SPEECH INTELLIGIBILITY MEASUREMENT – THE CURRENT STATE OF THE ART

Peter Mapp Peter Mapp Associates, Colchester Essex, UK

1 INTRODUCTION

The need for an accurate and portable machine based method of measuring the potential Intelligibility of a classroom, auditorium or sound system has long been recognised. Furthermore with the ever increasing use of sound systems for Voice Alarm and emergency purposes, together with the heightened interest in classroom acoustics and so called 'Soundfield' systems, the need for a simple but reliable method of intelligibility verification has never been greater. A number of methods are potentially available for assessing intelligibility, these include the Articulation Index (AI), Percentage Loss of Consonants (% Alcons), D/R ratios (including C50 & C35) Word Scores and STI and it's derivatives RaSTI & STIPa. Whilst word scores are fundamentally the most accurate, they are cumbersome and expensive to conduct and in many situations completely impractical. Furthermore, they are not as straightforward to conduct as many appear to think. Considerable effort has therefore been made over the years into developing machine or computer based or indirect intelligibility assessment techniques:

2 SPEECH INTELLIGIBILITY MEASURES

2.1 Articulation Index

The Articulation Index was the earliest of the instrumentation based approaches, being developed in the 1960s, though its origins date back well before this. It was primarily derived for testing single channel communication systems. Whilst dealing well with noise it is not able to handle speech degradation due to reverberation or poor direct to reverberant ratios, although some (inaccurate) corrections have been proposed. [1] The method is based on measuring the background noise and wanted speech signal either in terms of octave or 1/3 octave bands. The resultant signal to noise ratios are then weighted according to their intelligibility contribution and then combined to provide a single number index. