function of deterministic modulation reduction (m_{det}). Dots are derived from SRT measurements; the curve is a least-squares fit tot the data according to Eq. (3).

When no noise is added, the threshold (50% intelligibility) for complete deterministic modulation reduction equals 0.10. This result agrees very well with measurements of deterministic modulation reduction for speech in quiet, where $m_{thr} = m_{det} = 0.11$.

When considering the constant C = 0.81, this can be interpreted that 81% of the effect of noise on intelligibility can be explained by the reduction of relevant speech modulations.

For the remaining 19%, other factors have to be considered: the introduction of spurious noise modulations and the corruption of the speech fine structure.

In conclusion, the origin of the modulation reduction appears to be a crucial factor in predicting speech intelligibility. When calculating the STI, modulation reductions are expressed in apparent signal-to-noise ratios, irrespective of the origin of the distortions (IEC 60268-16, 1998):

$$SNR_{app} = 10 \cdot \log(m/(1-m)) \quad (4)$$

This equation does not hold for deterministic modulation reduction. A simple revision of the formula would make it

$$SNR_{app} = 10 \cdot \log(1-C \cdot (1-m)/C \cdot (1-m)), \quad (5)$$

where C equals 1 in case of noise and 0.81 in case of (uniform) deterministic modulation reduction.

REFERENCES

Drullman, R. **(1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585-592.

Hohmann V. and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J. Acoust. Soc. Am. **97**, 1191-1195.

Noordhoek, I.M. & Drullman, R. (**1997**). "Effect of reducing temporal intensity modulations on sentence intelligibility," J. Acoust. Soc. Am. **101**, 498-502.

IEC 60268-16 (**1998**). *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index* (International Electrotechnical Commission, Geneva, Switzerland).

Plomp, R. and Mimpen, A.M. (**1979**). "Improving the reliability of testing the Speech Reception Threshold for sentences,", Audiology **18**, 43-52.

Steeneken, H.J.M. (**1992**). *On measuring and predicting speech intelligibility* (Doctoral dissertation, University of Amsterdam).

Steeneken, H.J.M. & Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318-326.

Van Buuren, R.A., Festen, J.M., and Houtgast, T. (**1998**). "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," J. Acoust. Soc. Am. **105**, 2903-2913.

Chapter 5. Application of the Speech Transmission Index to the Hearing Impaired

Joost M. Festen and Reinier Plomp[a]

Department of Otolaryngology, VU medical center, Amsterdam
[a]*Emeritus professor of the Vrije Universiteit, Amsterdam*

5.1. INTRODUCTION

Commonly, hearing-impaired listeners are expected to need a louder signal than normal-hearing listeners, but more often they need a signal of higher quality, i.e. a signal with less degradation. In terms of masking they need a better S/N ratio, reflected in a higher Articulation Index, AI (ANSI, 1969) or Speech Intelligibility Index, SII (ANSI, 1997). In terms of spectro-temporal details they need a larger part of the speech variability, reflected in a higher Speech Transmission Index, STI (Houtgast and Steeneken, 1973). While STI = 0.5 may be 'fair' in general, for some hearing-impaired listeners this may be far too poor. In recent years several attempts were made to correct for this disparity among listener groups by including the characteristics of the ear as a part of the transmission channel. A first and obvious step was to account for audibility by the introduction of an equivalent internal noise (Pavlovic, 1984). Good results were obtained for listeners with minor impairment, but for the larger impairments in general performance was less than predicted by the modified AI. To account for the progressively poorer results with increasing hearing loss Pavlovic et al. (1986) introduced two additional factors in AI: a speech desensitisation factor [a fixed function of hearing loss to account for reduced auditory processing capabilities] and a listening proficiency factor to account for differences in individual listening skill. On average for a large group of listeners this procedure gave adequate results, but for individual hearing-impaired listeners a lot of variance remained for the proficiency factor. This was unsatisfactory because in this way the index required an individual correction that was based only on an actual speech-intelligibility test.

In predicting intelligibility, STI seems more powerful than AI or SII because it uses a more general approach towards speech interference. However, when it comes to predicting intelligibility for hearing-impaired listeners, all three indices need similar adjustments, one for loss of sensitivity and an additional correction for speech-processing deficits. In the calculation of STI for hearing-impaired listeners Ludvigsen (1987) proposed a correction for reduced temporal resolution. More recently, Holube and Kollmeier (1996) measured spectral and temporal resolution for individual listeners and used the results in a perception model to predict intelligibility. The results were compared to predictions from AI and STI. Because only minor improvements were obtained from the individually measured spectral and temporal parameters, Holube and Kollmeier concluded that suprathreshold processing deficiencies played only a minor role in their group of listeners.

In the present paper a procedure will be presented to distinguish between hearing-impaired listeners with and without suprathreshold deficiencies. Additionally, two

applications of STI for hearing-impaired listeners will be discussed. In these examples hearing-impaired listeners need a higher-than-normal STI but, as for normal hearing, the transmission of modulations is the factor on which various conditions are comparable.

5.2. SIMPLE VERSUS COMPLEX HEARING IMPAIRMENT

An elevated Speech Reception Threshold (SRT, i.e. level for 50% intelligibility of short sentences) in noise does not necessarily imply that the listener suffers from suprathreshold processing deficits. For hearing-impaired listeners, with sometimes steep audiograms, at least we have to take care that all parts of the speech are above threshold.

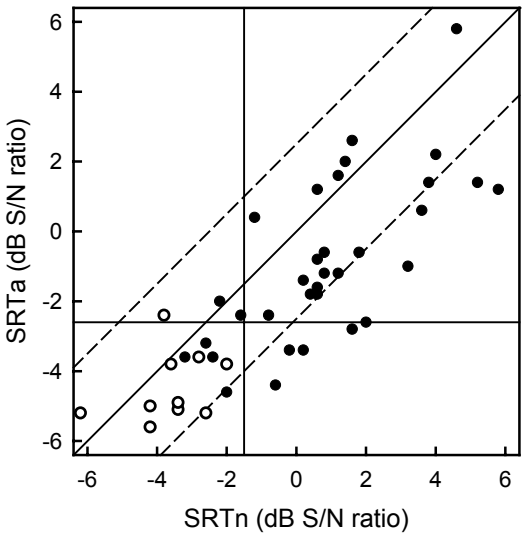


Figure 1. Speech-reception threshold in noise for signals with a spectrum adapted to the hearing loss (SRTa) versus the condition with unmodified signals (SRTn) for normal-hearing listeners (*open circles*) and hearing-impaired listeners (*closed circles*). Solid horizontal and vertical lines represent the one-tailed 95% confidence limit for the normal-hearing data. For data between the dashed lines SRTa and SRTn are not significantly different ($p<0.05$).

Fig. 1 shows a scatter diagram of two types of speech-reception threshold, SRTn measured in noise at a level 20 dB above the SRT in quiet and the SRTa measured after careful correction of the speech and noise spectra for the individual audiogram, such that all relevant speech information from 250 - 4000 Hz is above threshold (Noordhoek et.al., 2000). As expected SRTa is lower than SRTn for nearly all listeners, and even when higher the difference is statistically not significant. The figure also shows that, although spectral adaptation lowers the SRT, still hearing-impaired listeners have in general higher thresholds than normal-hearing listeners. To a first approximation there are two possible grounds for this phenomenon. First, despite all efforts the listener may not have available all of the speech dynamic range. Parts of the signal may be below threshold, other parts may be too loud causing level distortion or, in particular for steep spectral slopes, parts of the speech may be affected by auditory masking as found in normal-hearing listeners. If these are the only causes for the elevated SRT, i.e. the ear performs a normal processing of the available signal, we call this *simple* hearing loss. A second possible cause for the elevated SRT is a reduced suprathreshold processing capability of the ear. In contrast to the first category we call this *complex* hearing loss.

To distinguish between these two types of hearing loss we use the Speech Intelligibility Index (SII). The index may be interpreted as the proportion of the total speech information available to the listener. For normal-hearing listeners, all conditions of equal intelligibility should give the same SII. Thus, when the measured SRTn and SRTa are expressed in SII values, and the SII model is consistent with the results of normal hearing listeners, the model will yield identical SII values for the various tests. Then, an elevated SII at 50% intelligibility for a hearing-impaired listener (thus, including the audiogram in the SII calculation) can be considered an indication of deteriorated suprathreshold speech processing, because a listener with a higher-than-normal SII needs more speech information than normal-hearing listeners to reach the 50% intelligibility score.

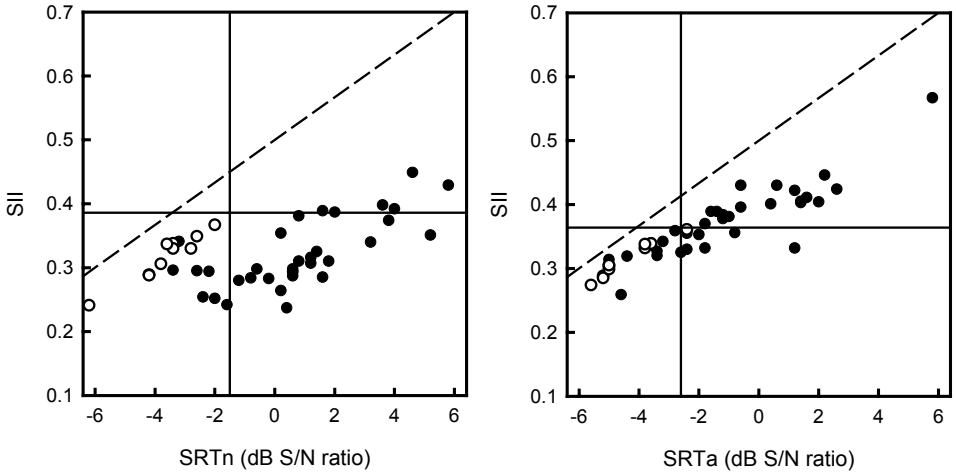


Figure 2. Speech intelligibility index versus the thresholds on two intelligibility tests (SRTn and SRTa) for normal-hearing listeners (*open circles*) and hearing-impaired listeners (*filled circles*). The solid lines represent the one-tailed 95% confidence limit for the data of the normal-hearing listeners. Dashed lines represent the maximum possible SII as a function of SRT, i.e. the SII that would have been calculated if the audibility of the speech had not been affected by the hearing threshold, upward spread of masking, and level distortion.

SII calculated for the data shown in Fig. 1 is plotted as a function of SRTn and SRTa in Fig. 2. For normal-hearing listeners comparable SII values are obtained from these tests. However, for the hearing-impaired listeners the mean SII for the SRTa is *higher* than for SRTn, although their mean SRTa (-1.0 dB) is 1.8 dB *lower* their mean SRTn (+0.8 dB). This means that by adapting the spectrum, the SRT in noise did not decrease as much as predicted from audibility. A possible explanation for SII variation across conditions is that the suprathreshold deficit is frequency dependent. For example, in frequency regions where the hearing loss is greater, the suprathreshold deficit may be larger. In the standard condition listeners only use the frequency region where the original speech spectrum is above their hearing threshold. In the SRTa condition the entire frequency range from 250 to 4000 Hz should contribute. If a suprathreshold deficit is present in the frequency region that is below threshold in the standard condition but above in the SRTa condition, this could explain the observed effects.

The horizontal and vertical lines in Fig. 2 (95% confidence limits) can be considered as separation between normal and impaired results. These lines divide the graphs in four quadrants. No data fall in the upper-left quadrant. Data points in the lower-left quadrant correspond to a normal threshold. Data points in the lower-right quadrant correspond to an

elevated SRT, but the elevation can be explained on the basis of audibility. Finally, data points in the upper-right quadrant correspond to elevated thresholds that cannot be explained by the SII model. The results on SRTn yield a rather constant SII. Only six of the listeners performed slightly poorer than predicted based on audibility. SII calculations show that the hearing threshold is the primary cause for the difference between the individual SII and the dashed line. In contrast many of the elevated thresholds obtained for SRTa cannot be explained by audibility.

Apparently, a speech-reception test with signals corrected for audibility is more sensitive to suprathreshold deficits than the standard test. Calculation of SII from SRTa offers the possibility to distinguish between listeners with simple and complex hearing loss. For nearly half of the hearing-impaired listeners in Fig. 2 the elevated SII for the SRTa test shows a suprathreshold speech-processing deficit. Currently, research on the understanding of auditory processing deficits and on signal processing for compensation is focused on these listeners.

5.3. TRADE-OFF BETWEEN NOISE AND REVERBERATION FOR LISTENERS WITH HEARING LOSS

In contrast to SII, STI seems particularly designed for predicting speech intelligibility also in conditions with temporal distortions, like reverberation or amplitude compression. In these conditions the transmission of modulations, i.e. the basis of the STI model, plays an essential role in defining speech quality. If we accept that hearing-impaired listeners need a better STI than normal-hearing listeners, we can still ask whether the transmission of modulations determines intelligibility in conditions with, for instance, noise and reverberation. To investigate whether the STI model holds in these conditions for hearing-impaired listeners, Duquesnoy and Plomp (1980) conducted an experiment measuring speech reception in noise under conditions with reverberation for a large and heterogeneous group of elderly hearing-impaired listeners. Before data collection the listeners were divided in subgroups according to the maximum amount of reverberation they could tolerate in speech perception. With an adaptive procedure the SRT in noise was measured for a number of conditions with fixed reverberation.

The average results for five groups of listeners are plotted in Fig. 3 as limit conditions for mixed amounts of noise and reverberation. The smooth curves represent iso-STI contours. Clearly, the different groups of listeners are characterised by a minimum STI they can tolerate. Within groups, the threshold S/N ratio for different reverberation conditions essentially corresponds to a fixed STI value. Thus for these groups of elderly listeners and for this speech material the signal distortions by noise and reverberation can be interchanged as described by the STI model. Different groups of listeners can be characterised simply by different limiting values of STI.

5.3.1. Matching acoustic job environment and individuals

If we assume that the trade-off between noise and reverberation as found for elderly by Duquesnoy and Plomp holds also for other groups of hearing-impaired listeners, we can use STI to anticipate conditions in which problems with speech intelligibility will be encountered. For instance, such an approach can be used in counselling on employability for hearing-impaired people. If the SRT in noise is known as well as the STI of a possible job environment, a prediction can be made of the individual speech intelligibility in that condition and thus of problems that may arise, depending on the type of work.

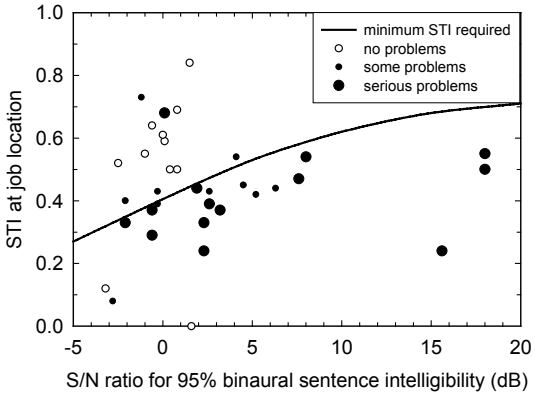


Figure 3. Iso-STI contours for conditions of steady-state speech-spectrum noise combined with reverberation. The interconnected data point series are mean speech reception thresholds for sentences measured from one-microphone recordings in a diffuse sound field for normal-hearing listeners (*open symbols*) and four groups of hearing-impaired listeners (*solid symbols*). Groups consist of between 15 and 30 listeners. Data adopted from Duquesnoy and Plomp (1980).

Fig. 4 shows some results for 35 hearing-impaired people in different job environments. For each listener the binaural SRT in noise was measured in a diffuse sound field with reverberation time T=0.5 seconds. For comfortable listening 2 dB is added to the SRT to achieve a level at which about 95% of sentences are understood without error. From these data a minimum required STI can be calculated. Additionally, STI was measured at the job location and the listeners filled in a questionnaire on problems with intelligibility encountered in their work. With only a few exceptions the minimum requited STI plotted in Fig. 4 gives the separation between people who encounter problems and those who don't.

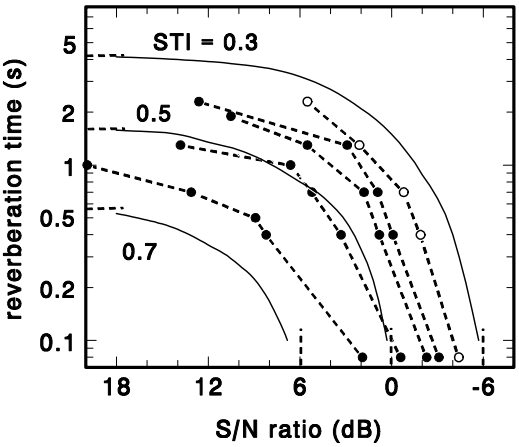


Figure 4. STI values measured at job locations for 35 hearing-impaired people plotted against the signal-to-noise ratio needed by these individuals for 95% binaural sentence intelligibility (binaural SRT + 2 dB in a T=0.5 s environment). The curved line represents the minimum STI required, i.e. the STI associated with the S/N ratio on the abscissa and the reverberation. Various symbols are used to indicate the level of problems experienced with speech intelligibility while at work.

5.4. SYLLABIC COMPRESSION IN HEARING AIDS

Fast-acting amplitude compression in hearing aids is another area were the transmission of modulations is at stake. Sensorineural hearing impairment is often accompanied by an abnormally steep increase of loudness with stimulus level, called recruitment. To restore normal loudness perception and to fit speech in the limited dynamic range of the impaired ear many hearing aids apply some form of amplitude compression. In this processing the level variations in dB above the compression onset level (knee point) are reduced by a factor (the compression ratio). Modern hearing aids often use multi-band compression, because both the needs of the ear and the time-varying demands of the signal are a function of frequency. However, the effects of fast-acting, or syllabic, compression on intelligibility as measured in various studies (see Dillon, 1996) are equivocal and range from negative effects to no effect. These confusing results may be a consequence of the strong dependence of compression effectiveness on its time constants and on the bandwidth of the compression channels.

5.4.1. Intelligibility of amplitude compressed speech

The intelligibility of amplitude-compressed speech in quiet was measured for 16 normal-hearing listeners and 16 listeners with sensorineural hearing loss (Festen and van Dijhuizen, 1998). The hearing-impaired listeners had flat to moderately sloping audiograms with an average hearing loss of 55 dB and a residual dynamic range of about 45 dB. Short sentences (250 - 4000 Hz) were amplitude compressed with compression ratios of 1, 2, 4, and 4 in a single wideband channel or after subdivision in 2, 4, 8, or 16 frequency channels. The compression time constants were very short with an attack time of a few milliseconds and a release time of 20 ms (*syllabic compression*). The compression onset was at a level 30 dB below the average speech level (*full-range compression*). The modified speech was presented to the normal-hearing listeners at 75 dBA and to the hearing-impaired listeners halfway their individual dynamic range. The average intelligibility scores for both groups are presented below. Apart from a ceiling effect for the normal-hearing group, these results show a gradual decline of intelligibility with increasing compression ratio and with an increasing number of channels.

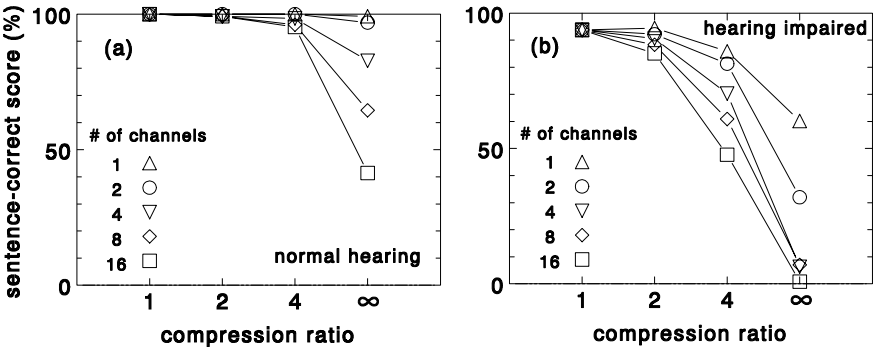


Figure 5. Intelligibility scores for sentences with syllabic compression as a function of compression ratio. The conditions range from single-channel compression (4 octaves) to compression in 16 quarter-octave channels. Panel (a) for normal-hearing listeners and Panel (b) for hearing-impaired listeners. Data from Festen and van Dijkhuizen (1998).

5.4.2. Modulation transfer

For transmission channels with noise and reverberation the quality of speech transmission can be very well expressed by the transfer of modulations. Because noise and reverberation mainly reduce modulations, the integrity of the modulations after transmission can be estimated sufficiently from an evaluation of the attenuation of modulations. With amplitude compression the situation is different: modulations may not only be attenuated, but also new modulations may be added. An example is given in Fig. 6. Panel (a) shows the envelope of a sentence in an octave frequency band before and after compression. In the compressed envelope the modulations are a reduced version of the input modulations. Panel (b) shows envelopes from the same octave band, only now the compressor was fed with the wideband signal. Two effects are apparent: first, the overall modulation reduction in panel (b) is much less than in panel (a), and second, the shape of the envelope has considerably changed. These new modulations are introduced because the compressor control signal contains modulations different from those in the observed frequency band.

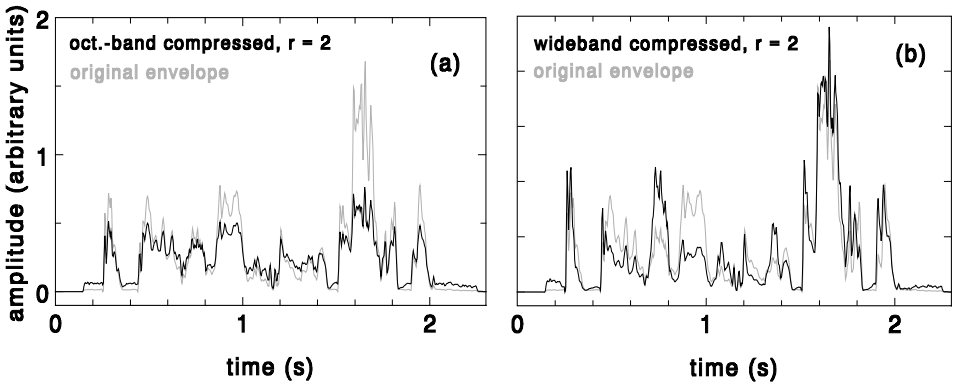


Figure 6. Envelope of a single sentence in an octave band from 2-4 kHz. Panel (a) shows the original envelope (light) and the envelope after octave-band compression (bold). Panel (b) shows the original envelope again together with the envelope after wideband compression (bold).

Modulations may be important for intelligibility, but it seems reasonable to regard modulations without a relation to the modulations in the input signal as spurious. For speech intelligibility it is assumed that only those modulations contribute that are in phase with input modulations. Therefore, when evaluating syllabic compression in terms of modulation transfer we focus on phase-locked modulations. In the calculations presented below modulations are taken from quarter-octave frequency bands. The phase-locked modulation transfer was calculated for a number of compression conditions from the intelligibility experiment. The procedure for calculating the phase-locked modulation transfer was a quarter-octave version of the calculations described by Drullman *et al.* (1994)[1]. The calculations were performed on 32 concatenated sentences with amplitude compression. The original and compressed signals were split up in 16 quarter octaves and envelopes were determined.

The phase-locked modulation transfer function is defined as the cross-spectral density of these envelopes relative to the input-envelope auto-spectral density. A single transfer index is obtained by unweighted averaging over the relevant modulation frequencies from 0.5 to 20 Hz and a weighting over frequency bands according to the ANSI SII (1997). In Fig. 7 the

[1] Note by the editor: also see chapter 4 of this book

intelligibility scores for wideband, octave-band, and ¼-oct.-band compression from the above experiments are plotted as a function of phase-locked modulation transfer. In this figure the different curves from Fig. 5 are, for each group of listeners, reduced to a single curve relating speech intelligibility and temporal modulation transfer.

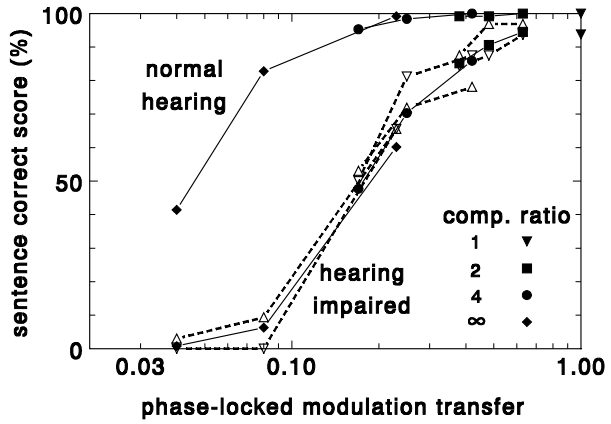


Figure 7. Intelligibility scores as a function of phase locked modulation transfer for sentences processed with syllabic compression. Different compression ratios are indicated with different symbols. For each compression ratio calculations were performed for the single-band, four band and 16-band conditions yielding gradually poorer modulation transfer. Open triangles show average results for subgroups of the hearing-impaired listeners; up-pointing triangles for four listeners with a dynamic range wider than 55 dB, down-pointing triangles for four listeners with a dynamic range narrower than 40 dB.

For both listener groups Fig. 7 shows a gradual decrease of intelligibility with decreasing modulation transfer. The only difference between the two groups of listeners is that, for the same score, the hearing-impaired group needs a better preservation of the modulations then the normally hearing group. If positive effects of syllabic amplitude compression were to be expected, it would be for listeners with a narrow dynamic range. Therefore, from the 16 hearing-impaired listeners in Fig. 7 the data of four listeners with the narrowest dynamic range and four with the widest range are plotted separately. As the figure shows, the differences with the average data are extremely small.

5.5. SPEECH RECEPTION IN FLUCTUATING NOISE

While for some distortions that affect the signal temporal envelope STI and MTF give an adequate prediction of speech intelligibility, for other, quite common, conditions these predictions fail completely. If speech is masked by fluctuating or modulated noise the SRT for normal-hearing listeners drops for speech-like modulations from about –5 dB for steady noise to thresholds below –10 dB (Festen and Plomp, 1990). This strong effect on SRT occurs while both the masker spectrum an the average masker intensity are constant. STI calculations are therefore not affected by these masker modulations. Also a possible redefinition of STI based on envelope correlations or phase-locked modulation transfer will not solve the problem. In contrast to the data, predictions from such an approach will be that modulated noise affects intelligibility even more than unmodulated noise.

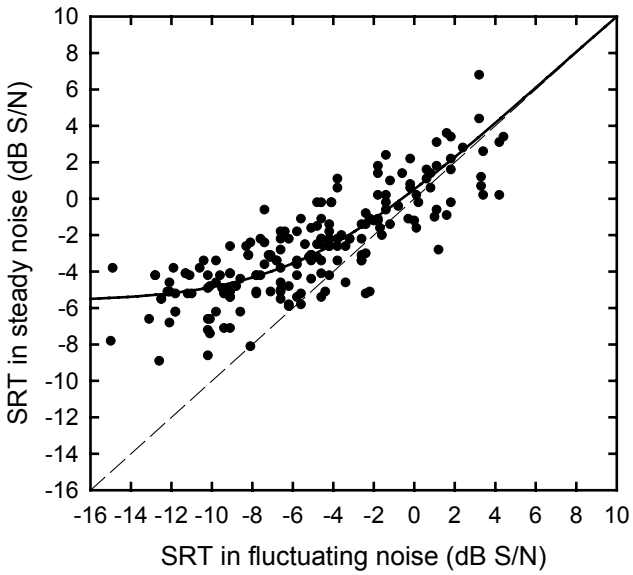


Figure 8. Scatter diagram of speech-reception thresholds for 194 ears measured in steady noise and in fluctuating noise, both with a speech-like spectrum. The fluctuating noise consists of two frequency bands, with 1000-Hz as cross-over frequency and each with modulations from the corresponding frequency band of running speech.

Hearing-impaired listeners further complicate the situation because with increasing hearing loss the advantage of masker modulations gradually disappears. Fig.8 shows a scatter diagram of speech-reception thresholds measured in steady noise and in fluctuating noise for 194 ears ranging from normal to impaired. The best performing ears have a speech-reception threshold of –6 dB in unmodulated noise and about –12 dB in modulated noise. However, for listeners with an elevated SRT the advantage of masker fluctuations is quickly lost, while the threshold in unmodulated noise rises more gradually. Because a fluctuating interference, like speech, is so common in everyday conditions there is a real need for an extension of STI towards conditions with a time-varying background.

5.6. DISCUSSION

Functionally, hearing impairment can be divided in two categories: simple and complex. A listener with simple hearing impairment can function as normal hearing after a correction for the elevated hearing threshold. To put it in a different way, this hearing loss can be modelled with an equivalent masking noise. For these ears the speech variations that are presented above threshold are processed just as efficient as in a normal ear. When the equivalent masking noise is included in the STI or SII calculations, this leads to normal values. For listeners with complex hearing impairment speech becomes intelligible only if more of the speech variability is above threshold than for normal hearing. This indicates that for these listeners the speech-processing capabilities are reduced. Examples are a poor spectral or temporal resolution.

In terms of the speech transmission index listeners with a hearing loss need a better STI than normal hearing listeners. But even for this higher STI the trade-off between noise and reverberation as predicted by the model seems to hold. Also the reduction of intelligibility

with different parameters of fast amplitude compression, like bandwidth and compression ratio, can be adequately described in term of modulation transfer, provided that not only the strength of the modulations is accounted for but also their timing or phase (phase-locked MTF).

Serious problem arise for the STI model when the masker is a fluctuating noise or a interfering voice. In these conditions the SRT for normal-hearing listeners drops by 6-8 dB compared to the SRT in steady noise. Unfortunately, the current STI model is not designed for these very common conditions. For a correct development of a modified STI that incorporates time-varying masking effects more data on the effects of modulated maskers need to be available.

REFERENCES

ANSI (**1969**). ANSI S3.5-1969, "American national standard methods for the calculation of the articulation index" (American National Standards Institute, New York).

ANSI (**1997**). ANSI S3.5-1997, "American national standard methods for calculation of the speech intelligibility index" (American National Standards Institute, New York).

Dillon, H. (**1996**). "Compression? Yes, but for low or high frequencies, for low or high intensities, and with what response times?", *Ear & Hearing* **17**, 287-307.

Drullman, R., Festen, J.M. and Plomp R. (**1994**). "Effect of reducing slow temporal modulations on speech reception", *J. Acoust. Soc. Am.* **95**, 2670-80.

Duquesnoy, A.J. and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbycusis", *J. Acoust. Soc. Am.* **68**, 537-44.

Festen J.M. and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing", *J. Acoust. Soc. Am.* **88**(4), 1725-36.

Festen, J.M. and van Dijkhuizen, J.N. (**1998**). "Modeling the speech-reception threshold for amplitude-compressed speech", In: *Psychophysics, Physiology and Models of Hearing.* edited by: T. Dau, V. Hohmann and B. Kollmeier, pp 249-54, World Scientific, Singapore, ISBN 981-02-3741-3.

Houtgast, T. and Steeneken, H.J.M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acoustica* **28**, 66-73.

Holube, I. and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impared listeners based on a psychoacoustically motivated perception model", *J. Acoust. Soc. Am.* **100**(3), 1703-16.

Ludvigsen, C. (**1987**). "Prediction of speech intelligibility for normal-hearing and cochlearly hearing-impaired listeners", *J. Acoust. Soc. Am.* **82**, 1162-71.

Noordhoek, I.M., Houtgast, T. and Festen, J.M. (**2000**). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. II. Hearing-impaired listeners", *J. Acoust. Soc. Am.* **107**, 1685-96.

Pavlovic, C.V. (**1984**). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment", *J.Acoust.Soc.Am.* 75(4), 1253-8.

Pavlovic, C.V., Studebaker, G.A., and Sherbecoe, R.L. (**1986**). "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals", *J. Acoust. Soc. Am.* **80**, 50-7.

Chapter 6. Implementation of intelligibility algorithms into acoustic simulation programs

Wolfgang Ahnert
ADA Acoustic Design Ahnert, Berlin, Germany

ABSTRACT

In acoustic simulation programs very different algorithms are used to calculate the intelligibility of speech and music. The desired results are obtained by post-processing fixed energy ratios as well as time-depending impulse responses. As an example, in EASE 4.0 there have been implemented now all usual intelligibility measures, derived by means of simulated high resolution data or by applying statistical estimations. This paper compares all these measures and methods like STI, AL_{cons}, definition etc., by means of the results obtained within a common computer model. Recommendations for applications will be given.

6.1. INTELLIGIBILITY MEASURES USED FOR COMPARISONS

6.1.1. Definition Measure C_{50} for Speech

The *definition measure* C_{50} describes the intelligibility of speech and also of singing. It is generally calculated in a bandwidth of 4 octaves between 500 Hz and 4000 Hz from the tenfold logarithm of the ratio between the sound energy arriving at a reception measuring position up to a delay time of 50 ms after the arrival of the direct sound and the following energy:

$$C_{50} = 10\lg(\frac{E_{50}}{E_{\infty} - E_{50}})dB \quad (1)$$

the intelligibility of speech is presumed to be good when

$$C_{50} \geq 0 \text{ dB}$$

The frequency-dependent definition measure C_{50} (Ahnert, 1975) should increase by approx. 5 dB with octave center frequencies above 1000 Hz (starting with the octave center frequencies 2000 Hz, 4000 Hz and 8000 Hz), and decrease by this value with octave center frequencies below 1000 Hz (octave center frequencies 500 Hz, 250 Hz and, 125 Hz).

An equivalent, albeit less used criterion is the *degree of definition D*, also called D_{50}, that results from the ratio between the sound energy arriving at the reception measuring position up to a delay time of 50 ms after the arrival of the direct sound, and the overall energy (given in %; Thiele, 1953).

The correlation with the definition measure C_{50} is determined by the formula

$$C_{50} = 10\lg\left(\frac{D_{50}}{1 - D_{50}}\right)dB \quad (2)$$

6.1.2. Speech Transmission Index STI

The determination of the STI-values is based on measuring the reduction of the signal modulation between the location of the sound source, e.g. on stage, and the reception measuring position with octave center frequencies from 125 Hz up to 8000 Hz.

Houtgast and Steeneken (1985) proceeded on the assumption that not only reverberation and noise reduce the intelligibility of speech, but generally all external signals or signal changes that occur on the path from source to listener. For ascertaining this influence they employ the *modulation transmission function* (MTF) for acoustical purposes. The available useful signal S (signal) is put into relation with the prevailing interfering signal N (noise). The *modulation reduction factor* m(F) determined this way is a factor that characterizes the interference with speech intelligibility[1]:

$$m(F) = \frac{1}{\sqrt{1+(2\pi F \cdot RT/13{,}8)^2}} \cdot \frac{1}{1+10^{-(\frac{S/N}{10dB})}} \tag{3}$$

with F modulation frequency in Hz,
RT reverberation time in s,
S/N signal/noise ratio in dB.

To this effect one uses modulation frequencies from 0.63 Hz to 12,5 Hz in third octaves. In addition, the modulation transmission function is subjected to a frequency weighting (WMTF - weighted modulation transmission function) in order to achieve a complete correlation to speech intelligibility. In doing so, the modulation transmission function is divided into 7 frequency bands which are each modulated with the modulation frequency (14). This results in a matrix of 7 x 14= 98 modulation reduction factors m_i.

For obtaining the STI value, the (apparent) effective signal-noise ratio X can afterwards be calculated from the modulation reduction factors m_i:

$$X_i = 10\lg\left(\frac{m_i}{1-m_i}\right)dB \tag{4}$$

(To get RASTI you have to consider only 9 m_i values in the 500 and 2000 Hz bands in this calculation).
According to definition you get, after averaging, the STI-value according to:

$$STI = \frac{X+15}{30} \tag{5}$$

[1] Note by the editor: Eq. (4) only gives the modulation reduction factor correctly under the assumption of pure exponential decay. Reverberation time RT must be the early decay time. In more complex cases (composite decay curves, echoes) the modulation reduction factor must be derived directly from the (squared) impulse response. See [Houtgast, T., Steeneken, H.J.M, and Plomp, R. (**1980**). Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics, Acustica **46**, 60-72] for more details.

Based on the comparison of subjective examination results with a maximum possible intelligibility of syllables of 96%, the RASTI-values are graded in subjective values for syllable intelligibility according to Table I.

Table I. Relation between STI and syllable intelligibility

Syllable intelligibility	*STI-value*
poor	0 to 0.3
satisfactory	0.3 to 0.45
good	0.45 to 0.6
very good	0.6 to 0.75
excellent	0.75 to 1.0

6.1.3. Articulation Loss AL_{cons} with speech

Peutz (1971) and Klein (1971) have ascertained that the articulation loss of spoken consonants AL_{cons} is decisive for the evaluation of speech intelligibility in rooms. Starting from this discovery they developed a criterion for the determination of intelligibility:

$$Al_{cons} \approx 0{,}652\left(\frac{r_{QH}}{r_H}\right)^2 \cdot RT\% \tag{6}$$

r_{QH}: Distance sound source - listener
r_H: Reverberation radius or, in case of directional sound sources, critical distance r_R
RT: Reverberation time in s

Assigning the results to speech intelligibility yields:

$AL_{cons} \leq 3\%$	ideal intelligibility,
$AL_{cons} = 3$ to 8%	very good intelligibility,
$AL_{cons} = 8$ to 11%	good intelligibility,
$AL_{cons} > 11\%$	poor intelligibility,
$AL_{cons} > 20\%$	worthless intelligibility (limit value 15%).

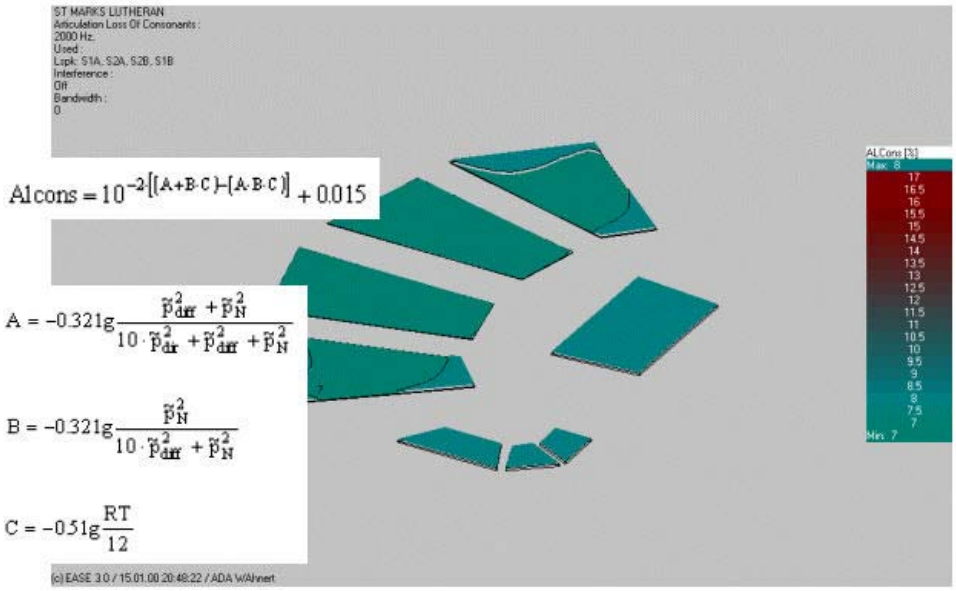
Long reverberation times entail an increased articulation loss. With the corresponding duration, this reverberation acts like noise on the following signals and thus reduces the intelligibility. In EASE4.0, the Peutz long formula is implemented too, according to Fig. 1.

Via the equation[2]

$$RASTI = 0{,}9482 - 0{,}1845\ln(Al_{cons}) \tag{7}$$

it is also possible to establish an analytical correlation between the two quantities RASTI and AL_{cons}.

[2] Note by the editor: this equation, referred to as the Farrel Becker equation, is often used to relate STI to AL_{cons} scores. It appears that the source of this equation is not documented in open literature. However, a remarkable correspondence is observed with the empirical data reported (in a figure rather than as an equation) by Houtgast, Steeneken and Plomp (1980; Fig. 3. see previous footnote for reference). It seems reasonable to assume that the equation was either obtained through similar experiments, or derived from the data reported by Houtgast et al.



$$ALcons = 10^{-2\left[(A+B\cdot C)-(A\cdot B\cdot C)\right]} + 0.015$$

$$A = -0.321g\frac{\tilde{p}_{dir}^2 + \tilde{p}_N^2}{10\cdot\tilde{p}_{dir}^2 + \tilde{p}_{diff}^2 + \tilde{p}_N^2}$$

$$B = -0.321g\frac{\tilde{p}_N^2}{10\cdot\tilde{p}_{diff}^2 + \tilde{p}_N^2}$$

$$C = -0.51g\frac{RT}{12}$$

Figure 1. AL_{cons} calculation in EASE4.0

6.2. APPLICATION OF INTELLIGIBILITY MEASURES IN EASE4.0

6.2.1. Methods used in EASE 4.0

To calculate all the mentioned measures you may go back to simple statistical approaches or make use of sophisticated calculation routines (CAESAR), developed by the Aachen University and implemented into EASE. Thus not only simple omnidirectional sound sources but loudspeakers, line arrays and complete clusters have been introduced to work with CAESAR. Therefore this new module in EASE was named AURA, which stands for **A**nalysis **U**tility for **R**oom **A**coustics. So the following three methods will be compared:

- Standard: Here the direct sound and the statistical tail energy will supply the results
- Standard with reflections: Additionally to the normal standard calculation a ray-tracing module (here set to 3rd order, 1000 rays) is added to find short time reflections
- the AURA algorithm

6.2.2. Computer model

For the following computer simulations a known example has been used: St. Stephens church in Gladstone, Oregon, USA. It is a simple model; only one central speaker is in use. The volume is only about 2000m³ and for the investigations there have been chosen 4 so-called probe seats named A1#5, A2#15, A3#27 and A4#31 (see Fig. 2).

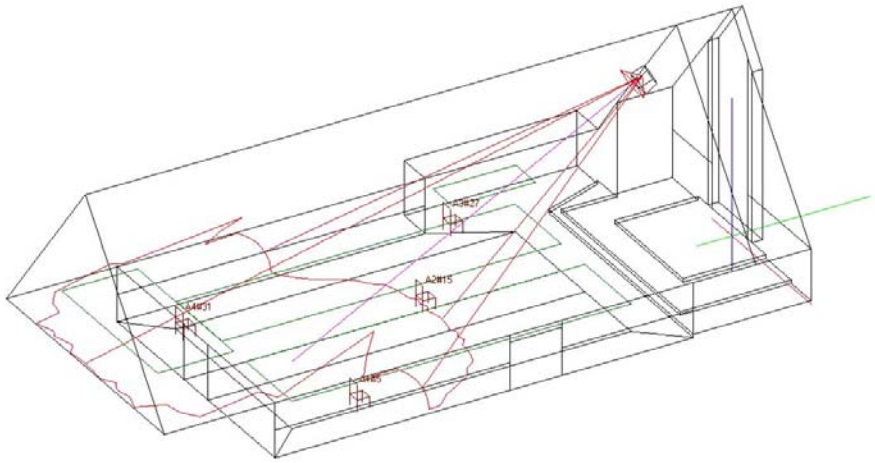


Figure 2. Model of St. Stephens church in Gladstone, Oregon, USA.

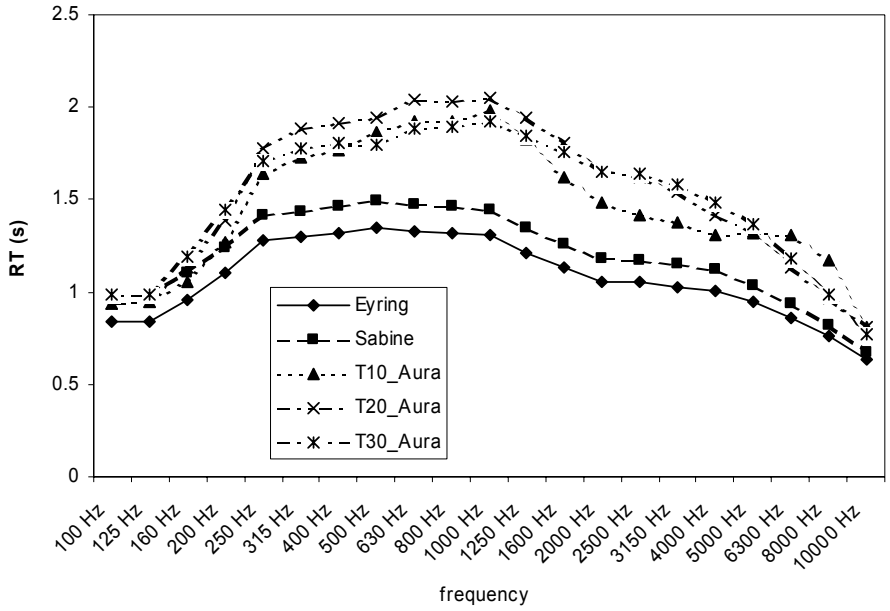


Figure 3. Reverberation time calculated under different conditions and averaged over 35 probe positions.

You can see that the statistical reverberation times are shorter than the three other ones, calculated with the new AURA algorithm in EASE4.0. This higher values result from the unequal distribution of absorption materials along walls and ceiling. Also flutter echoes have been observed.

The Direct SPL at all 4 probe positions is exactly the same and shown in Fig. 4. The total sound is, of course, different and shown in Fig. 5.

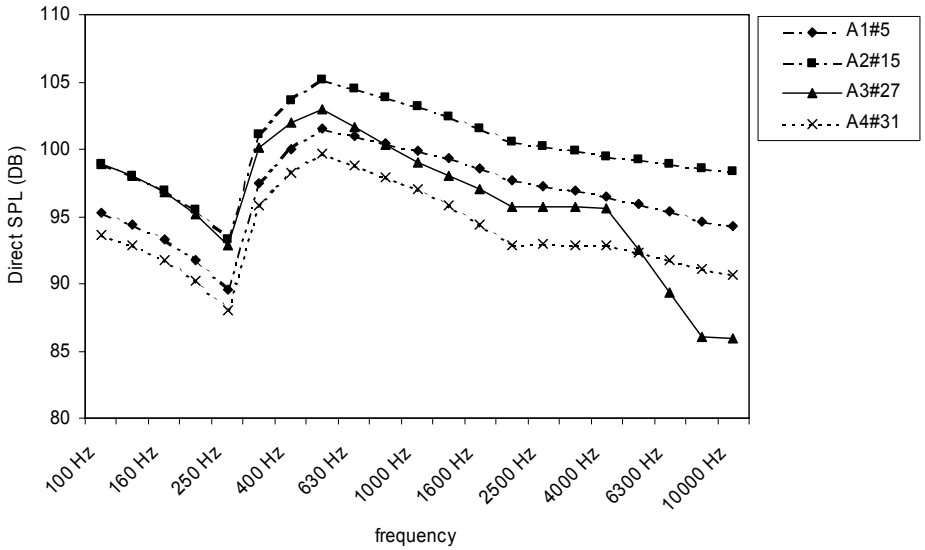


Figure 4. Direct SPL at all four probe positions in the St Stephens church model

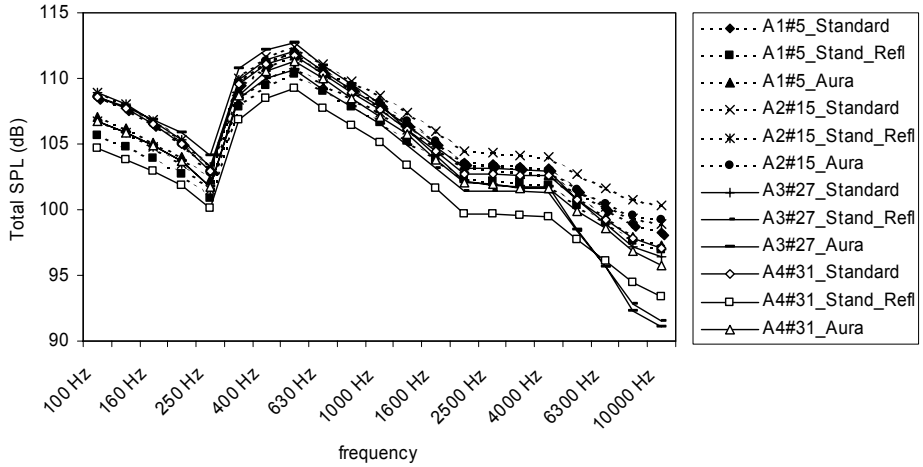


Figure 5. Total sound level (dB) at four probe positions for various calculation algorithms

The highest values for low frequencies can be observed at probe position A3#27 (low direct sound of the speaker), for higher frequencies in the center of the hall at A2#15.

These values influence the results of speech clarity and intelligibility. This shall be discussed now.

6.2.3. Comparison of different intelligibility measures in EASE4.0

6.2.3.1. Definition measure C_{50} and speech definition D

A first classical measure is the C_{50} measure and its results at the 4 probe positions (PP) are shown in Fig. 6.

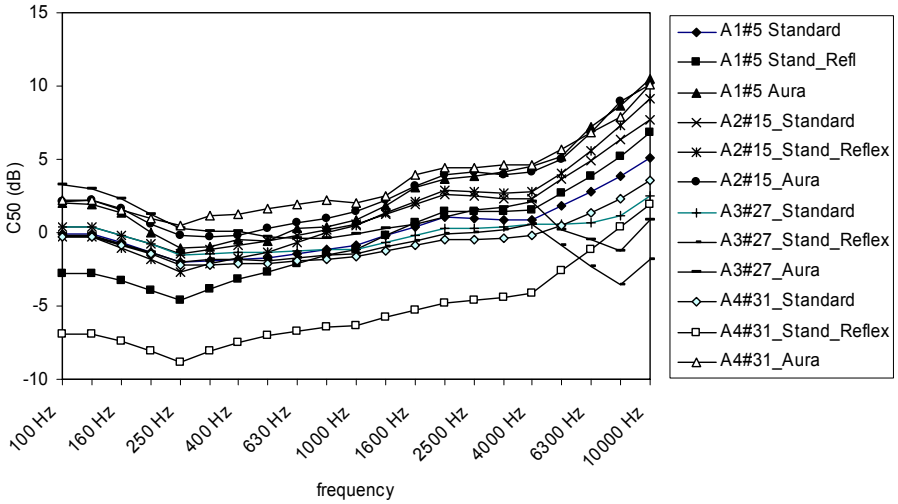


Figure 6. C50 at four probe positions for various calculation algorithms

The highest C50 values and certainly the most accurate ones are obtained with the AURA algorithm, all AURA curves are shown above. Only at PP A3#27 there is a decay at high frequencies, because the speaker is highly directional and does not radiate enough high frequency signal to this probe position. This becomes more clear from Fig. 7, where the Deutlichkeit D (speech definition) is depicted. The lowest curve is calculated with option Standard + Reflections at PP A4#31, that means on the balcony. Here certainly the numbers of rays were not sufficient.

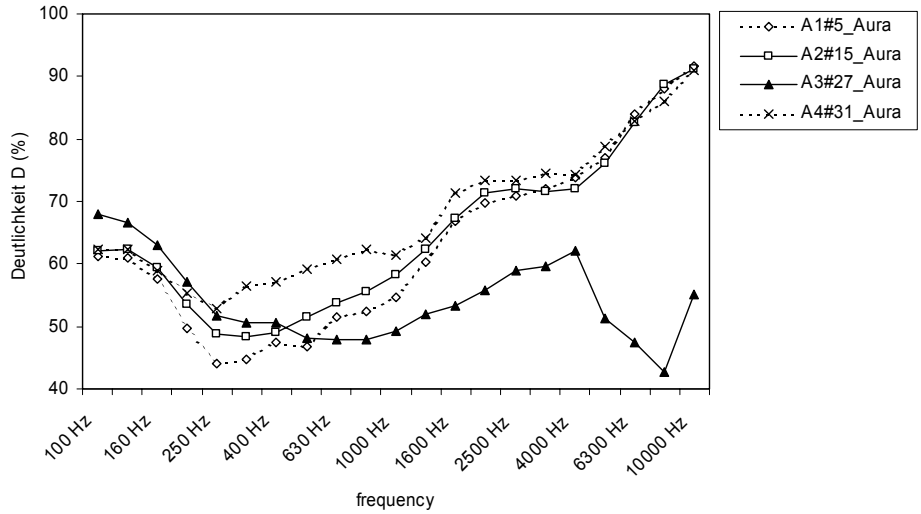


Figure 7. Deutlichkeit at four probe positions

6.2.3.2. Articulation loss of consonants

A more important measure used for simulation jobs in sound reinforcement is the so called Articulation Loss of Consonants (see above). Here the magic threshold is often set to 10% and this value is also used in tenders to describe the speech transmission quality of a new sound system intended to be installed.

This AL_{cons} number is, of course, a single number and cannot alone guarantee the quality of sound transmission, but describes quite well the quality level of the system. Therefore it is important to get a fast impression of the transmission behavior if you run a simulation. EASE included already in the DOS days algorithms for evaluating the speech transmission. Since 1971, as AL_{cons} was introduced, a couple of changes have been made to adapt these measures to new needs regarding a better description of real speech transmission. So we distinguish today a short form (Eq. (1)), a long form (Fig. 1) and special modifications used by the TEF community. Mainly in use today is the long form, refurbished by Peutz in 1990. But anyway all AL_{cons} measures are included in EASE 4.0 and the long form is used here for the evaluations of our example. Calculated STI values transferred to AL_{cons} by means of the Farrel-Becker equation (Eq. 7), are additionally used. The results are depicted in Fig. 8.

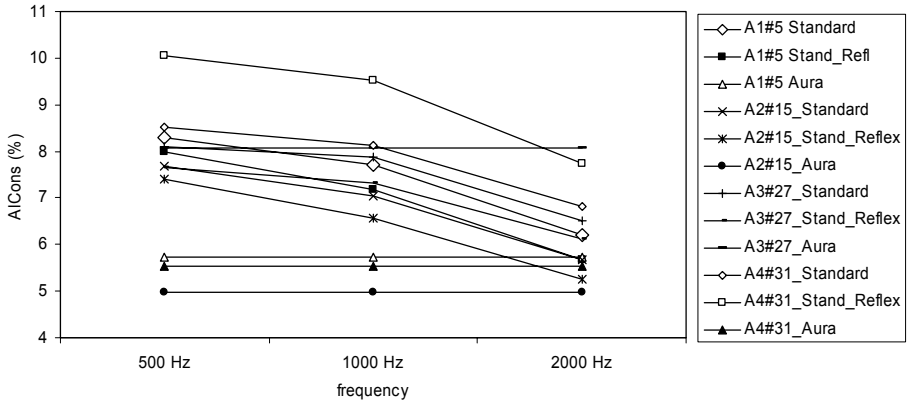


Figure 8. AL_{cons} at four probe positions for various calculation algorithms (Calculated Aura-STI values are transferred to AL_{cons} using the Farrel Becker equation (Eq. 7))

The decaying curves are calculated with the Peutz long form accordingly for the three mid-frequencies 500, 1000 and 2000 Hz. The horizontal lines belong to the STI recalculation, because the single number of full STI supplies only one equivalent AL_{cons} number.

Thanks to the full response calculation (including diffusion) the AURA results should come very close to reality. To get a quick overview the statistical approach using the long form is recommended. Fig. 8 shows that the 2000 Hz value corresponds well with the value derived from full response calculation; but this is only valid as long as the listener seats (here probe positions PP) are inside the coverage cone of the speaker. Outside this cone and still more with omnidirectional sources the 500Hz long form value gives a better representation.

All this corresponds to our experience and is thus confirmed by these calculations.

6.2.3.3. STI and RASTI

The measure STI is becoming more and more important because of its advantage resulting from calculating the Modular Transfer Function MTF directly from the impulse response. After having a full response calculation been included into EASE with the CAESAR algorithm it was a logic step to add the STI evaluations as well. In contrast to the measured impulse responses we deal here with discrete impulses (response or echogram files), which required special routines for deriving the 98 MTF values.

A response in EASE is given by a set of pulses. Each pulse is defined by an arrival time and a complex frequency response in third octave band resolution. To obtain a broadband ETC several steps of calculation have to be done.

As a first step the complex frequency response has to be transformed into an energy distribution in the time domain by means of a FFT. The phase used in addition to the magnitude can be one of the following three:

- run time phase
- true phase, which uses the loudspeaker phase data in addition to the run time phase
- minimum phase, which is calculated using the minimum phase algorithm (see Ahnert et al., 2001).

After having calculated the impulse response of a single pulse in the time domain, the arrival time has to be taken into account. This time is used as an offset for the final summation of all impulse responses for all pulses.

As a result one finds the complete impulse response of the system in the time domain. It is common knowledge that this impulse response can then be used as a basis for computing the MTF values.

In the comparisons there were used not only the full STI, but also the weighted male and female ones. Also the RASTI values derived from AL_{cons} calculations and transferred to RASTI via eq. (7) were additionally included.

The results are shown in Fig. 9.

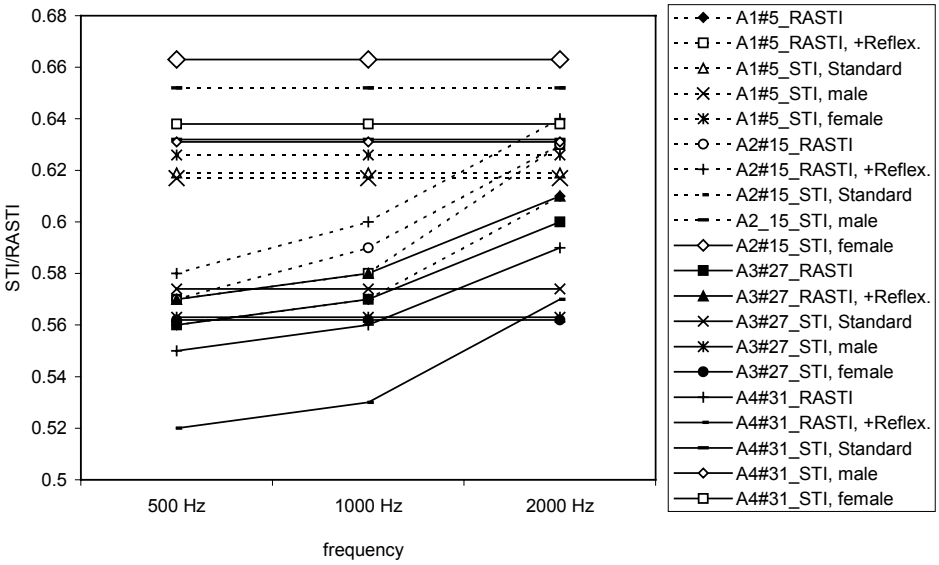


Figure 9. STI/RASTI at four probe positions for various calculation algorithms (Calculated AL_{cons} values are transferred to RASTI using the Farrel-Becker equation (Eq. 7))

In contrast to AL_{cons}, you see that the full response calculations with AURA supply the highest STI values and here especially the female weighted one for PP A2#15 (in the center of the coverage cone). The worst AURA result you get for PP A3#27 (very bad direct sound coverage). The transferred RASTI values fit well in the coverage areas for 2000 Hz, outside the coverage areas for 500Hz. These are the analogue results to AL_{cons} considerations.

6.3. CONCLUSIONS

With EASE 4.0 the new AURA module enables very time-effective full response calculations for deriving speech intelligibility measures. These speech transmission estimations can be done here not only for omnidirectional sound sources, but also for loudspeakers. We would like to emphasize especially that now line arrays or other cluster arrangements can act as sound sources for evaluating the improvements obtained by substituting "bad acting" speakers by better ones.

The paper shows convincingly that statistical approaches can lead to results coming close to full response simulation. You only have to know how to interpret the results and what to apply for the corresponding statement. The more measures and methods are included in simulation programs, the higher are the demands on the program user to select the correct ones and to draw the needed conclusions. To become an expert in acoustics it does not suffice just to buy a computer program, but it is equally important to attend some training courses.

REFERENCES

Ahnert, W. (**1975**): Einsatz elektroakustischer Hilfsmittel zur Räumlichkeitssteigerung, Schallverstärkung und Vermeidung der akustischen Rückkopplung (Use of Electro- acoustical Means for Enhancement of Spaciousness, Sound Reinforcement and Feedback Suppression), Techn. University Dresden, PHD work, pp. 87-89.

Ahnert, W.; Bourillet, C.; Feistel, S. (**2001**). Phase Presentation in the Acoustic Design Program EASE. Presented at the 110th AES Convention, Amsterdam, Holland, May 2001.

Houtgast, T. and Steeneken, H.J.M. (**1985**). A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria. J. Acoust. Soc. Amer. **77**, pp. 1060-1077.

Klein, W. (**1971**). Articulation Loss of Consonants as a Basis for the Design and Judgement of Sound Reinforcement Systems. J. Audio Eng. Soc. **19.**

Peutz, V.M.A.(**1971**) Articulation Loss of Consonants as a Criterion for Speech Transmission in a Room. J. Audio Eng. Soc. **19**, H.11, pp. 915-919.

Thiele, R. (**1953**).: Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen. Acustica (Directional Distribution and Time Sequence of Sound Reflections in Rooms), Beiheft Nr. 2, p. 291.

Chapter 7. Development of an Accurate, Handheld, Simple-to-use Meter for the Prediction of Speech Intelligibility

Kenneth Jacob, Chief Engineer, Bose Professional Systems Division
Steve McManus, Senior Engineer, Gold Line Corporation
Jan A. Verhave, Research Scientist, TNO Human Factors
Herman J.M. Steeneken, Senior Research Scientist, TNO Human Factors

ABSTRACT

International, European, and North American codes and standards now require sound systems used for emergency purposes to meet or exceed a minimum level of speech intelligibility. Intelligibility standards are also routinely being used for non-emergency systems. In both kinds of systems, however, verifying compliance is time consuming and requires the use of complex instrumentation and highly skilled technicians. As a result, the routine measurements of intelligibility necessary to enforce the codes and standards are rarely if ever made. With this problem in mind, the authors set out to develop a dedicated instrument that would accurately measure speech intelligibility according to international standards, and do so quickly and simply, thus making it possible for experts and non-experts alike to quickly measure intelligibility. Their results show clearly the feasibility of such an instrument.

7.1. INTRODUCTION

Speech intelligibility is arguably the most important dimension of sound quality in most commercial and professional audio installations. Unfortunately, it is also the most common source of complaints from owners, operators, and the public. Why?

The answer can not be based on a lack of attention from the scientific, engineering, and regulatory communities. Speech intelligibility has received enormous attention from these communities for at least fifty years. For example, researchers have developed a number of proven methods for predicting intelligibility. Software programs for predicting the intelligibility of a design before it is installed – even before a building is constructed – have been developed and are used extensively. And technical codes and standards covering speech intelligibility are widely used by governments, facility owners, architects, audio consultants, and other professionals throughout the world.

In spite of this progress, measuring intelligibility using any of the proven methods remains a relatively difficult and expensive undertaking. As a consequence, routine surveys of existing facilities are rarely performed; sound system engineers design for good intelligibility but then rarely measure to verify performance; clients make intelligibility a high priority but typically do not have the final intelligibility level confirmed; and finally, because intelligibility measurements are so difficult and expensive to make, codes and standards writers are reluctant to require periodic compliance measurements even though they recognize that such measurements are essential to effective enforcement.

Fundamentally, then, achieving a desired level of speech intelligibility is essentially an open-loop process because feedback in the form of confirmatory measurements is so rare. As with any open-loop process, convergence can be slow and uneven, and is in fact not guaranteed. This suggests why, in the presence of almost universal agreement as to the importance of speech intelligibility, actual performance continues to generate a large number of complaints.

The problem is that speech intelligibility is not a simple parameter to measure. If it were, engineers would have developed accurate and easy-to-use instruments years ago. The proven methods of predicting speech intelligibility from physical quantities require relatively complex instrumentation and post-measurement signal processing. What has made the problem more tractable recently is 1) the availability of powerful, yet cost-effective microprocessors, and 2) research that shows how the computational demands of a proven measurement method can be reduced without sacrificing accuracy in the intended applications. Armed with both, a collaborative effort was undertaken to develop a fast, accurate, portable, and easy-to-use intelligibility prediction meter. The remainder of this paper describes the approach taken and the results obtained.

7.2. SPEECH TRANSMISSION INDEX (STI) CONSIDERATIONS

A proven method of predicting intelligibility from physical quantities is the Speech Transmission Index, or STI method, developed at TNO Human Factors in the Netherlands (Houtgast and Steeneken, 1985a; Steeneken and Houtgast, 1999). Its accuracy, robustness, diagnostic capabilities, and wide applicability have made it a nearly unanimous choice by writers of codes and standards throughout the world.

In systems where linearity can be guaranteed, the STI can be measured relatively quickly by measuring the system's transfer function, and then using a mathematical formula to obtain the data necessary to compute the STI (Schroeder, 1981; Steeneken and Houtgast, 1980). Unfortunately, linearity can not in general be guaranteed; in fact many systems have diminished intelligibility due for example to amplifier clipping and overdriven loudspeakers. Regardless, a test for linearity must be conducted first before the transfer-function method of calculating the STI can be used safely. Such a linearity test has complexities that would greatly compromise our goal of developing a simple yet accurate intelligibility prediction meter. Any technique that relies on measurement of a system's transfer function must therefore be rejected given that systems in which non-linearities are responsible for degradation of speech intelligibility are a fact of life, and given our goal of creating a fast and easy-to-use instrument.

To measure the STI of an arbitrary sound system, including systems where non-linearities play a role, a more direct approach must be taken. The STI method is based on the amplitude modulation of octave bands of noise, where the modulation frequencies and octave bands are chosen to match those of natural speech. Optimally, fourteen different modulation frequencies and seven octave bands are used, making a total of 98 different combinations of modulated noise. The STI method is based on research showing that the loss of modulation from a system's input to its output – which represents the loss of the modulations in natural speech – is also a measure of the loss of intelligibility. These so-called modulation reduction factors for the 98 combinations are clipped, weighted, and averaged to obtain the final STI value.

To measure the modulation reduction factor for just one of the 98 combinations requires that at least several periods of the lowest modulation frequency of interest be passed through the system. Other considerations such as error detection (discussed in more detail below) mean that a measurement that includes all 98 combinations would take on the order of 10-15 minutes to complete. Such a measurement, which would yield a value for only one

position in a room or one condition in a sound system, misses the target for a fast measurement by at least an order of magnitude, and must also therefore be rejected as an option given our goals.

Measurement time considerations are not new to the researchers and users of the STI method. In the early nineteen eighties, research was conducted at TNO to determine if a smaller set of the 98 combinations of amplitude-modulated noise bands could be used in certain applications with an acceptable loss of accuracy. A subset of nine combinations, called RASTI (for Room Acoustics Speech Transmission Index or Rapid Speech Transmission Index) was shown to be effective in some applications, and a commercial instrument was produced[1] embodying the technique (Houtgast and Steeneken, 1985b). Only about eight seconds were required to make a RASTI measurement. However, the instrument was considered as a screening device for person-to-person communications because the excitation signal was limited to just two octave bands. This made the system unsuitable for measurements on sound systems in rooms, which are known to vary widely in their performance from octave band to octave band.

The explosive advancements in micro-circuitry and software in the years since RASTI was developed compelled these authors to revisit the question of making fast and accurate STI measurements suitable for use in the professional and commercial audio field. First, the team felt it was possible to drastically reduce the bulk of equipment required on the reception-and-processing end of the measurement by using digital micro-circuitry; a bulky piece of equipment requiring a power cord could conceivably be reduced to a lightweight, handheld, battery-operable device. Second, the team felt a new subset of the 98 combinations of modulated noise could be designed that was optimized for public address systems.

7.3. STI-PA: AN EFFICIENT FORM OF THE SPEECH TRANSMISSION INDEX METHOD FOR PUBLIC ADDRESS AND SOUND REINFORCEMENT SYSTEMS

Given these considerations, research was undertaken to determine if a new subset of modulated noise bands could be shown to be an acceptable substitute for the full set of 98 modulated bands over a set of conditions designed to represent commercial and professional sound systems.

As an initial starting point, the team chose to include all seven octave bands rather than the two used in RASTI measurements. Eliminating whole octave bands at the outset was avoided because professional and commercial sound systems can suffer from degradation in intelligibility in any of the relevant frequency bands. However, it is well known that the lower octave bands of speech play a much smaller role in speech intelligibility than the middle and higher bands (Houtgast and Steeneken, 1985c). Thus a portion of the research focused on the question of whether separate 125 and 250 Hz octave bands were justified. Additionally, the team considered simultaneous modulation by two or more frequencies in each octave band as an option. Tempering this potential advantage, however, was the knowledge that addition of a second, third or greater number of modulation frequencies in a single octave band would have the effect of reducing the dynamic range of the excitation signal and thus reducing the accuracy of measurements in situations where the ambient noise level would be very low.

In the end, the team arrived at a six-band, two-modulation-frequencies-per-band signal. The six bands consist of the 125 and 250 Hz octave bands combined, together with the 500 Hz - 8 kHz octave bands. The STI value obtained using the new subset was compared to the STI value obtained using all 98 combinations in a large number of conditions intended to

[1] B&K Speech Transmission Meter 4225, and Receiver 4419.

represent systems in which intelligibility is degraded by reverberation, noise, and band limiting. The results, showing more than 800 conditions, are shown in Fig. 1. The agreement is outstanding, proving the value of the new subset as a computationally efficient and viable substitute for the full 98 combinations of modulated noise in public address and sound reinforcement applications.

The name "STI-PA" (for Speech Transmission Index – Public Address) was chosen to differentiate this subset of modulated bands from others already in use (e.g. RASTI, or STITEL, which is designed for testing telephone systems).

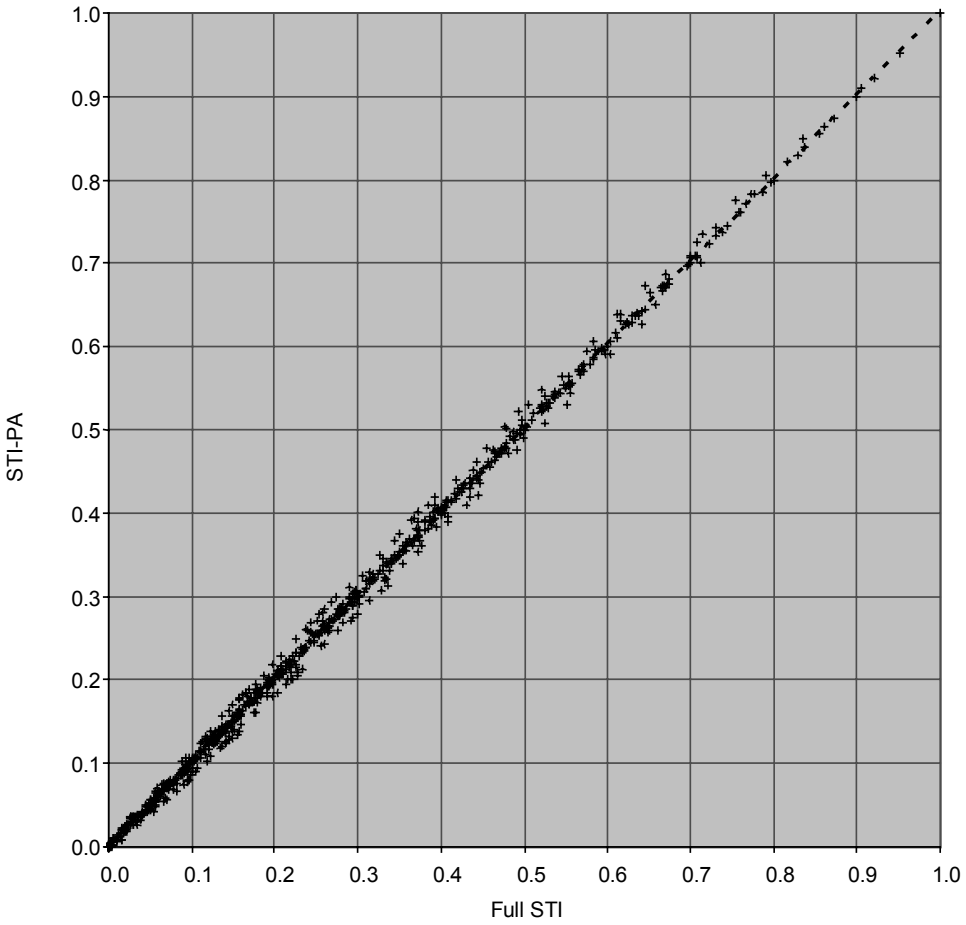


Figure 1. STI values calculated using the new STI-PA excitation signal are compared to STI values calculated using the full 98 combinations of modulated noise. The 45° line represents perfect agreement. A total of 880 different conditions varying reverberation times (0.00, 0.25, 0.50, 1.00 2.00, 4.00, and 8.00 seconds), signal-to-noise ratios (-15, -12, -9, -6, -3, 0, 3, 6, 9, 12, and 15 dB), and bandpass filters (2 kHz, 1-2 kHz, 1-4 kHz, 1-8 kHz, 500 Hz - 2 kHz, 500 Hz - 4 kHz, 500 Hz - 8 kHz, 125 Hz - 2 kHz, 125 Hz - 4 kHz, and 125 Hz – 8 kHz) are shown. In 833 out of 880 conditions (95%) the error is less than ± 0.02. The maximum error is ± 0.03. The agreement is outstanding, proving the suitability of the STI-PA signal in public address and sound reinforcement applications.

7.4. EMBEDDED SYSTEM VS. GENERAL PURPOSE COMPUTER

With the results showing that a new, computationally efficient form of the STI could be used, attention turned to the issue of implementing the reception and signal processing functions on a micro-processor-based measurement device. While the necessary processing certainly could be made to fit on a general-purpose computer, an embedded system was selected in order to guarantee an instrument that was simple to use and reasonable to calibrate and service.

From one perspective, this decision may seem unintuitive since the cost of a powerful laptop computer is now comparable to an embedded-system device, and can do so much more. And yet that is exactly why the general-purpose computer was rejected. The fact that it can 'do so much more' is the very reason that each user-customized unit tends to behave differently, and why one machine's results often fail to match another's, both contrary to the fundamental principle that measurements should be repeatable from one like machine to another.

Moreover, the instrument envisioned is meant to be used in code-compliance measurements (among other things) which means that proper service, calibration, and maintenance will be essential if owners, insurance companies, government officials, and others are to completely trust the results. The idea of providing the required level of service on a multitude of general-purpose computers, each of which had been customized by the user, was rejected as unmanageable.

7.5. USER INTERFACE

To reach the goal of making a meter that would be truly easy to use, we envisioned a user interface consisting of a single button that would initiate a measurement and a counter that would inform the user how much time remained to complete the measurement. With the excitation signal running continuously, the user would initiate a test, and about fifteen seconds later would get an intelligibility score based on the STI.

One of the most important and widely applied standards, *IEC 60849: Sound Systems for Emergency Purposes*, allows the use of a number of different methods for measuring speech intelligibility, including the STI. In order to relate the results of the different methods, a Common Intelligibility Scale (CIS) was created. The actual language in IEC 60849 calls for a minimum CIS value, which is therefore what we also chose to display. The meter computes the STI and then converts it to the CIS.

7.6. ERROR CHECKING

Because the anticipated users of the meter include non-experts in acoustics (inspectors, local authorities, etc.) as well as acoustical experts, the team felt it was important that the meter have robust error detection capabilities. In the case of an intelligibility prediction meter upon whose output could depend the granting of an occupancy permit, the outcome of a lawsuit, or the rightful collection of professional fees, the desire to notify the user when accurate measurements could not be made was considered of paramount importance.

To address this need, error-checking algorithms were developed and implemented in the meter. For example, a common occurrence when making acoustical measurements of any kind, including STI measurements, is for conditions to be interrupted by spurious noise. If someone talks next to the microphone during a measurement, or if a worker drops a piece of construction material, for example, the data generated are suspect. Rather than rely on the judgement of the user in such situations, the meter automatically invalidates measurements

when the interfering noise produces an unacceptable error. In most cases, the test can then simply be re-run.

7.7. INSTRUMENT TESTING

Prototypes of the meter [2] were tested to ensure that they behaved according to theory and met our goals for accuracy, portability, and simplicity. STI values obtained on the meter were compared to STI values obtained using the full 98 combinations of modulated noise as calculated on a reference system at TNO. A number of representative test conditions were used, including different bandpass conditions, noise levels, non-linear effects, reverberation and echo profiles. The results are shown in Fig. 2. The data show conclusively that the prototype meter is accurate in a wide range of conditions typical of public address and sound reinforcement systems.

7.8. CONCLUSION

The central assumption here is that if the world wants better intelligibility – which it apparently does based on the proliferation of codes and standards requiring it – then we must have a simple and effective means to measure it. Without such means, intelligibility will continue to be of intense scientific and technical interest but the public will not benefit in any fundamental or widespread way.

With these concerns in mind, these authors felt that it might be possible to develop a speech intelligibility prediction meter that worked almost as easily as a sound level meter. The approach and results presented here are intended to build the confidence of other experts that such a meter is feasible and has been successfully demonstrated.

Such a device may trigger concern among the community of experts who currently make intelligibility measurements. If others who are not necessarily acoustical experts can now make intelligibility measurements, does this threaten the expert's income? We believe the opposite is true – that the existence of a simple and effective way to measure intelligibility will greatly increase the demand for expertise, not jeopardize it. Who will be called on to diagnose and recommend changes to a system that fails an intelligibility test? Similarly, who will be called on to create a design that is certain to pass the test in the first place? Frankly, we suggest the problem may be that we lack enough experts to meet the demand that will be created.

Fundamentally, our hope is that the single-most important dimension of sound quality – speech intelligibility – undergoes significant improvement in the coming years. Spoken announcements remain one of the most effective and vital means of communication in public places and places of business – for both emergency and non-emergency purposes – but only if they are intelligible. Together with the scientific knowledge of how to measure intelligibility, and many of the codes and standards that allow the public and private sectors to require it, we submit the 'third leg' now exists: a simple and effective means of measuring it.

[2] Gold Line DSP30 with STI-CIS™ module.

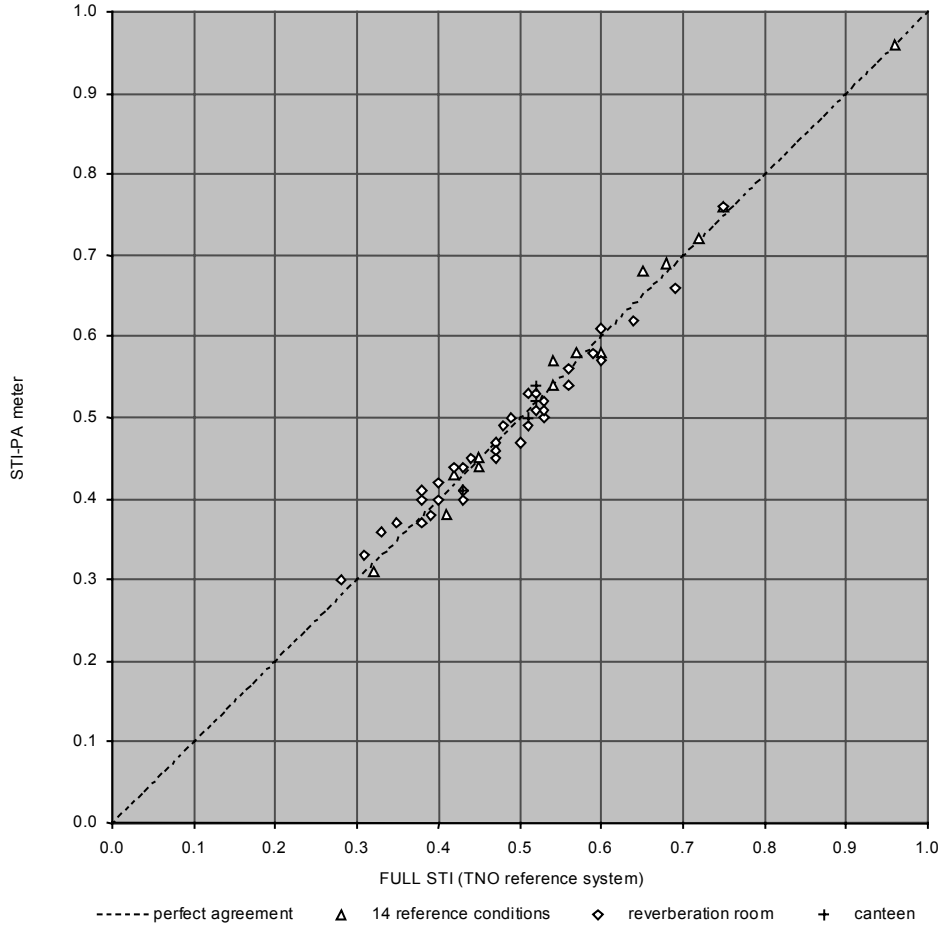


Figure 2. For a variety of conditions, STI values from the prototype meter are compared to values obtained on an STI reference system at TNO that uses the full 98 combinations of modulated noise. The test conditions are: 1) Fourteen reference conditions which use a mixture of band pass limiting, noise, peak clipping, and echoes[3]. 2) Measurements made in a TNO reverberation room using different numbers of absorbers in order to vary the reverberation time (0, 1, 2, 4, 8, and 16 absorbers resulting in a reverberation time range of 0.2 – 4.0 seconds), different bandwidths (500 Hz - 4 kHz and 1-4kHz) and different noise conditions (no noise and 50 dBA of speech babble), all using an artificial talker as a source producing 50 dBA at one meter.
3) Measurements from the TNO canteen, with either a PA system or an artificial talker as a source, and at two locations (5 and 15 meters from the artificial talker). Most data are within ± 0.02 and the worst case error is only ± 0.03. The results show that the meter is very accurate over a broad range of conditions common in public address and sound reinforcement systems.

[3] The fourteen conditions used are described in Steeneken, H. J. M., (1992) "On Measuring and Predicting Speech Intelligibility," Ph.D. thesis, University of Amsterdam, 143-144. The abbreviations for the fourteen conditions used here are: BP01 - BP1 + SNR inf; BP04 - BP1 + N2 + SNR = 0; BP08 - BP1 + N4 + SNR = 3; BP17 - BP2 + N4 + SNR = 6; NL01 - PC1 + BP1 + SNR = inf; NL04 - PC1 + BP1 + N4 + SNR = 9; NL08 - PC1 + BP2 + N1 + NSR = 6; NL14 - CC2 + BP1 + N4 + SNR = inf; E01 - E1 + BP1 + SNR = inf; E04 - E2 + BP1 + SNR = inf; E06 - E2 + BP1 + N4 + SNR = 6; E07 - E3 + BP1 + SNR = inf; E08 - E3 + BP1 + N4 + SNR = 12; E09 - E3 + BP1 + N4 + SNR = 6.

REFERENCES

Houtgast, T., and Steeneken, H. J. M. (**1985a**) "A Review of the MTF Concept in Room Acoustics and Its Use for Estimating Speech Intelligibility in Auditoria," J. Acoust. Soc. Am., **77**, (3).

Houtgast, T., and Steeneken, H. J. M. (**1985b**) "RASTI: A Tool for Evaluating Auditoria", B&K Technical Review.

Houtgast, T., and Steeneken, H. J. M. (**1985c**) "A Review of the MTF Concept in Room Acoustics and Its Use for Estimating Speech Intelligibility in Auditoria," J. Acoust. Soc. Am., **77**(3).

Schroeder, M. (**1981**) "Modulation Transfer Functions: Definition and Measurement," Acustica, **49**.

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., and Houtgast, T. (**1999**) "Mutual Dependence of the Octave-Band Weights in Predicting Speech Intelligibility," Speech Commun., **28**, 109-123.

Chapter 8. Practical application of STI to assessing Public Address and Emergency Sound Systems

Peter Mapp, Peter Mapp Associates, Colchester, CO3 4JZ, UK

ABSTRACT

Over the past ten years, there has been an ever increasing awareness of the need for Public Address and Voice Alarm systems to provide emergency as well as general announcements with a high degree of intelligibility. STI (and RASTI) have been at the forefront of this revolution as the need to measure and verify system performance gathered importance and momentum. However, measuring PA system performance is quite different to assessing the natural intelligibility of an auditorium or classroom for example. The paper discusses the relationship between STI and RASTI, based on examination of the data taken from over 80 sound systems. It is shown that RASTI is generally an inaccurate predictor of STI for a wide range of conditions[1]. Some of the practical complications, and limitations when testing sound systems are also discussed. In particular, it is shown that some forms of signal processing and irregular sound system frequency responses can give rise to STI results with reduced accuracy.

8.1. INTRODUCTION AND BACKGROUND

The past few years has seen an increasing demand for the installation of Voice Alarm systems rather than traditional fire alarms using bells or sirens. The ability of a Voice Alarm system to provide specific information concerning safety related incidents, potentially improves the evacuation response and successful egress from buildings and public spaces. However, to be effective, voice alarm systems need to be able to provide intelligible announcements and messages. The introduction in 1989 of a standard for emergency sound systems (IEC 8049/BS 7743-1991) has had a significant positive impact, particularly in the UK. For the first time, the required intelligibility of the sound system was specified (0,5 STI) with RASTI being the designated measurement/verification method. Other standards and codes of practice codes soon followed including CAA specification No 15 1989 and RTCA/DP-214, 1993, which for example specify given RASTI performance values for the PA systems of all commercial aircraft. Many military aircraft, also adopt these specifications and RASTI requirements. Contractually, many systems became liable to meeting a given RASTI performance and the implications for both sound system design and the acoustic environment in which a given system has to operate, slowly emerged. The result in the UK

[1] Note by the editor: in this chapter, the author gives various examples of cases where the improper choice to use the RASTI method would lead to results that do not correspond well with the "full" STI method. As explicitly stated in relevant international standards, as well as in scientific literature, the RASTI method is a simplified version of the full STI method, which can *not* be used with some types of speech degradation, such as speech bandwidth limiting and non-linearities, that frequently occur due to PA systems. This chapter demonstrates what the consequences are if (national) standards and regulations call for application of RASTI in situations where this method is actually unsuitable.

has been a significant improvement of the quality and intelligibility of PA and VA systems, though at times this has been a painful process, as initially the acoustic implications were not fully understood. Indeed, it still comes as a surprise to many, that the acoustic conditions of a building or space, are the limiting and often the determining factor of the resultant intelligibility of an installed sound system.

Whereas subject based word score tests are the most accurate method of ensuring that a given sound system is intelligible, this is an extremely expensive and cumbersome technique to conduct and is not a practical assessment option for most sound system installations. The requirement for a simple to use, electro-acoustic method, whereby a specific test signal is either electronically or acoustically injected into the sound system and the resultant acoustic response measured was manifestly needed. Whereas the most common technique in Europe is RASTI, in the USA, a direct-to-reverberant ratio measure relating to percent AL_{cons} is still currently the most frequently employed. [Though the recent introduction in the USA of NFPA 72 and the adoption of the requirements of EN60849 is likely to change the USA's reliance on % AL_{cons}].

Concern has been expressed over the accuracy of both the above techniques when applied to sound systems (Mapp, 1997, 2001ab, 2002abc). As early as 1991, Mapp showed that it was possible to radically alter the frequency response of a sound system and for RASTI not to recognise this. Fig. 1 shows an example taken from this early investigation.

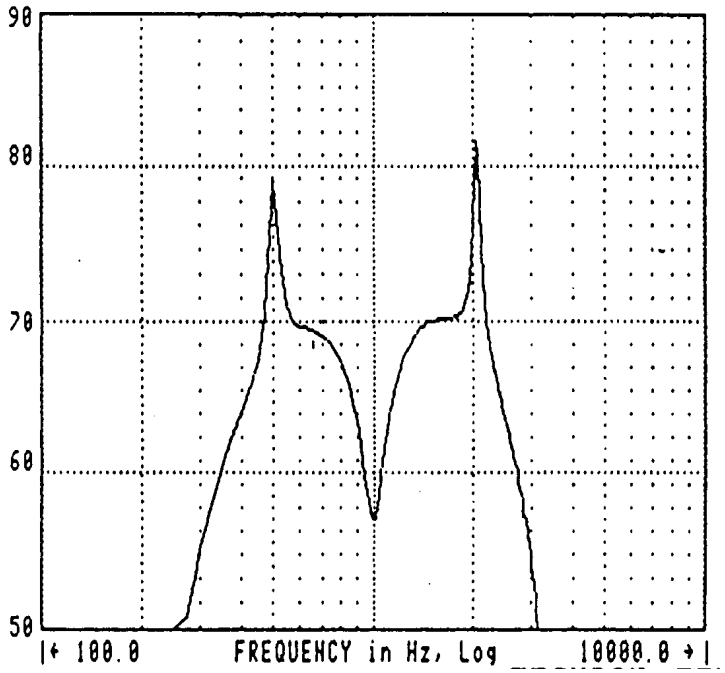


Figure 1. Effect of limited frequency information on RASTI[2]

[2] Note by the editor: the RASTI test signal, as specified by international standards, samples only two octave bands (500 Hz and 2 kHz), which makes the RASTI method insensitive to changes in frequency response. This is the main reason why these standards limit the application scope of RASTI to "pure" room acoustics (without electro-acoustics). IEC 60268-16 restricts the use of RASTI to cases where the level difference between any two adjacent octave bands is smaller than 5 dB.

Mapp also pointed out at this time that system non-linearity such as deliberately introduced signal compression could give rise to misleading results. It is interesting to note that both RASTI and % AL_{cons} were originally developed to assist with the assessment of natural speech transmission in auditoria or other acoustic spaces and not sound systems (e.g. IEC 268-16, 'The objective rating of speech intelligibility in auditoria by the RASTI method' - 1988). The application to the measurement of PA and sound systems followed later - without the realisation of the unique environment and conditions under which such systems operate and the potential for error that this might introduce.

Other measures, also based on auditorium acoustic assessment techniques have been tried but subjective value scales have not been sufficiently developed. (e.g. C_{50}). It has long been recognised that the full Speech Transmission Index (STI) method is significantly more accurate than its subset RASTI, however until very recently this required sophisticated analysis equipment and usually considerable operator skill and training. The revised version of EN 60849 (1998) – Sound systems for Emergency Purposes, recognised some of the potential problems with RASTI and introduced the concept of the 'Common Intelligibility Scale' or CIS. The concurrent revision of IEC 268-16 also recognised some of the problems and added some additional guidance with regard to measurement validity. (EN 60286-16 : 1998).

Although there is currently considerable opinion that RASTI should be abandoned with respect to sound system verification and full STI or STIPA employed instead, (e.g. the current draft revisions of EN 60849 and EN60268-16), there are many standards and codes that still directly cite RASTI and it will take several years to revise these. It therefore seemed timely (and this symposium offers an ideal opportunity) to present some hitherto unpublished research carried out by the author which illustrates the typical errors associated with RASTI and its accuracy in approximating the full STI.

8.2. RASTI VS. STI ERROR ANALYSIS

As it is the intelligibility performance of sound systems as opposed to rooms or auditoria, that is the current interest to the author, an analysis of a wide range of 'real world' sound systems was conducted. In total, the measurement data for eighty one sound systems was examined and the RASTI, STI and % AL_{cons} values have been compared, together with a number of other parameters. Subsets of the data (25 systems) have also been subject to further experimentation and analysis.

The sound systems involved covered a very wide range of types and applications, ranging from rail and transportation terminus systems to churches, theatres, concert halls, museums and stadiums. The range of associated STI values is also very wide but offers a skewed distribution as fewer systems exhibited very low performance results. Fig. 2 shows a regression plot of the STI and RASTI data taken for the situation where the signal to noise ratios were sufficiently high such as to ensure that reverberation is the dominant distortion or speech distractor. Although there is a strong relationship between the data, as would be expected, the correlation is only 0.9097, a significant discrepancy clearly being present.

Fig. 3 plots the data in terms of the error between the two measures. The mean error for the 81 systems is 0.08, which might at first glance seem fairly small, though both contractually and subjectively, this is certainly not the case. Examination of the curve shows that at around 0.5 STI (the target criteria for many VA systems) the error is as much as $\pm$ 0.1.

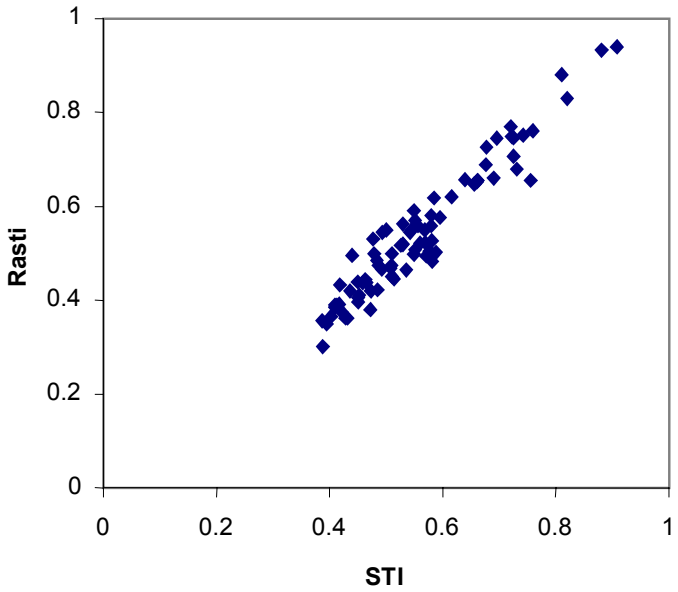


Figure 2. Correlation between RASTI and STI

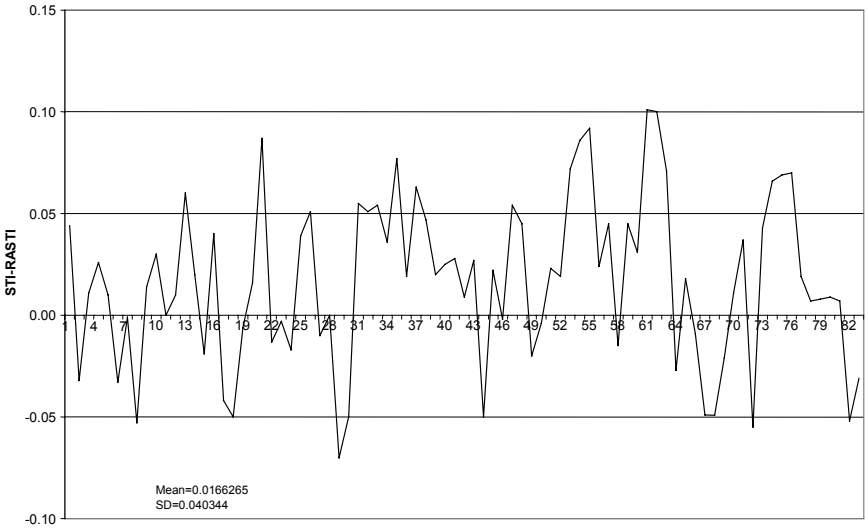


Figure 3. STI Vs RASTI error curve

Inspection of the error curve (Fig. 3) shows that the error distribution is not linear but in approximately $^2/_3$ of the cases, the RASTI values underestimate the STI but over estimate it in the other $^1/_3$. An initial review of the data shows no immediate factor that causes either the under or over estimation of the STI by RASTI. Fig. 4 shows the measured STI values versus the range of the Reverberation Times of the spaces employed (approximately 0.2 to 3.5 seconds). It is interesting to note from the figure that no relationship exists between the mean reverberation time (over the range of 500 Hz to 4 kHz) and the STI, although there is clearly a trend showing higher STI values with lower reverberation times – a not unsurprising finding (correlation coefficient = 0.4983).

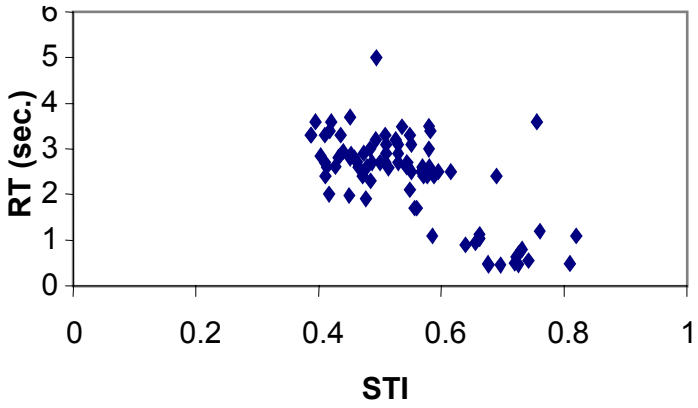


Figure 4. Range of RT values and STI

Possibly surprising however, is that no stronger correlation was found to exist between the EDT and the STI, as shown in Fig. 5.

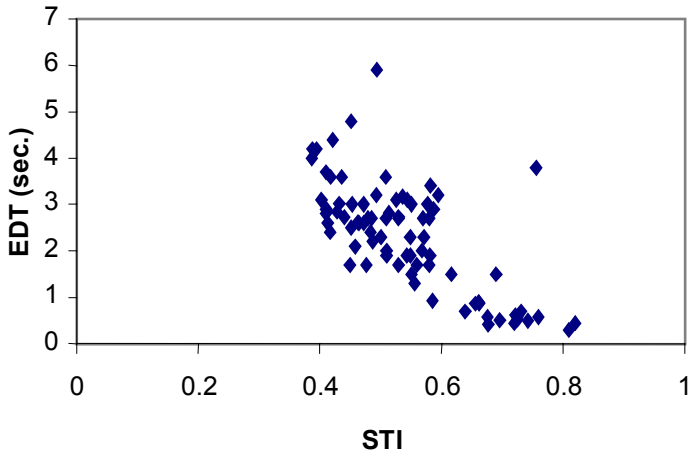


Figure 5. Relationship between STI and EDT

The effect of noise and reverberation as distractors was also examined at two different signal to noise ratios (+6 & +10 dBA) – but using a smaller sample of 25 systems. In the presence of noise and reverberation, the previous trend of underestimating the STI completely changes and in 90 % of the cases, RASTI now overestimates STI. (The mean error is 0.05). To put this into context, if we assume that RASTI was predicting a value of 0.50 STI i.e. a 'pass' as far as most emergency sound systems are on concerned, the actual value would in fact be just 0.45 – not only on the limit of practical intelligibility but also a clear 'fail' with respect to many current codes and system requirements! (The legal and contractual ramifications of this are of significant interest). At +6 dBA S/N, a similar trend was found to occur with the mean error increasing to – 0.065. (further details are given by Mapp, 2002a).

The effect of noise only as the intelligibility reducing parameter was also investigated. This particular study related to the performance of aircraft PA systems. The reverberation time of the cabin was very short (< 0.2 seconds) this coupled with the short distance between

the loudspeaker and listener resulted in very high initial STI and RASTI scores – as shown in Table I.

Table I : Aircraft PA system RASTI and STI Measurements

LS Type	RASTI (no noise)	RASTI (with noise)
A	0.93	0.76
B	0.91	0.65
C	0.95	0.71
	STI (no noise)	STI (with noise)
A	0.88	0.60
B	0.92	0.64
C	0.93	0.57

The agreement for the 'no noise' case between RASTI and STI is good, the variations being within 0.02 to 0.05 (mean = 0.03). However under typical aircraft operational background noise conditions, the discrepancy between RASTI and STI increased markedly. In this case the error range was between 0.1 and 0.16 – again very significant errors. This is particularly so when it is considered that in the case of loudspeaker type 'C' the situation changes from a good pass of 0.71 RASTI against the CAA requirement of 0.60 and the full STI value of 0.57! Again technically a failure, but as the CAA certification is in terms of RASTI and not STI, this becomes a contentious point. Particularly when it is realised that the better performing loudspeaker (B) both subjectively and in terms of STI was the worst performer in terms of RASTI! Word score testing carried out in conjunction with the measurements showed good agreement with the STI values. (Further details can be found in Mapp, 2001b).

The effect of non-linear behaviour in terms of signal clipping was also investigated for RASTI and STI. An interesting discrepancy was found to occur between laboratory tests and real world sound systems, with a 'straight wire' set up underestimating the effect found under field conditions. In each case the signal level was increased to clip the same component in the signal chain. Noticeable differences were found o occur between the reductions in STI and RASTI. Fig. 6 shows the effect of clipping on STI for five different sound systems. The behaviour is highly non-linear, with the same degree of clipping having a different effect on the different systems. Comparing the results with Fig. 7**,** which presents the data plot for the effect on RASTI, immediately shows the reductions for RASTI to be far smaller. The associated errors range from around 0.04 to 0.14 STI.

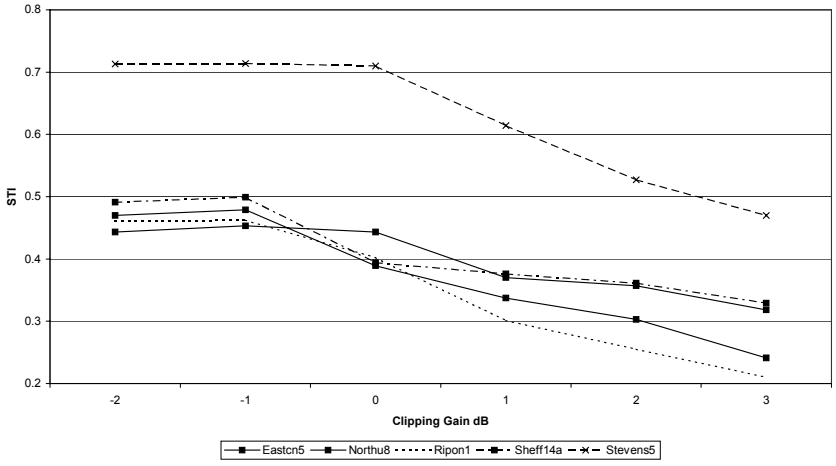


Figure 6. Effect of clipping on STI (MLS measurement)

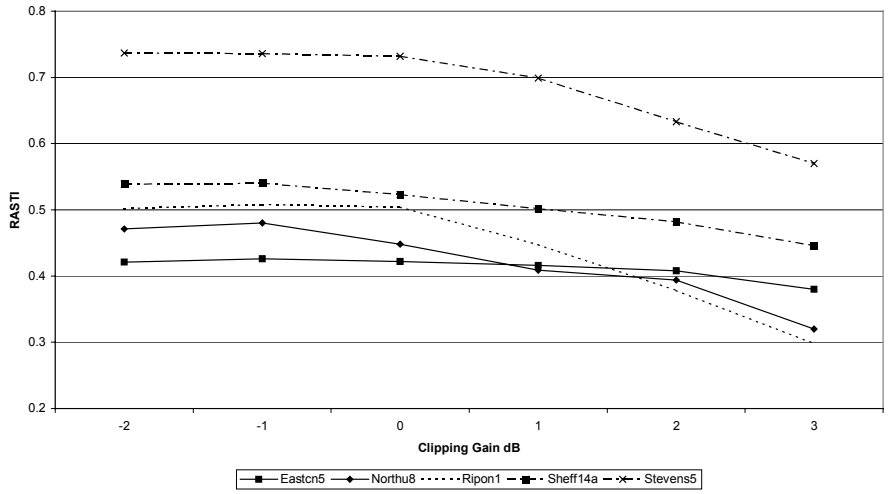


Figure 7. Effect of clipping on RASTI (MLS measurement)

The above test was conducted using an MLS test signal. An experiment was also run to compare the effects of signal stimulus. Fig. 8 shows the comparison between the results obtained using MLS STI and a modulated test signal (STIPA). As can be seen, the STIPA signal is far less sensitive to clipping.

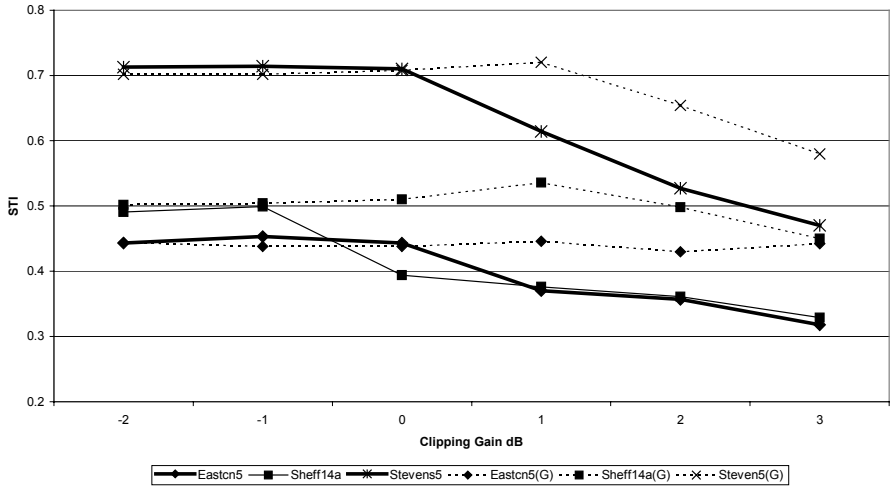


Figure 8. Comparison of clipping for STI MLS and modulated (dotted lines) signals

The above examples show not only the potential magnitudes of the errors between RASTI and STI under typical working conditions but clearly also show the need to retire RASTI for all forms of PA system assessment. However, the above measurements and results assume that STI is correct and infallible – but how robust is STI? can it be fooled? Some preliminary investigations have therefore been carried out and are reported below.

8.3. IS STI AN INFALLIBLE INDICATOR OF SOUND SYSTEM INTELLIGIBILITY ?

It is perhaps worth reviewing at this point, the main parameters that can affect the intelligibility of a sound system and seeing which ones STI is able to account for. This may then help to highlight any potential factors which STI may not be sensitive to. There two distinct groups of parameters: (A) those that relate to the production and reception of the speech signal itself and (B) those that relate to the sound system and its environment[3]. The latter category also includes measurement and stimulus related artefacts.

Primary factors include

- Sound system bandwidth and frequency response
- Loudness (absolute SPL)
- Signal to noise ratio (S/N)
- The direct-to-reverberant (D/R) ratio[4]

  This is determined by
  - Room reverberation time (RT60)
  - Volume and size and shape of the space
  - Distance from the listener to a loudspeaker
  - Directivity of the loudspeaker
  - The number of loudspeakers operating within the space

0 Talker annunciation/rate of delivery
0 Listener acuity

Secondary factors include

- System distortion (e.g. harmonic or intermodulation)
- System equalisation
- Uniformity of coverage
- Presence of very early reflections (< 1-2 ms)
- Sound focussing or presence of late or isolated higher level reflections (> 70 ms)
- Direction of sound arriving at the listener
- Direction of any interfering noise
- Amplitude Compression

0 Gender of talker
0 Vocabulary and context of speech information
0 Talker microphone technique

---

[3] The bulleted parameters marked • are building or system related, whilst those marked 0 relate to human factors outside the direct control of the system itself

[4] Strictly speaking a more complex characteristic than the simple D/R ratio should be used which should include the ratio of the direct sound plus early-reflected energy to late reflected sound energy & reverberation (eg C50 or C35). C50, whilst widely used in auditorium acoustics has yet to meet its full potential in terms of PA and sound system assessment, where it could be a very useful further assessor.

Although system equalisation should inherently be covered by the main factor of frequency response/bandwidth, it is separately identified as it is a parameter that can be readily adjusted and is known to affect subjective evaluation and intelligibility.

Referring to the four primary factors of bandwidth/frequency response, absolute SPL, direct-to-reverberant ratio and signal to noise ratio, STI is stated to take these into account (EN 60268-16, 1998). An STI measurement of course cannot directly account for talker annunciation/rate of delivery and listener acuity, though this can to a certain account be accounted for by adjustment of the criterion value required. (listener acuity can also incorporate language familiarity/fluency).

With respect to the secondary parameters, STI should account for distortion, equalisation and uniformity of coverage but may not deal with the coloration due to very early reflections, nor the effects of sound arrival direction. The effects of speech amplitude compression are also not catered for, although such processing can affect word scores by an equivalent of more than 0.1 STI (Barnett 1997)[5]. The gender of the talker is taken into account by running separate STI measurements with different male and female MTF weightings. The vocabulary/complexity of the speech information can be accounted for by setting an appropriate criterion and therefore is not part of the measurement itself. Talker microphone technique is highly variable and cannot readily be accounted for but measurements made with different source positions relative to the microphone can give an indication.

From the above analysis, it can be seen that STI should be able to cope with a wide range of sound system parameters. However, it is the author's experience that on occasion, the measured system STI does not agree with a subjective evaluation – albeit an informal assessment. In particular this seemed to relate to issues regarding the frequency response of a system. This is particularly noticeable when equalising a system, as almost instantaneous comparisons between different frequency response balances and speech clarity / intelligibility can be appraised. Whereas the improvement in the clarity of speech broadcast over the system after appropriate equalisation could generally be readily heard, the measured STI did not appear to reflect this, indeed often not showing any measurable difference at all!

8.3.1. Effect of non-linear frequency response on STI

A set of pilot experiments was therefore conducted to investigate this apparent discrepancy. Although the problem was mainly encountered in reverberant spaces (Mapp 1997), similar effects were also noted to occur in acoustically well controlled environments. (See Mapp, 2001, 2002).

Fig. 9 shows the results of one of the pilot experiments. The upper dotted curve shows the measured frequency response of a sound system operating in reverberant conditions (RT = 1.8 seconds) which results in an STI of 0.46 (CIS = 0.66). Subjectively this equated to understandable speech provided the listener concentrated on what was being said. The lower solid curve was taken with the response deliberately restricted such that it was -3 dB down at 1 kHz, –6 dB down at 2 kHz and –9 dB down at 4 kHz. Subjectively, the intelligibility degraded markedly with the word score reducing to an of equivalent of 0.3 STI, yet the measured STI remained at 0.46 ! A number of curves were examined and in each case the STI did not change, yet perceptually, the differences were significant.

[5] Note by the editor: also see chapter 4 of this book

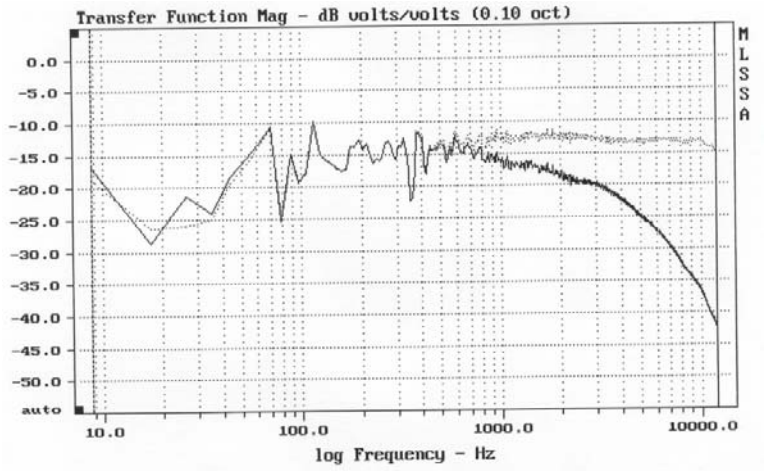


Figure 9. Frequency response of reverberant system, lower curve shows high frequency roll-off.

The above results indicate that STI does not appear to be able to take account of perceptually significant changes in intelligibility, brought about by modifications of the frequency response of a sound system, under quiet reverberant conditions (i.e. equalisation). This raises the question as to whether the problem is inherent in STI or associated with the measurement technique employed? The STI measurements were primarily made using a MLSSA analyser, although TEF TDS measurements were also carried out and shown to exhibit the same effect. Both MLS and TDS measurement techniques enjoy high noise immunity and it was wondered if it was a consequence of this or the fact that the STI was derived from the impulse response rather than a directly modulated signal. The experiment was therefore repeated using the Bose/Goldline STIPA signal and dedicated measuring instrument. (Jacobs et al 2001). A series of low pass (high frequency roll off) conditions was investigated based on the initial and final conditions presented in Fig. 11. The same effect was found with STIPA, ie the system was insensitive to a loss of high frequency information under quiet reverberant conditions. Good agreement was found to exist between STIPA and STI as measured using a MLSSA measurement system (0.45 and 0.46 STI respectively).

The above result therefore supports the theory that it is an inherent problem with STI rather than the current measurement platforms. This has major implications for sound system intelligibility testing and is currently being further investigated.

The above discrepancies relate to the frequency response of the sound system itself – but what happens when the input signal response is modified as for example is the case with male and female speech. Different octave band MTF weightings are available to cater for this, as given in EN 60286-16 1998. The original data sample was therefore re-analysed and the STI 'male' and STI 'female' values computed[6]. Very little difference was found to occur and the mean values for the 81 samples varied by only +/- 0.01 from the original STI. The 25 system sub set was then also re analysed but this time the +10 and + 6 dBA S/N ratio data was also included. This time some immediate differences between the data were apparent. in approximately $^2/_3$ of the cases the 'male' weighting increased the overall STI slightly. Fig. 10

[6] Note by the editor: the 'male' and 'female' weightings referred to by the author, belong with the revised version of STI, named STI_r (chapter 3 of this book, IEC 60268-16 2nd edition, 1998). The difference with the 'old' STI is not just the distinction between speaker gender. As shown in chapter 3, the STI_r correlates better with subjective intelligibility across a wide range of conditions.

shows the effect of applying the 'female' weighting. Here the scatter increases significantly – particularly in the presence of noise. The overall effect of the weighting is to increase the STI value. In 20 % of the cases, the increases with noise are in the region of 0.1 STI. As seen earlier, this is a very significant improvement with respect to a voice alarm criterion of 0.5 STI for example.

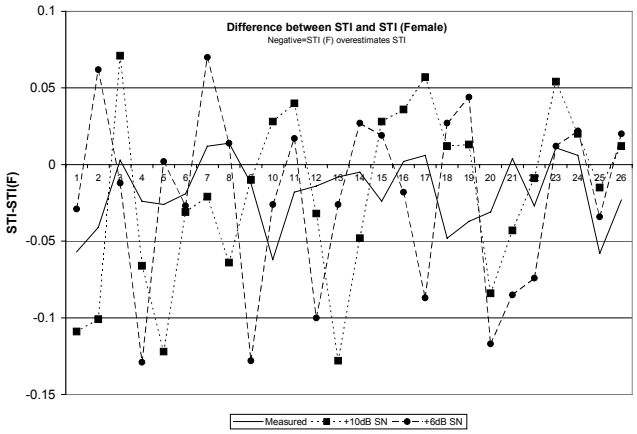


Figure 10. Comparison of STI and STI 'female' weighting of data Reverberation + Noise

8.3.2. Effects of signal processing

Most modern sound systems employ some form of signal processing, which can be either digital or analogue, or a combination of the two. Typical processing includes, equalisation, compression, filtering, and limiting. Such processing may not necessarily be obvious, as it may be hidden within either a general DSP unit also employed for routing and signal distribution or within devices such as dedicated loudspeaker controllers. AGC (automatic gain control) and automatic noise sensing and gain adjusting circuits may also be incorporated into paging and voice alarm systems. The microphone stations on such systems may also incorporate compression, limiting or expansion circuitry. Some sound systems may also incorporate feedback suppression circuitry or processors. Level dependent frequency response shaping may also occur. The result is that many systems are highly non-linear. Whereas it may be possible to bypass or switch out some processing, this is often not possible in practice when auditing a completed system. Some brief experiments were therefore carried out in order to obtain an idea of the potential problem. Different measurement techniques were also tested so that it would be possible to differentiate between measurement errors or artefacts as opposed to underlying discrepancies with STI itself.

The effects of equalisation were effectively dealt with in the preceding section 3.1 on frequency response. Compression is another very common form of processing. In that it deliberately changes the dynamic range and crest factor of the signal, it would seem likely that such processing could affect the STI. Compression was applied to a variety of STI test stimuli including MLS, TDS sine sweep and the modulated STIPA signal. A maximum of 6 dB was initially applied as this represents a typical value used in practice. No effect was found to occur with either the MLS or TDS signals but a noticeable reduction in STI did occur with the STIPA signal. Typically, on average, the STIPA of a 'straight wire system', reduced to around 0.74 i.e. a reduction of 26 %. This is a very significant reduction, particularly as it occurs before the transmission of the test signal to the loudspeaker and out into the space. It means for example that an intelligible system, with a nominal STI of say 0.55 – well inside the

usual safety system criteria of 0.5 STI (0.7 CIS) would reduce to 0.29 STI or 0.46 CIS – a total failure with indications of being almost totally unintelligible! The need to check that the system is behaving totally linearly is therefore essential when using such a signal. (A similar effect was also found to occur with an amplitude modulated RASTI signal). Some further tests are currently being conducted to quantify the problem for a wider range of situations.

Feedback is a limitation of almost any sound reinforcement system and various techniques have been developed to improve the gain before feedback margin. Currently, broadband equalisation used in conjunction with narrow-band notch filtering is probably the most common technique and generally employs digital filters. Other techniques include phase and frequency shifting and signal decorrelation. Tests on typical digital filter set ups resulted in STI reductions of around 0.05 STI when using MLS signals and so again care clearly needs to be taken when evaluating systems with nominal values in the region of 0.5 STI. However, little or no difference was found when using the STIPA signal. A number of tests were also conducted on phase/frequency shifters. Using the impulse based techniques e.g. MLS/TDS led to completely anomalous results being created with the received signal either being regarded purely as noise or as a highly distorted artefact. MLS based signals for example computed STIs of around 0.03 to 0.04! Clearly, there is a need to either bypass or deactivate such devices when testing. However, at least in this case the result is obviously wrong, whereas with other forms of processing, this may well not be the case. Interestingly, no problem was encountered when the tests were repeated with the STIPA signal and analyser, which completely ignored the frequency/phase shifting processing. Further information on potential system and measurement errors can be found two other papers by the author (Mapp, 1996; Mapp, 2002).

8.4. CONCLUSIONS

1. It has been shown that considerable discrepancies occur between the STI and RASTI values when measuring or verifying the performance of typical sound systems. Of the 81 systems studied, under reverberant and high S/N ratio conditions, RASTI underestimated the STI in 66% of the cases. However a brief analysis of the data showed there to be no immediately apparent consistent mechanism responsible for this.
2. In the presence of background noise at +10 and +6 dBA and reverberation, the above trend was completely reversed and in the vast majority of cases RASTI overestimated the STI.
3. In the case of low reverberation and background noise, RASTI was also found to significantly overestimate the STI.
4. When high levels of harmonic distortion were present (i.e. signal clipping) RASTI was found to underestimate STI.
5. The findings clearly show that RASTI is an unreliable predictor of STI when measuring the performance of sound systems. This has been shown to be the case over a wide range of typical conditions and system formats.
6. Both anecdotal experience and limited subjective testing using 3 different sets of conditions and test subjects, indicate that STI does not appear to accurately predict intelligibility for low noise, reverberant conditions with irregular or falling high frequency responses.
7. Good agreement was found to occur between impulse based STI measurements (as measured using MLSSA and TEF TDS techniques) and the recently introduced STIPA modulated signal with a sparse MTF analysis matrix. This indicates that the discrepancies noted above are inherent within STI as opposed to being a measurement artefact or systematic measurement error.

8. Good agreement was found to occur between STI and word scores for systems operating under low reverberation conditions but in the presence of background noise (> 10dBA).
9. Little difference was found to occur between conventional STI and the male weighted STI_r under reverberant and reverberation with noise conditions. However the female weighting did produce slightly improved values, with a mean improvement of 0.02 for the low noise reverberant condition. In the presence of background noise and reverberation, the scatter increased significantly. Increases of 0.1 STI were noted in twenty percent of the cases when background noise was present and overall there was an increase in the STI for sixty percent of the samples when the female weighting was applied.
10. Typical sound system signal processing was found to have stimulus dependent effects and anomalies. Whereas for example, simple amplitude compression affected a modulated stimulus (as would be expected) up to 6 dB of such compression had little or no effect on MLS or TDS based signals and measurement systems. Equally however, time variant processing had no discernible effect on STIPA but very badly affected MLS and TDS FFT based analysis to the extent that no useful signal could be recovered.
11. It was noted that some forms of A-D converter and processing could affect MLS based STI measurements, with reductions of around 0.05 being typical. No such reductions in STIPA were found. This suggests that the STI of the electronic signal chain should always be measured when using an MLS based measurement and the overall acoustic performance result corrected.
12. The research shows that although STI is a good predictor of sound system intelligibility under many conditions, measurements made under low noise reverberant conditions on systems exhibiting irregular frequency responses may give rise to erroneous results with the potential intelligibility being over estimated.
13. The above research findings raise a number of interesting contractual and legal implications relating to system verification and performance measurement.
14. Further research is required to increase the accuracy of STI for a wider range of typical sound system responses and operating conditions.

REFERENCES

CAA, (**1989**) 'Public Address Systems' Specification no. 15.

IEC 268 pt 16 (**1988**) The objective rating of speech intelligibility in auditoria by the RASTI method. Later to become EN 60268-16, (**1998**). Objective Rating of Speech Intelligibility by Speech Transmission Index.

IEC 849 (**1989**) (later EN60849) and BS 7443 1990 (later BS EN6084 : 1998) 'Sound systems for emergency purposes'.

Barnett, P, (1997) Implications of Amplitude Compression on RASTI Performance. Proc IOA Vol 19, Pt 6. p. 213

Jacob, K, Steeneken,H, Verhave, J, McManus, S, (**2001**) ' Development of an Accurate, Handheld Simple-to-Use, Meter for the Prediction of Speech Intelligibility. Proc IOA Vol 23 Pt 8.

Mapp, P (**1991**), reprinted in Handbook for sound engineers 2nd edition, Ed Ballou, Pub. Focal Press, Chapter 32, Speech Intelligibility by Davis & Davis.

Mapp, P (**1996**), 'A Comparison between STI and RASTI Speech Intelligibility Measurement Systems'. 100th Convention AES, Copenhagen.

Mapp, P, (**1997**), 'Limitations of Speech Intelligibility Methods.' 133rd Meeting ASA, Pennsylvania,

Mapp, P, (**1997**), Some Effects of Equalisation and Spectral Distortion on Sound System Intelligibility. Proc IOA Vol 19, Pt 6. p. 245

Mapp, P, (**2001a**), 'Limitations of current sound system intelligibility performance measurement techniques and metrics'. 142nd Meeting ASA Fort Lauderdale, Florida.

Mapp, P, (**2001b**), 'Improving the Intelligibility of Aircraft PA Systems'. 111th AES Convention New York.

Mapp, P, (**2002a**) 'The measure of Intelligibility', Sound &Video Contractor, Vol **20** No 4,

Mapp, P, (**2002b**) 'Relationships between Speech Intelligibility Measures for Sound Systems', 112th AES Convention Munich.

Mapp, P, (**2002c**) 'Limitations of current sound system intelligibility verification techniques' 113th AES Convention Los Angeles.

RTCA (**1993**) 'Audio Systems Characteristics and Minimum Operational Performance Standards for Aircraft Audio Systems and Equipment', Document no. RTCA/DP-214.

Chapter 9. Measurement and prediction of speech intelligibility in traffic tunnels using the STI

Sander J. van Wijngaarden and Jan A. Verhave

ABSTRACT

Traffic tunnels in the Netherlands are generally equipped with public address systems. Tunnels are difficult environments for reproducing speech, because of their physical dimensions, lack of acoustic absorption, and high ambient sound levels. This chapter describes how in-situ measurements of the STI are used to accept or reject PA-systems in tunnels. It also indicates how acoustic prediction models can be used to make design choices.

9.1. INTRODUCTION

Designers of Public Address (PA) systems know that a number of signal degrading factors threaten the intelligibility of the speech that the system produces. The main factors are reverberation, ambient noise, audio bandwidth limiting and signal distortion. In all of these aspects, traffic tunnels represent more or less the worst-case situation.

The reverberant sound field in tunnels can be very strong, because of the physical dimensions of tunnels, but also because surfaces in tunnel are not designed to absorb acoustic energy. Materials with high absorption coefficients are usually difficult to clean, not very water resistant, sensitive to exhaust fumes, or otherwise unsuitable for an outdoors environment. Reverberation times longer then 10 seconds in the lower octave bands are quite often observed.

The ambient noise caused by moving traffic, reinforced by the tunnel acoustics, is usually considerable (80 – 95 dBA), while high-power ventilators used for smoke, fumes and temperature control, can even be worse noise sources (sometimes over 105 dB(A) at worst-case locations).

The electro-acoustic components must also be chosen to be durable in an aggressive outdoors environment. This calls for rugged and robust horns, powerful enough to produce sound levels that can compete with the noise. Unfortunately, this is not the type of loudspeaker that features the wide range frequency response that is desired for high-quality speech reproduction. Moreover, at the high sound levels that need to be produced, non-linear distortion components introduced by these loudspeakers have a considerable impact.

9.2. TYPICAL DESIGN OF A TUNNEL PA SYSTEM

Although PA systems are usually specifically adapted to each tunnel environment, a general schematic representation is given in Fig. 1.

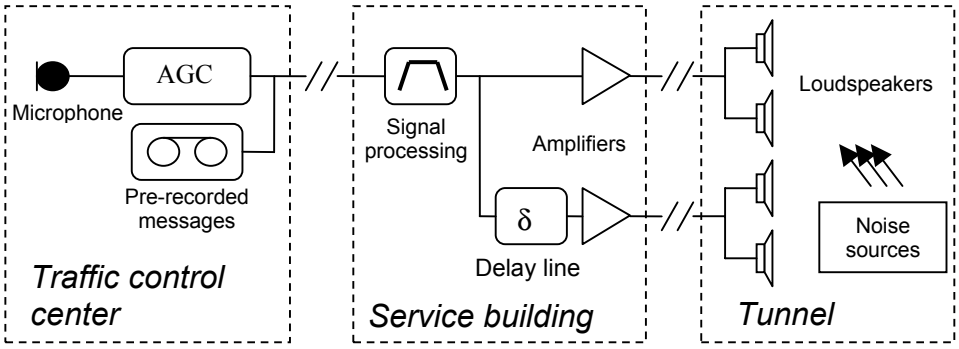


Figure 1. Schematic representation of a PA system in a traffic tunnel

Tunnels are normally divided into sections of 100 – 200 m in length. Each individual section is normally covered by a single closed-circuit TV camera. These cameras are used by traffic control center personnel to monitor whatever is happening in the tunnel. The PA system is usually designed to address one section at a time. Each section contains two or three arrays of loudspeakers (sometimes, in special cases, even just one). Fig. 2 shows an example of the arrangement of loudspeaker arrays in a tunnel.

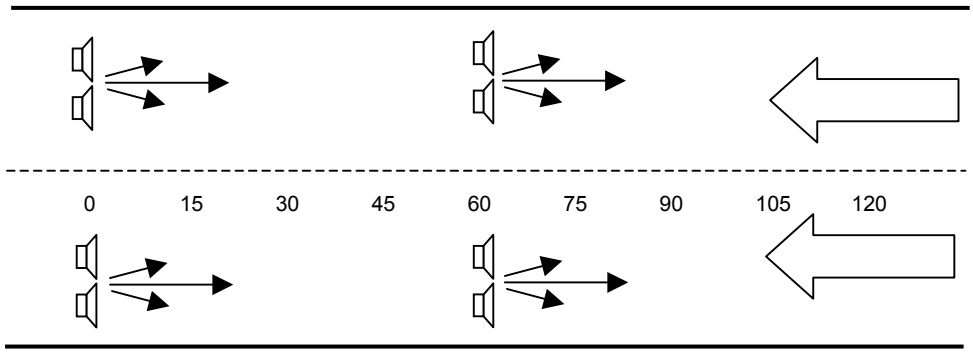


Figure 2. Example of an arrangement of loudspeaker arrays in a two-lane traffic tunnel, with a section length of 120m. The traffic on both lanes is travelling in the same direction, indicated by the large arrows. Clusters of loudspeaker are mounted on racks attached to the ceiling, above the middle of each lane, and aimed toward the moving traffic. At the position marked "0" the signal is presented without delay; the sound reproduced by the clusters 60 m further away is delayed.

The physical delay associated with acoustic wave propagation requires the use of delay lines to make sure that the sound from different speakers arrives more or less aligned in time. This also requires the loudspeakers to be highly directional.

9.3. MEASURING THE STI IN TUNNELS

9.3.1. Minimum criteria in the Netherlands

Until recently, the applicable standards of the Netherlands Ministry of Transportation called for a minimum STI of 0.35, at any position in the tunnel. A limit of STI=0.35 is much lower than most experts would like to see, in view of the importance of the PA system in

emergencies. Emergency drills have shown that a PA system that just barely complies with the STI=0.35 limit, is of little practical use when you really need it. For this reason, and also because of international standardisation developments (ISO 9921), the current limit is set at an average STI of 0.45, averaged across a representative section of the tunnel (also see chapter 10 of this book). By working with an average STI, instead of a hard minimum that must be met at any position, the measurement protocols for compliance testing are simplified; when working with a hard limit, the measurement error (typically 0.02 to 0.03 STI for a single STI measurement) sometimes leads to extensive re-measuring of single points. When working with an average criterion, the measurement error is averaged out across the whole tunnel section.

9.3.2. Outline of the measuring protocol

The first step in verifying compliance with minimum STI criteria, is to select the conditions under which this minimum must be reached. The main variables are potential listener positions and noise sources. Will the PA system be used when the traffic is moving, or only when the tunnel is blocked up in an emergency? Are the listeners expected to be inside or outside of their vehicles? These questions can only be answered by considering the scenarios in which the PA system is used. For all of these scenarios ('emergency', 'tunnel maintenance', etc.) the noise sources and possible listener positions need to be identified, and the minimum STI needs to be decided.

The next step is to actually measure the STI under all of the conditions, and at all of the positions, indicated by the scenarios. This generally involves setting up a measuring grid, and moving a measuring microphone along this grid while carrying out STI measurements. Noise sources (real or simulated) are switched on and off, depending on the requirements following from the scenarios.

Practice has shown that measuring the STI at positions 10-15 meters apart along the length of the tunnel, in the middle of each lane, yields a sufficiently dense measuring grid. In all cases, STI measurements are also carried out with all noise sources switched off, even when none of the scenarios predicts that this situation occurs in practice. From 'clean' STI measurements (no noise, just reverberation and signal distortion effects), STI values in noise with an arbitrary spectrum and level may be derived mathematically.

9.4. PREDICTING THE STI USING RAY-TRACING SIMULATIONS

Simulation models are not suitable for verifying compliance with a minimum standard., even when a validated model is used. The outcome of these models relies on the correctness and accuracy of many of the underlying assumptions, such as acoustic absorption of the materials used, and directivity of the loudspeakers. Moreover, validated prediction models that are capable of dealing with the peculiarities of a tunnel, which may be 10 km in length, with a tilted road surface and curved along two dimensions, are hard to find.

However, simulation can be a powerful tool to predict how effective certain measures can be in improving speech intelligibility. Simulation of well-chosen scenarios can result in much better design decisions, even very early on in the PA system design process.

A specific issue with commercially available acoustic simulation packages is the way the Modulation Transfer Function (MTF) is calculated. Acoustic simulation packages are invariably capable of prediction the impulse response between a (simulated) source and receiver. The MTF can be calculated in a straightforward manner from the (squared) impulse response (Houtgast, Steeneken and Plomp, 1980), after which the IEC-standardised procedures can be used to calculate the STI. Unfortunately, many software packages choose to calculate the MTF from the well-known approximation, also given in the IEC standard,

that relates the MTF to the reverberation time in case of pure exponential decay. This approximation leads to unacceptable errors in some cases, due to echoes from reflective surfaces, or sharp bumps in the decay curve due to backward propagation from a delayed loudspeaker array.

Another complication is that, even to the present date, some software packages calculate the RASTI rather then the "full" STI. RASTI is not suitable for tunnels. The bandwidth limiting due to the use of horn speakers alone results in considerable errors when using RASTI.

Because of these complications, TNO developed a set of software procedures to calculate the STI from a predicted impulse response. This means that simulated impulse responses exported from any simulation package can be used (we used Odeon; Lynge, 1998). These software tools were specifically optimised for application to traffic tunnels.

To validate the prediction tools, STI values from simulations were compared to results from in-situ measurements. Fig. 3 shows the correlation between measured and predicted STI values for 154 measuring points, differing in tunnel geometry, surface materials, noise spectrum, loudspeaker type, loudspeaker position and listener position. Fig. 3 shows that the procedure is effective and sufficiently accurate in predicting STI values. This conclusion is supported by Fig. 4, which shows measured and predicted STI values (in a single condition without noise) for various positions in the same tunnel. The predictions of Figs. 3 and 4 were obtained without using a priori knowledge of measurement results for the corresponding conditions.

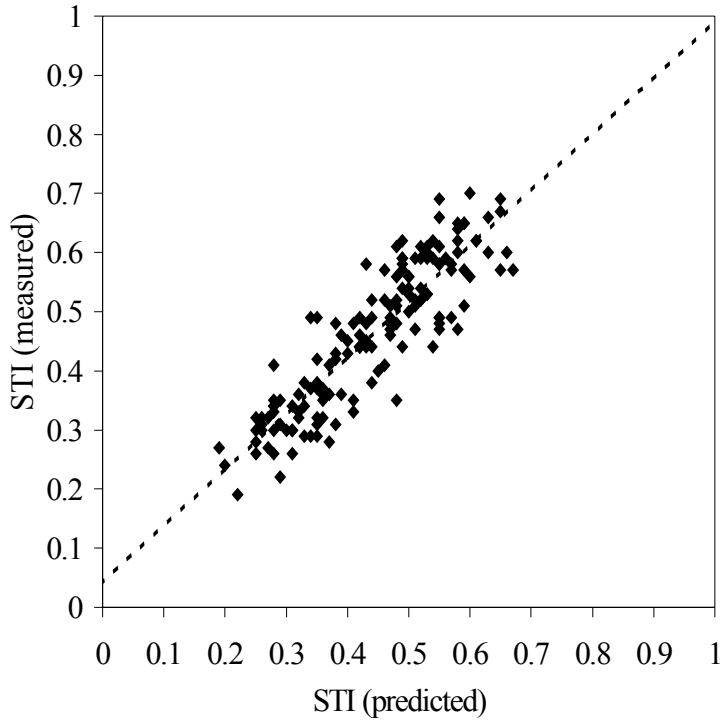


Figure 3. Correlation between predicted and measured STI values (correlation coefficient r=0.89).

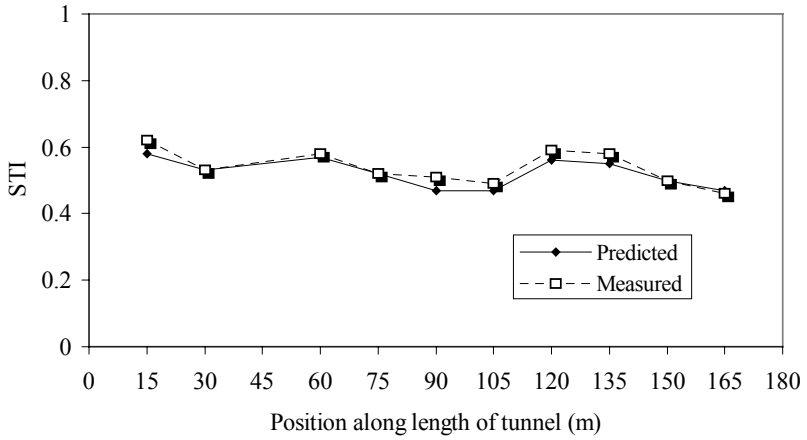


Figure 4. Measured and predicted STI values for several listener positions (indicated in meters along the length of the tunnel) for a condition without noise, at a height of 1.50 meter above the road surface.

Despite the good results shown in Figs. 3 and 4, simulation is still no replacement for measurement. The fact that there is no way to verify the validity of the model input before one can actually measure in a tunnel, makes simulation results inherently less reliable than measurement results.

9.5. PREDICTING THE STI FOLLOWING AN EMPIRICAL REGRESSION APPROACH

The acoustic tunnel environment differs in quite a few ways from the more commonly studied environments (such as concert halls and classrooms). On the other hand, the acoustic characteristics of most tunnels (at least traffic tunnels in the Netherlands) seem to be very similar from one tunnel to the next. Differences observed between tunnels are readily explained from known parameters, such as the tunnel dimensions, equivalent "open window area" of surface materials, and the characteristics of the loudspeakers.

Acoustic simulation models, depending on computationally expensive algorithms such as ray-tracing or source-image modelling, require either a lot of time or a powerful computer. As computers are becoming faster, this is becoming less of an issue. However, there is always something to be said for "light-weight" simulation tool. Such a tool was developed using a multiple regression approach.

Using available data of 17 different configurations (either measured data, or obtained with validated simulation tools), linear equations for predicting modulation indices were derived statistically. The equivalent speech-to-noise ratio and sound level were predicted separately for each octave band.

$$S = a_{(0)} + a_{(1)}A + a_{(2)}O + a_{(3)}Q + a_{(4)}I + a_{(5)}\log(\frac{D}{15}+1) + a_{(6)}D + a_{(7)}P + a_{(8)}(\frac{N}{Q})^{a_{(9)}} + a_{(10)}F + a_{(11)}N \quad (1)$$

$$L = b_{(0)} + b_{(1)}A + b_{(2)}O + b_{(3)}Q + b_{(4)}I + b_{(5)}\log(\frac{D}{15}+1) + b_{(6)}D + b_{(7)}P + b_{(8)}N^{b_{(9)}} + b_{(10)}F \quad (2)$$

By applying the linear prediction equations (1) and (2) for equivalent SNR and level in the octave bands 125 Hz – 8 kHz, the STI can be calculated. Values of the parameter a(n) and b(n) were determined using multiple linear regression. The variables in Eqs (1) and (2) are explained in Table I.

Table I. Explanation of variables in Eqs. (1) and (2)

Variable	Explanation
S	Equivalent speech-to-noise ratio
L	Sound level at listener position due to the PA system
A	Area of the tunnel (tube) cross-section
O	Acoustic "open window area" per meter tunnel length
Q	Octave band dependent Q-factor of the loudspeaker
I	Octave band index: 1 for 125 Hz, 7 for 8 kHz
D	Forward distance to the nearest loudspeaker array
P	Absolute position within each loudspeaker section
N	Backward distance to the nearest loudspeaker array
F	Linear octave band centre frequency (125–8000)

Values of the parameters a(n) and b(n) are only valid when the variables are restricted to carefully determined boundaries. In short, the values of all variables have to be within the limits represented by the set of tunnel configurations from which the statistical parameters are derived.

Using Eqs (1) and (2), a low-cost model for carrying out rough calculations is easily programmed. Such a model is limited in flexibility and accuracy. The benefit is that STI prediction results can be obtained almost instantly on any PC, without the need to use CAD tools for modelling the tunnel geometry.

9.6. DISCUSSION AND CONCLUSION

The STI measuring method is well-suited for the evaluation of PA systems in complex acoustic environments. Experiences with traffic tunnels have shown that compliance with minimum standards for speech intelligibility can be tested reliably using the STI method, even under the hardest conditions. All that is really needed is hard- or software capable of STI measurements, and a good measuring protocol.

STI predictions are useful for optimising intelligibility, but not as suitable for verification against minimum standards. Sophisticated (validated) simulation tools will produce accurate predictions of the STI, but the validity of the input parameters of the model will always remain an issue. For rough estimates of intelligibility, when predictions are needed quickly for a great number of conditions, a statistically derived regression model can be used.

REFERENCES

Houtgast, T, Steeneken, H.J.M. and Plomp, R (**1980**)., "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics", Acustica **46**, 60-72.

Lynge, C. (**1998**) "Odeon room acoustics program user manual", Department of Acoustic Technology, Technical University of Denmark,

Wijngaarden, S.J. van & Verhave, J.A. (**2001**). The influence of fan and traffic noise on speech intelligibility in Dutch traffic tunnels. In: Proc. Internoise 2001.

Chapter 10. Standardisation of performance criteria and assessment methods for speech communication

Herman J.M. Steeneken

ABSTRACT

The purpose of standardisation of the performance criteria for speech communication, is to assure a certain level of speech communication quality, as required for various applications. The quality of speech communications is routinely assessed in cases where warning, danger, or information messages may be produced; at work places, but also in 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.

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

10.1. INTRODUCTION

Safe operation of warning and alert systems requires a certain level of speech communication quality. Therefore, generally accepted criteria and performance measures have to be used. This is the goal of standardisation 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 contrast 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 standardisation network, such as ANSI (American National Standards Institute).

The communications may be directly between humans, through public address or intercom systems, or by using pre-recorded 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 acceptation or rejection. If two third of the supporting nations have voted in favour 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.

10.2. SELECTION OF CRITERIA FOR SPEECH COMMUNICATION QUALITY

A commonly posed requirement for the understanding of spoken messages is a correct recognition of each utterance. In technical terms, this 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).

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).

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

10.3. 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 syllable recognition, and gives a qualification of the predicted speech intelligibility. Unlike the other methods, the STI also 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 2nd edition, ANSI S3.5].

10.4. 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 meet 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_{EQB}-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[1] score %	CVC_{EQB}-nonsense word score %	STI	SIL[2] dB	SII[3]
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.

[1] According to Anderson and Kalb (1987).

[2] Qualification of the SIL-method is converted to a five-point scale rather than the original seven-point scale.

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

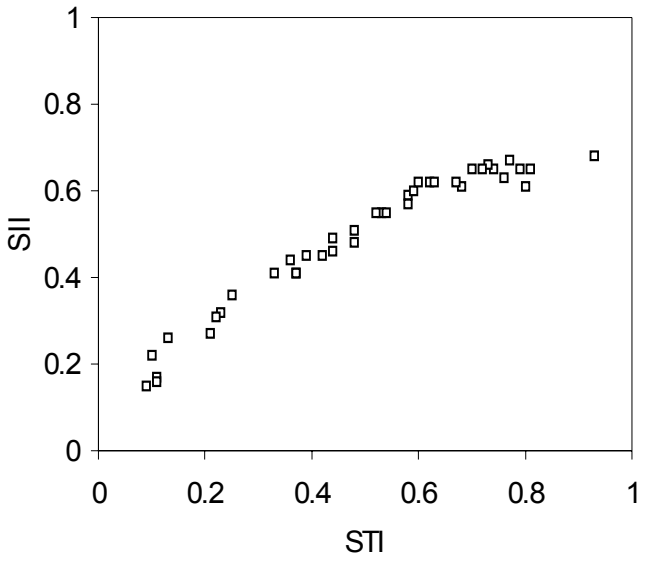


Figure 1. Relation between STI and SII. The correlation coefficient r = 0.93.

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.

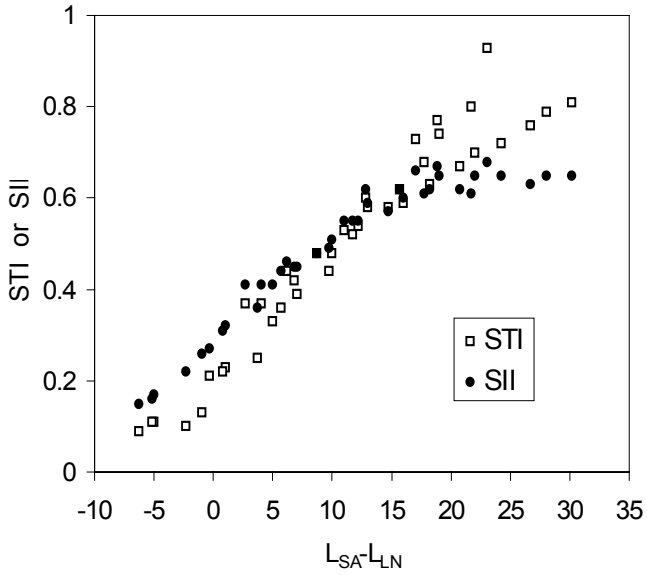


Figure 2. Relation between SIL, STI and SII. The correlation coefficients between SIL and STI is r=0.97 and between SIL and SII r=0.95.

10.5. 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 11 September 2001, 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."

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 if 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.

Chapter 11. Computing the STI using speech as a probe stimulus

Karen L. Payton [a], Louis D. Braida [b], Shaoyan Chen [a], Peninah Rosengard [c] and Raymond Goldsworthy [c]

[a] ECE Dept., UMass Dartmouth; [b] EECS Dept., MIT; [c] Harvard-MIT Health Science & Tech. Prog.

11.1. INTRODUCTION

The purpose of the STI is to predict the preservation of speech intensity modulations after transmission through an acoustic environment. Our research efforts have focused on determining the extent to which those modulations can be measured directly from speech waveforms. From our work and others, we will report how well a variety of speech-based Modulation Transfer Functions (sMTFs) match the more traditional room-acoustics based approach. In addition, we will present and discuss MTF data for conditions not easily predicted by the traditional approach and compare the resultant STI values with subject intelligibility scores in those environments.

We have a variety of motivations for pursuing a speech-based MTF. First, in the long run we are interested in predicting intelligibility differences due to speaking style. Clearly articulated speech has been shown to be more intelligible than conversationally articulated speech in noise and reverberation by both normal-hearing and hearing-impaired listeners (Picheny et al., 1985, Payton et al., 1994). Envelope spectra (intermediate byproducts of the STI computations) are one of the few overall acoustic measures that differentiate between the speaking styles. Although Steeneken and Houtgast (1983) appear to show that different speaking styles have very similar envelope spectra we, in contrast, have found that clearly articulated speech typically has envelope spectra with larger contributions at low modulation frequencies (less than 5 Hz) than conversationally articulated speech. Fig. 1 (adapted from Fig. 3.6 in Krause, 2001) depicts averaged envelope spectra of one talker speaking nonsense sentences three different ways: Conversationally articulated, clearly articulated and clearly articulated but at the faster conversational speaking rate[1],[2]. While some of the differences between the clear envelope spectra (dashed lines) and the conversational envelope spectra (solid lines) may be due to the slower production rate for clear speech, differences are also apparent when the talker speaks clearly at conversational speaking rates (dotted lines).

A second reason to investigate a speech-based STI is to study the response of environments not well characterized by modulated noise probe stimuli such as that used in the RASTI procedure. An example that has been of interest to several researchers (e.g. Plomp, 1988, Villchur, 1989, Ludvigsen et al., 1990, Ludvigsen, 1993, Hohmann &

[1] Nonsense sentences are grammatically correct but syntactically meaningless, e.g. "A right cane could guard an edge."

[2] All data presented were computed on a corpus of 50 nonsense sentences, which were concatenated together to simulate running speech. The total duration of the speech materials was about 1.5 minutes.

Kollmeier, 1995) is predicting the intelligibility of amplitude-compressed speech, such as that produced by many modern hearing aids.

Normal-hearing listener thresholds are incorporated as internal noise levels in the IEC standard for computing the STI (IEC, 1998). Humes et al. (1986) showed that representing elevated listener thresholds as internal noise also allows the STI to predict hearing-impaired subject performance quite well. Given that the listener could be adequately represented in the STI computations, it would be useful to be able to predict the performance of a wide range of hearing aid processors, including those with amplitude compression. A predictor correlated with intelligibility would allow one to screen a wide variety of compression parameters such as attack and release time constants and compression ratios for a variety of hearing-loss profiles without requiring the use of human listeners for every condition.

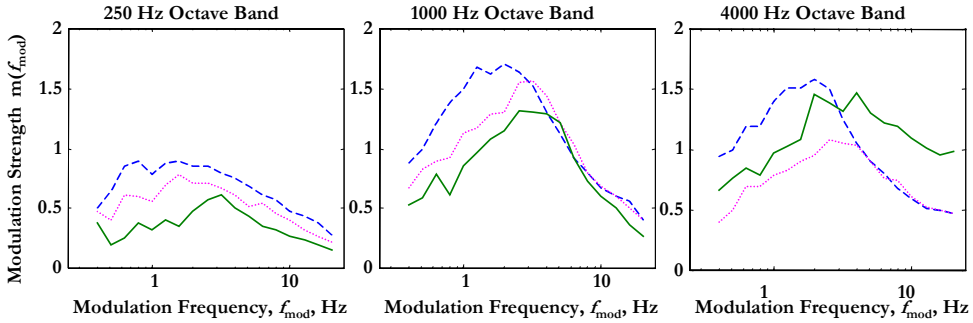


Figure 1. Comparison of clear (dashed lines), conversational (solid lines) and fast clear speech (dotted lines) envelope spectra in three octave bands, from Fig. 3.6 of Krause, 2001.

Other researchers (e.g. Lundin, 1982, Ludvigsen et al., 1993, Hohmann & Kollmeier, 1995, Holube & Kollmeier, 1996) have attempted to use a speech-based STI to predict the intelligibility of amplitude compressed speech without success. Given that their results were due, in part to artifacts (discussed in the next section), this topic is worth additional study.

11.2. BACKGROUND

While the primary focus of the STI has been to use artificial probe stimuli to evaluate room acoustics, as early as 1972 Houtgast and Steeneken published a conference paper that discussed envelope spectra of octave-band filtered speech and the influence of room acoustics on the envelope spectra (Houtgast & Steeneken, 1972). That paper was the first to show the specific effects of noise, reverberation and echoes on speech envelope spectra. The paper also showed one of the limitations of using speech envelope spectra directly: the existence of a noise floor. Although not apparent when speech was presented in quiet, the envelope spectrum due to the addition of noise did not fall below the noise floor. In later papers, the noise floor issue was not discussed although it was often mentioned that speech could be used as a probe stimulus (Steeneken & Houtgast, 1980, 1983, Houtgast & Steeneken, 1985).

In most cases, the STI using speech was computed in a manner very similar to that using modulated noise. First, the signal was filtered into seven octave bands. The averaged magnitude of the output envelope spectrum in each band was divided by the averaged magnitude of the input envelope spectrum in that band. The only additional computation was binning of the envelope spectra magnitudes over $^1/_3$ octaves. One exception was in Steeneken and Houtgast (1980) where they suggested using speech in all the bands NOT under test and a modulated noise signal be used in the band under test.

As has been found by several researchers, the direct application of speech as a probe stimulus can be problematic. The noise floor is one issue, although the noise floor is not unique to speech stimuli. It is also present when modulated noise is used as a probe stimulus. It is not a problem in modulated noise since noise is initially 100% intensity modulated and the modulation components are therefore well above the noise floor, even when severely degraded. To demonstrate this, Fig. 2 depicts the envelope spectra of speech-spectrum noise; intensity modulated at either 100% or 50% of modulation depth in 3 octave bands. The modulation frequency is 3 Hz The noise floor in each octave band has a different characteristic slope due to the bandwidth of the band. The lower bands are narrow and consequently have more residual noise modulation. The noise floor level increases with modulation frequency due to the summing across $^1/_3$ octaves.

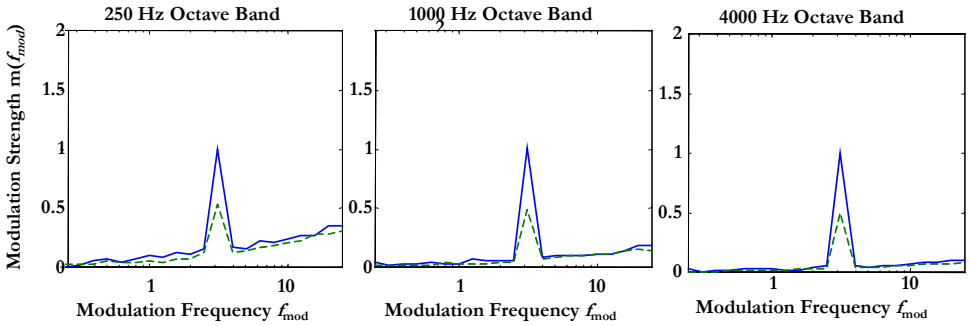


Figure 2. Envelope spectra of speech-spectrum noise intensity modulated at 3 Hz. Modulation is either 100% (solid lines) or 50% (dashed lines). The 50% condition is representative of the 100% modulation signal after a 0dB SNR environmental degradation.

The noise floor impacts speech MTFs because speech modulations are well below 100% at most modulation frequencies, particularly above 5 Hz. The consequence is that the noise floor comes into play at moderate signal to noise levels. Figure 3 depicts envelope spectra of conversational speech, conversational speech in 0dB SNR speech-shaped noise and the envelope spectra of the noise alone. In each octave band, the speech-in-noise envelope spectrum is an attenuated version of the speech envelope spectrum in that band until the noise floor level is comparable in value, then it tends to follow the noise floor envelope spectrum.

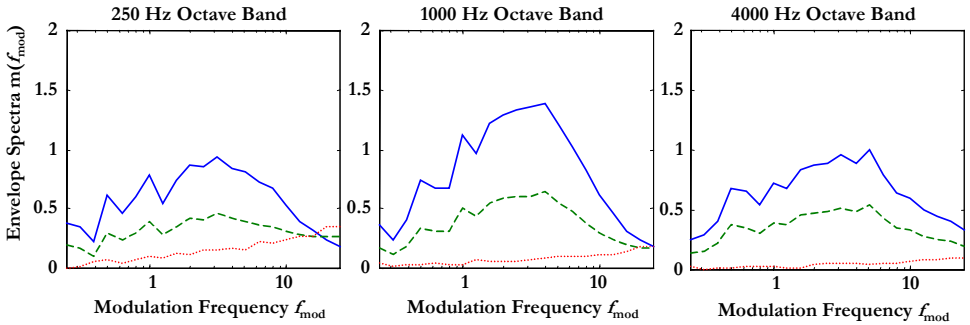


Figure 3. Averaged envelope spectra of speech (solid lines), speech plus noise at 0dB SNR (dashed lines) and unmodulated noise (dotted lines) in three octave bands.

A second issue that arises when using speech as a probe stimulus is what happens to the envelope spectra in the presence of reverberation. Theory predicts a low-pass filtering effect on the MTF, with the roll-off modulation frequency dependent on the amount of reverberation. When speech is used as a probe stimulus, the roll off starts to happen, but there are dips or, in some cases, increases in the sMTF at the higher modulation frequencies. Fig. 4 shows speech envelope spectra with and without reverberation. At low modulation frequencies, the two are very similar. At higher modulation frequencies, the speech-in-reverberation envelope spectra start to decrease relative to the clean-speech envelope spectra but then they rise and eventually most become greater than the clean-speech envelope spectra, particularly in the lowest octave bands. This phenomenon is less well understood than the noise-floor problem, but is most likely due to contributions from isolated, early, room reflections and to the quasi-periodic nature of voiced speech. For the reverberation condition studied, early reflections are isolated and impulsive. Only the late reflections exhibit exponential decay. In the presence of reverberation, early reflection arrivals could either degrade or reinforce voiced speech segments. We have found high-frequency ripples in the intensity envelopes of voiced reverberant speech when none were present in the original intensity envelopes. In Fig. 5, for example, five peaks occur in the reverberant speech between .1 s and .2 s while only 2 exist in the original waveform during that time interval. While reinforcement of the voiced segments might be beneficial to intelligibility, listening tests do not substantiate such a hypothesis. Including this portion of the sMTF in an STI computation results in an over-prediction of intelligibility for that environment.

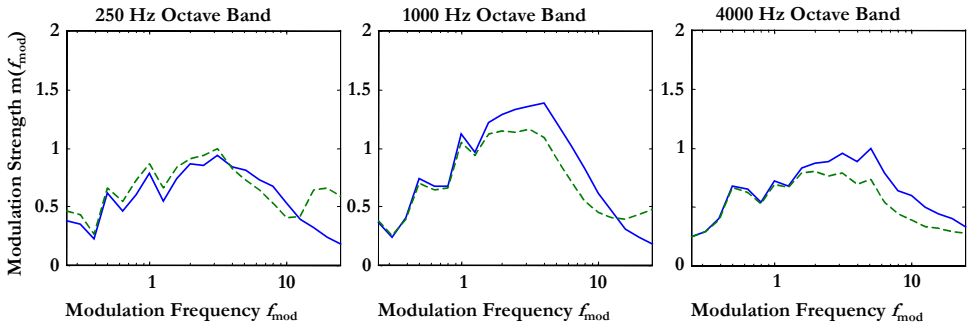


Figure 4. Envelope spectra of speech before (solid lines) and after reverberation (RT=0.6s) (dashed lines) for 3 octave bands.

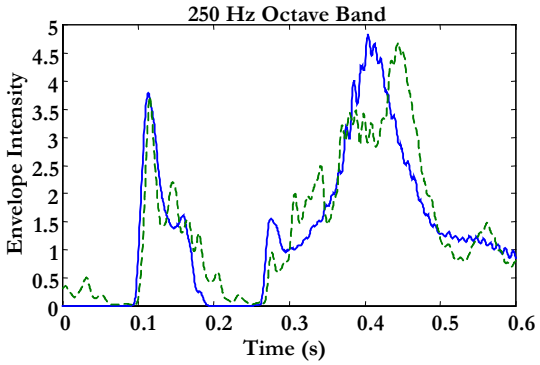


Figure 5. Segment of voiced speech intensity envelopes before (solid line) and after (dashed line) reverberation (250 Hz octave band).

To mitigate such problems, several different techniques have been proposed to compute an STI using speech as a probe stimulus. Ludvigsen et al. (1990) proposed computing the STI using speech based on the time-domain intensity envelopes in 7 octave bands without computing envelope spectra. They computed the cross correlation between input and output envelopes and formulated their version of an apparent SNR from the slope and intercept terms of the cross correlation. For situations with stationary noise and/or reverberation, their resultant STI matched the Houtgast and Steeneken STI very closely. Holube and Kollmeier (1996) used a similar technique but filtered the speech into 23 frequency bands instead of 7 bands.

Drullman et al. (1994) proposed a phase-locked MTF based on the ratio of the real part of the cross power spectrum (between input and output speech envelopes) to the input envelope power spectrum of octave-band filtered speech. They argued that, for their nonlinear processing conditions, only the portion of the envelope spectrum in phase with the input envelope was relevant for intelligibility prediction.

In our own work on the use of speech as a probe stimulus, we initially considered modifications to the original Houtgast and Steeneken technique (Payton et al., 1993, Payton & Braida, 1999). We initially focused on determining a modulation frequency at which we would truncate the MTF. The modulation frequency range considered in the earliest Houtgast and Steeneken papers was 0.25 Hz up to 25 Hz, (e.g. 1973). In later papers, the STI was computed based on a reduced modulation frequency range: 0.63 Hz to 12.5 Hz (Steeneken & Houtgast, 1980, Houtgast & Steeneken, 1985). Since hearing-impaired individuals are affected by mild to moderate reverberation, we wished to preserve as much of the MTF frequency range as possible. We noted that the coherence between input and output intensity envelopes dropped off rapidly at modulation frequencies where there were artifacts due to noise and/or reverberation. Using the coherence as a cutoff criterion, we were able to generate sMTFs that were very similar to the theoretical predictions over the modulation frequency ranges where coherence was high. The problem with this approach was that it resulted in some octave bands having very few MTF values.

We then decided to reconsider the Drullman and Ludvigsen techniques to understand how they related to each other and to the Houtgast and Steeneken method. Both Ludvigsen's and Drullman's approaches were important because they attempted to address the noise floor and reverberation problems and because they used the STI to predict the intelligibility of nonlinearly processed speech.

Work done by Goldsworthy (2001) has shown the mathematical relationship between the apparent SNR obtained via Ludvigsen's correlation method ($aSNR_L$) and the apparent SNR obtained from traditional MTF calculations ($aSNR_{HS}$):

$$aSNR_L = 10\log_{10}\left(\frac{A\mu_x}{B}\right) = 10\log_{10}\left(\frac{r^2}{1-r^2}\right), \quad (1)$$

where the output envelope, $y[n]$, was modeled as $y[n] = Ax[n] + B$, $\mu_x = E[x[n]]$ and $x[n]$ was the input envelope. The term $r^2 = \left(\frac{\mu_x}{\mu_y}\right)\left(\frac{\lambda_{xy}}{\lambda_x}\right)$ is the correlation coefficient between the input (clean) and output (degraded) intensity envelopes. The term μ_y is the mean of the output envelope, λ_{xy} is the cross-covariance of the input and output envelopes, and λ_x is the auto-covariance of the input envelope. Goldsworthy noted that the equivalent form of $aSNR_L$ was similar to the $aSNR$ computed by Houtgast and Steeneken, e.g.

$$aSNR_{HS} = \frac{1}{N}\sum_{f_{\mathrm{mod}}} aSNR_{HS}(f_{\mathrm{mod}}) = \frac{1}{N}\sum_{f_{\mathrm{mod}}} 10\log_{10}\left(\frac{MTF(f_{\mathrm{mod}})}{1-MTF(f_{\mathrm{mod}})}\right), \tag{2}$$

where N is the number of modulation frequencies averaged. The differences are that the latter is computed as a function of modulation frequency, the log is then taken and summing across modulation frequencies occurs after clipping the logged data to restrict the range to be from –15dB to +15dB (*MTF* from .031 to .969).

To investigate the relationship between the two techniques, Goldsworthy went on to expand the correlation coefficient, r^2, as follows:

$$r^2 = \frac{\mu_x}{\mu_y}\frac{\lambda_{xy}}{\lambda_x} = \frac{\mu_x}{\mu_y}\frac{\sum_{f_{\mathrm{mod}}} S_{xy}(f_{\mathrm{mod}})}{\sum_{f_{\mathrm{mod}}} S_{xx}(f_{\mathrm{mod}})}, \tag{3}$$

where the normalization terms in the numerator and denominator have been cancelled. The terms $S_{xy}(f_{\mathrm{mod}})$ and $S_{xx}(f_{\mathrm{mod}})$ correspond to the cross-power spectrum between input and output envelope and the power spectrum of the input envelope respectively. The numerator can be further expanded to give:

$$r^2 = \frac{\mu_x}{\mu_y}\sum_{f_{\mathrm{mod}}}\frac{S_{xx}(f_{\mathrm{mod}})}{\sum_{f_{\mathrm{mod}}} S_{xx}(f_{\mathrm{mod}})}\cdot\frac{S_{xy}(f_{\mathrm{mod}})}{S_{xx}(f_{\mathrm{mod}})}. \tag{4}$$

The ratio of μ_x to μ_y accounts for normalization of DC level of the envelopes. The first ratio inside the summation corresponds to a weighting factor that weights the MTF more heavily at modulation frequencies where the speech envelope spectral density is high (rather than equal weighting as the MTF does). The last ratio reduces to the $MTF(f_{\mathrm{mod}})$ for the case of deterministic signals. The first two ratios can be combined into a single scaling factor that is a function of f_{mod} to give:

$$r^2 = \sum_{f_{\mathrm{mod}}} w(f_{\mathrm{mod}})\frac{S_{xy}(f_{\mathrm{mod}})}{S_{xx}(f_{\mathrm{mod}})}. \tag{5}$$

This result is reminiscent of Drullman's result although the summation in this representation of r^2 is inside the log function. The ratio $\frac{S_{xy}(f_{\mathrm{mod}})}{S_{xx}(f_{\mathrm{mod}})}$ corresponds to an estimate of the complex transfer function of the system.

11.3. CURRENT INVESTIGATIONS

We are currently investigating various representations of the sMTF that are derived from the ratio of the cross power spectrum to the input power spectrum of the speech envelopes. This paper reports on those investigations for speech in speech-shaped noise, multi-talker interference, reverberation and amplitude compression plus the other degradations. Speech-shaped noise was generated by digitally filtering white Gaussian noise to roughly match the average broadband spectrum of clear and conversational speech. In

some octave bands, the noise level was slightly higher than the speech level and in others it was slightly less. For the multi-talker interference, restaurant babble was used (multi-talker babble with dish and silverware clanking in the background). For reverberation, a room was simulated to be moderately reflective with a reverberation time of 0.6 s and with the listener positioned at the critical distance from the talker.

11.3.1. sMTF Computational Considerations

Consider first the issue that $S_{xy}(f_{\text{mod}})$ is potentially complex whereas $S_{xx}(f_{\text{mod}})$ is real-valued. When summed over all modulation frequencies (both positive and negative) the imaginary terms in $S_{xy}(f_{\text{mod}})$ should cancel since $R_{xy}(\tau)$, the cross-correlation function, is real. On the other hand, transfer functions are usually represented in terms of their magnitude and phase rather than their real and imaginary parts. Arguing from a different perspective, Drullman et al. considered only the real part of $S_{xy}(f_{\text{mod}})$ because they felt the important part of the cross-power spectrum was that which was in phase with the input.. We consider both the real part and the magnitude of the cross-power spectrum as potential sMTF contributors.

A second computational issue is the location of the summation(s). The original method to compute an apparent SNR (aSNR) calls for summing across a range of $^1/_3$ octaves before taking the ratio and then summing the log of a transform of the ratios. Goldsworthy's evaluation of Ludvigsen's method suggests either summing across all frequencies before taking the ratio (Eqn 3) or taking a weighted sum of all ratios (Eqn 5). We will avoid this issue and consider only the different ways one might estimate the sMTF, upon which the aSNR is based. To remain consistent with the original perspective, we will focus on methods that sum across a range of $^1/_3$ octaves before taking a ratio. In other words, we will compute the following:

$$sMTF(f_{\text{mod}}) = \frac{\sum_{1/3\,\text{octave centered at } f_{\text{mod}}} \text{numerator}(f)}{\sum_{1/3\,\text{octave centered at } f_{\text{mod}}} \text{denominator}(f)}, \tag{6}$$

where $sMTF(f_{\text{mod}})$ corresponds to different estimates of the modulation transfer function at $^1/_3$ octave frequencies between .25 and 25 Hz. The terms 'numerator(f)' and 'denominator(f)' represent the various parameters we will examine as contributors to the $sMTF(f_{\text{mod}})$. For comparison purposes, we will also compare our new estimates with what would be obtained using the original Houtgast and Steeneken technique and with the theoretical prediction based on room acoustics (if it exists).

11.3.2. sMTFs of Speech in Speech-Shaped Noise

When speech is presented in speech-shaped noise, the signal to noise ratio in each octave band is the same and the theoretical MTF for an octave band is given in Houtgast and Steeneken (1980) by:

$$MTF = \frac{1}{1+10^{-SNR(\text{in dB})/10}}. \tag{7}$$

Note that this MTF is not a function of modulation frequency, i.e. interfering noise affects all modulation frequencies equally. For the purposes of demonstration, we consider an overall SNR of 0dB. Fig. 6 compares the theoretical MTF (dotted line) with three speech-based MTF algorithms in three octave bands: 250 Hz band, 1000 Hz band and 4000 Hz band. The solid curves were computed using the algorithm of Houtgast and Steeneken, i.e. $\frac{Y(f_{\text{mod}})}{X(f_{\text{mod}})}$. The dashed curves were computed using Drullman's phase-locked MTF algorithm, i.e. the real part of the cross-power spectrum divided by the input power spectrum. The dot-dashed curves were obtained from the magnitude of the cross-power spectrum divided by the input power spectrum. At most, only three curves can be seen because the phased-locked sMTF and the magnitude sMTF techniques give identical results for this type of degradation.

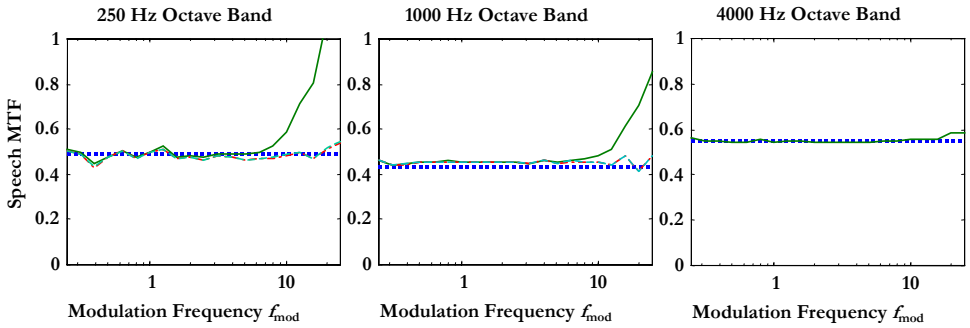


Figure 6. Speech-based Modulation Transfer Functions (sMTFs) for speech in noise using three different techniques in three octave bands. The dotted line depicts the theoretical prediction based on SNR in that octave band. The solid line plots application of the Houtgast and Steeneken method to speech, the dashed line plots Drullman's phase-locked MTF and the dot-dashed line plots our new approach using the magnitude of the cross power spectrum.

All speech-based methods result in sMTFs very close to theoretical predictions at low modulation frequencies. Drullman's method and the magnitude method closely match the theoretical predictions across all modulation frequencies, even in the lowest octave bands. The Houtgast and Steeneken method deviates from the theoretical values in the lowest octave bands due to the previously-mentioned noise-floor problem.

11.3.3. sMTFs of Speech in Noise plus Reverberation

For the condition of speech in noise plus reverberation, the speech-in-noise materials were convolved with the simulated room impulse response. To obtain the theoretical MTF for this condition, the room impulse response was bandpass filtered into octave bands, squared and the FFT taken for the signal in each band. The FFT was normalized by the DC term and values were averaged over $^1/_3$ octaves. The function was then scaled by the attenuation factor for noise alone, computed in the previous section (Eqn. 7). These theoretical functions are plotted in Fig. 7 as the dotted curves for the three octave bands. The prediction in each band is a low-pass filter shape with a low-frequency asymptote at the MTF for speech in noise. The three techniques to compute the sMTF are plotted using the same line types as in the previous figure.

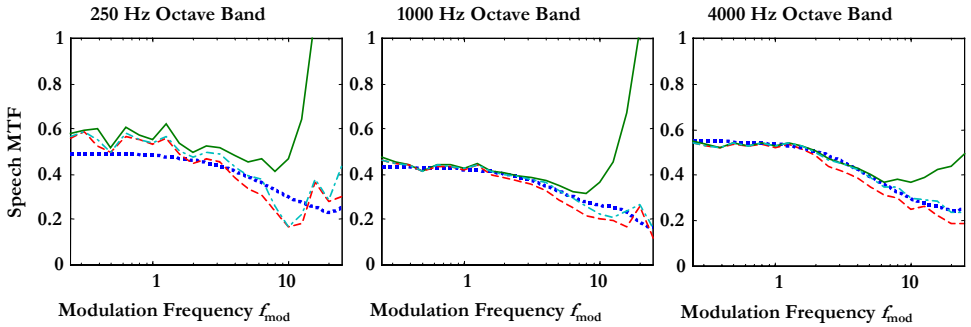


Figure 7. Speech-based MTFs for speech in noise plus reverberation using three different techniques, shown for three octave bands. The dotted line depicts the theoretical prediction based on SNR in that octave band times the MTF based on the room impulse response. The solid line plots application of the Houtgast and Steeneken method to speech, the dashed line plots Drullman's phase-locked MTF and the dot-dashed line plots our new approach using the magnitude of the cross power spectrum.

A few features in these plots are worth noting. First, all three sMTF techniques follow the general trend of the theoretical predictions up to a point. The sMTF using Houtgast and Steeneken's clearly deviates from the theoretical prediction starting around 8 Hz, even in the highest octave band. Also, all sMTF techniques slightly over-predict the MTF at low modulation frequencies in the lowest octave band. This is due to the sMTF in reverberation being slightly greater than 1.0 at low modulation frequencies in this band. The effect is slight and does not affect STI computations. In addition, Drullman's and our techniques slightly underestimate the theoretical MTF at higher modulation frequencies (Drullman's more so than ours) except for peaks at 16 Hz in the 250 Hz octave band and at 20 Hz in the 1000 Hz octave band.

11.3.4. sMTFs of Speech in Restaurant Babble

We also wished to determine the robustness of the various techniques in the presence of fluctuating background noise. Houtgast and Steeneken (1985) cited this as a limitation of their method using speech as a probe stimulus as did Ludvigsen et al. (1990) who came up with their own, correlation-based, approach. Fig. 8 plots the sMTFs for the three techniques on speech in 0dB, restaurant babble, noise. The dotted lines indicate the SNR in each octave band.

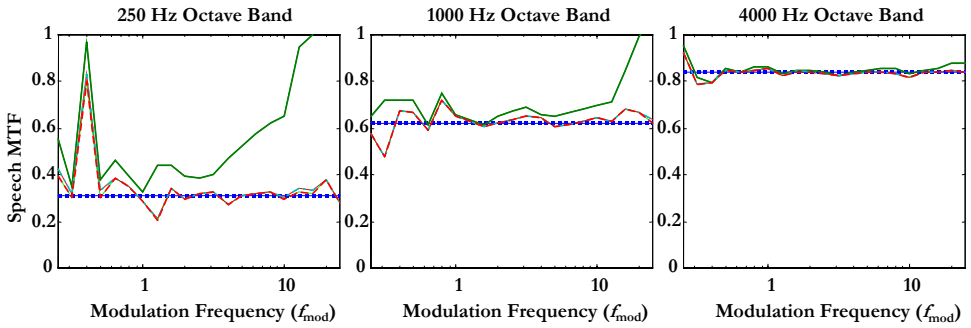


Figure 8. Speech-based MTF plots for speech in 0dB, restaurant babble, noise in three octave bands. The same line types are used as in the previous figure.

As expected, applying Houtgast and Steeneken's technique using speech as a probe stimulus overestimates the MTF, particularly in the lower octave bands. Both Drullman's technique and ours track the theoretical MTF quite well except at 0.4 Hz in the 250 Hz octave band where all three techniques show a large spike in the sMTF. This frequency happens to be very near the sentence rate for our speech materials, so the spike may indicate a robust representation of pauses by the talker in the presence of fluctuating background noise. While we have not yet analyzed the sMTF in other time-varying backgrounds, we are optimistic that both Drullman's technique and ours will accurately predict the MTF is such environments.

11.4. HEARING-AID PROCESSED SPEECH

We next examined how envelope spectra were affected by amplitude compression hearing aids for two different hearing loss profiles: A 50-dB flat loss and a moderate sloping loss. We treated listener elevated thresholds as internal noise and scaled the sMTFs accordingly.

Linear amplification and amplitude compression hearing aids were evaluated for both types of loss although we will only show the sMTFs for compressed conditions. All aids used the NAL-R frequency-gain prescription. The amplitude compression aids had attack times of 5 ms and release times of 200 ms. For the flat-loss profile, we examined compression ratios of 2 and 3. For the sloping loss we used a compression ratio of 1.5 in the low frequency bands (< 2.5 kHz) and 2.5 in the high frequency bands (> 2.5 kHz).

11.4.1. sMTFs of Amplitude-Compressed Speech in Restaurant Babble for Flat-Loss Profile

Fig. 9 depicts sMTFs for speech in restaurant babble at 5dB SNR through amplitude compression (CR=3) and the 50dB flat loss profile for the three processing algorithms. Note there is no theoretical prediction for this condition due to the presence of amplitude compression. All three sMTFs are relatively flat, with the Houtgast and Steeneken methods rising at high modulation frequencies and the curves based on Drullman's phase-locked and our magnitude methods falling at high modulation frequencies. Note that the phase-locked sMTF falls at or below the magnitude sMTF at each modulation frequency.

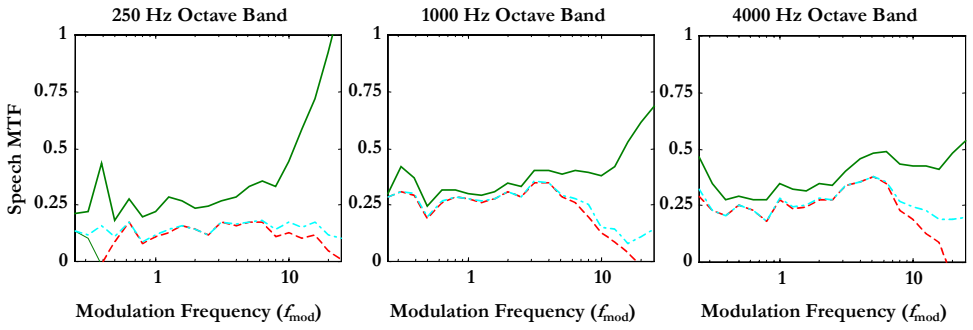


Figure 9. Speech-based MTFs for speech in the presence of restaurant babble (at 5dB SNR), as processed by an amplitude-compression hearing aid (CR=3) and a 50dB flat hearing loss. The solid lines plot the Houtgast and Steeneken method, the dashed lines plot Drullman's phase-locked method and the dot-dashed lines plot our new approach.

11.4.2. sMTFs of Amplitude-Compressed Speech in Restaurant Babble for Sloping-Loss Profile

In Fig. 10, we show speech-based MTFs for the same stimuli, only now they have been processed by an amplitude compression hearing aid designed for someone with a sloping loss and they include the corresponding sloping-loss thresholds. The biggest difference in the sMTFs for this condition is that the phase-locked sMTF of Drullman actually goes negative at 10 Hz. It appears that the output is exactly out of phase with the input at this modulation frequency. At frequencies above and below 10 Hz, the phase locked and magnitude sMTFs are very similar. The curves based on Houtgast and Steeneken's technique rise as they did for the flat-loss case.

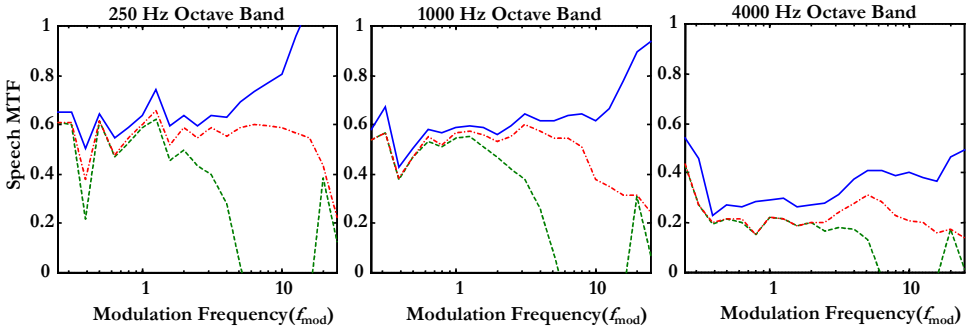


Figure 10. Speech-based MTF plots for speech in restaurant babble (5dB SNR), amplitude-compression hearing aid and sloping hearing-loss profile. The line types are the same as in the previous figure.

11.5. COMPARISONS OF SPEECH-BASED STI CALCULATIONS WITH INTELLIGIBILITY DATA

We next looked at how well the three sMTF techniques could predict listener intelligibility for hearing-aid processed speech. Two experiments were run. In the first, six normal-hearing subjects listened to nonsense sentences processed by linear and compression hearing aids and by a simulation of a 50-dB flat hearing loss that included recruitment (Duchnowski & Zurek, 1995). Two compression hearing aids were considered. One had a compression ratio of 2 and the other had a compression ratio of 3. Both the linear and compression aids were fit to the flat-loss profile using NAL-R frequency-gain characteristics. In the second experiment, three normal-hearing subjects listened to the same sentences processed through a linear and a compression hearing aid fit to a sloping-loss impairment, also using an NAL-R prescriptive fit which were then processed by a sloping-loss simulation.

Fig. 11 plots three STI vs intelligibility graphs, each uses a different sMTF technique to compute Modulation Transfer Indices (MTIs) in each octave band. The IEC standard octave band weights and redundancy correction factors for a male talker (α and β respectively) were then applied to obtain the respective STIs. The data look similar for all three techniques. In all three cases, the linear processing conditions (circles) have higher STI values than the compression conditions (stars, triangles and pluses) although the subject scores are comparable. The STI using Drullman's phase-locked sMTF appears to have the tightest clustering of the amplitude compression results but the data set is rather small to draw a firm conclusion.

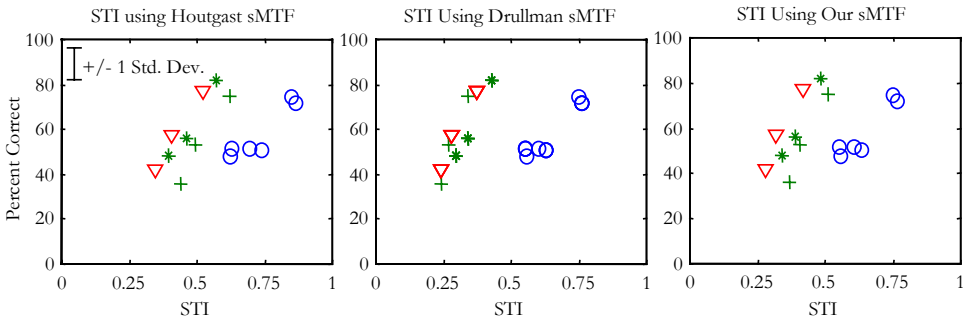


Figure 11. Speech-based STI vs intelligibility using IEC standard α and β weights and each of the three techniques considered in the previous section. The circles indicate linear processing conditions. The stars show CR=2 conditions, the inverted triangles plot CR=3 conditions and the pluses indicate sloping-loss compression conditions. The vertical bar in the first frame indicates +/- 1 standard deviation in the subject data.

The STI values were also computed based on weighting the sMTFs by the relative strength of the input power spectrum (as suggested in Eqn 4). While in some cases the resulting STI increased slightly and in others the STI decreased slightly, the overall impact appears to be negligible for the conditions considered. The linear processing conditions still have STI values much higher than comparable compression conditions for which subjects had similar intelligibility scores. Within a class of processing though (linear or amplitude compression) there appears to be a reasonable consistency between STI predictions and subject intelligibility scores.

11.6. DISCUSSION

In this paper we have considered three different techniques to compute a speech-based Modulation Transfer Function (sMTF). The one based on Houtgast and Steeneken's original algorithm shows well-documented deviations from theoretical predictions in every condition tested. The techniques based on cross power spectra (Drullman's phase-locked sMTF and our magnitude sMTF) appear to duplicate the theoretical predictions in every case where a prediction exists, including speech in the presence of fluctuating background noise. When speech is processed by an amplitude-compression hearing aid for a sloping-loss profile, a larger distinction between the techniques exists, with the phase-locked sMTF going negative at some modulation frequencies.

When the STI values predicted by the three techniques are compared to intelligibility data, Drullman's technique performs best, but the differences are slight. The better fit is based on the relatively low intelligibility scores subjects obtained in the sloping loss, amplitude compressed conditions for which Drullman's sMTF curves go negative and deviate the most from the other two techniques.

Overall, we find the results encouraging, even though there is not a unique clustering that relates a speech-based STI to intelligibility across all processing conditions. The linearly processed data and the amplitude compression data cluster well if they are analyzed separately. This may mean that a speech-based STI can be used to compare relative merits of parameters within a processing category, such as amplitude compressed speech.

We will continue this avenue of investigation, looking at a wider range of processing conditions, including more sentence lists in the envelope spectrum computations, and more subject experiments to provide intelligibility data.

ACKNOWLEDGEMENT

This work was supported by NIDCD.

REFERENCES

Drullman, R., Festen, J. M. and Plomp, R. (**1994**), "Effect of reducing slow temporal modulations on speech reception", *J. Acoust. Soc. Am.,* **95,** 2670-80.

Duchnowski, P. and Zurek, P. M. (**1995**), "Villchur revisited: another look at automatic gain control simulation of recruiting hearing loss", *J. Acoust. Soc. Am.,* **98,** 3170-81.

Goldsworthy, R. (**2001**), "Relationship of the MTF to the correlation-based STI calculation", Personal communication.

Hohmann, V. and Kollmeier, B. (**1995**), "The effect of multichannel dynamic compression on speech intelligibility", *J. Acoust. Soc. Am.,* **97,** 1191-5.

Holube, I. and Kollmeier, B. (**1996**), "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model", *J. Acoust. Soc. Am.,* **100,** 1703-16.

Houtgast, T. and Steeneken, H. J. M. (**1972**), "Envelope spectrum and intelligibility of speech in enclosures", In *Proceed. Conf. Speech Commun. Proc. IEEE-AFCRL*, 392-395.

Houtgast, T. and Steeneken, H. J. M. (**1973**), "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acustica,* **28,** 66-73.

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

Houtgast, T. and Steeneken, H. J. M. (**1985**), "A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *J. Acoust. Soc. Am.,* **77,** 1069-1077.

Humes, L. E., Dirks, D. D., Bell, T. S., Alhstrom, C. and Kincaid, G. E. (1986), "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners", *J. Speech Hear. Res.,* **29,** 447-462.

IEC (**1998**), *Sound System Equipment - Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index; 2nd Edition*, International Standard No. 60268-16.

Krause, J. C. (**2001**), *Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement*, Ph.D., Dept. Elect. Eng. Comp. Sci., M.I.T., Cambridge, MA, USA.

Ludvigsen, C. (**1993**), "The use of objective methods to predict the intelligibility of hearing aid processing speech", In *Recent Developments in Hearing Instrument Technology*, (Eds, Beilin, J. and Jensen, G. R.) Scanticon, 81-94.

Ludvigsen, C., Elberling, C. and Keidser, G. (**1993**), "Evaluation of a noise reduction method - Comparison between observed scores and scores predicted from STI", *Scand. Audiol. Suppl.,* **38,** 50-55.

Ludvigsen, C., Elberling, C., Keidser, G. and Poulsen, T. (**1990**), "Prediction of intelligibility of non-linearly processed speech", *Acta Otolaryngol Suppl,* **469,** 190-5.

Lundin, F. J. (**1982**), "The influence of room reverberation on speech - an acoustical study of speech in a room", *Speech Trans. Lab. - QPR,* **1982,** 24-59.

Payton, K. L. and Braida, L. D. (**1999**), "A method to determine the speech transmission index from speech waveforms", *J Acoust Soc Am,* **106,** 3637-48.

Payton, K. L., Uchanski, R. M. and Braida, L. D. (**1993**), "Computation of modulation spectra for the speech transmission index using real speech", In *Proceed. Appl. Sig. Proc. Aud. Acoust.*, 110-113.

Payton, K. L., Uchanski, R. M. and Braida, L. D. (**1994**), "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing", *J. Acoust. Soc. Am.,* **95,** 1581-92.

Picheny, M. A., Durlach, N. I. and Braida, L. D. (**1985**), "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech", *J Speech Hear Res,* **28,** 96-103.

Plomp, R. (**1988**), "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function", *J. Acoust. Soc. Am.,* **83,** 2322-7.

Steeneken, H. J. M. and Houtgast, T. (**1980**), "A physical method for measuring speech-transmission quality", *J. Acoust. Soc. Am.,* **67,** 318-26.

Steeneken, H. J. M. and Houtgast, T. (**1983**), "The temporal envelope spectrum of speech and its significance in room acoustics", In *Proceed. Eleventh Int. Cong. Acoust.*, 85-88.

Villchur, E. (**1989**), "Comments on "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function"", *J. Acoust. Soc. Am.,* **86,** 425-428.

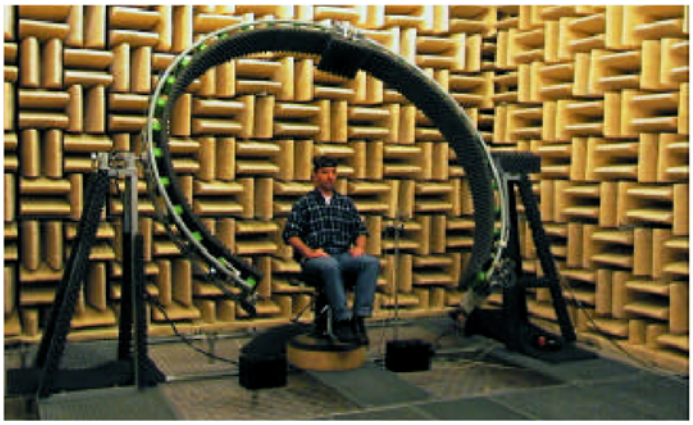
INDEX

Human Factors

TNO Human Factors is focussed at human behavior and performance in a technical environment. Through innovative research we improve performance, safety and comfort. We work for the Netherlands' Armed Forces and, world wide for private enterprises and governments.

Our primary activities include:

- *Perception:* Vision, Speech and Hearing.
- *Information Processing:* Usability Engineering, Cognition, Information Transfer.
- *Skilled Behavior:* Steering and Control, Traffic Behavior.
- *Work Environment:* Workplace Ergonomics, Thermal Physiology, Equilibrium and Orientation, Aerospace Medicine.
- *Training and Instruction:* Learning Processes, Teamtraining, Simulation and Modeling.
- *Group Work:* Distributed Decision Making, Psychosocial Interactions.



The Anechoic room, one of TNO Human Factors' research facilities with setup for sound localization experiments.

TNO Human Factors is the birthplace of the Speech Transmission Index STI

Website: www.tm.tno.nl

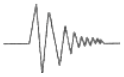
INTELLIGIBLITY CAN BE QUICKLY AND EASILY MEASURED WITH ACCURACY

Ideal for Voice Evacuation and Fire Alarm Systems, Paging for Arenas, Stadiums & Halls

- Conforms to International Standards
- Test Signal Developed by TNO in the Netherlands
- Analyzer Designed and Manufactured by Gold Line in Conjunction with Bose
- Gold Line Easy to Carry Kit Includes RTA, Microphone, Pink Noise Generator with Sine Wave and the New OPT STICis™ with the STI-PA™ Test Tone

GOLD LINE
MANUFACTURED IN U.S.A. SINCE 1961

e-mail: sales@gold-line.com
Website: www.gold-line.com



Remember 1977?

Dr. Ir. T. Houtgast studeerde aan de TH-Delft, Afdeling Technische Natuurkunde, en werkt sinds 1964 bij het Instituut voor Zintuigfysiologie TNO in Soesterberg. Zijn werkterrein omvat uiteenlopende onderwerpen op het gebied van auditieve perceptie

---

Ing. H.J.M Steeneken studeerde aan de H.T.S., studierichting elektronica, en trad in 1965 in dienst van TNO bij het Instituut voor Zintuigfysiologie. Hij houdt zich onder meer bezig met het beoordelen en evalueren van spraaktransmissiekanalen zoals radioverbindingen, digitale verbindingen, telefoons en microfoons. Ook zaalakoestische problemen op dit gebied behoren tot zijn werkterrein.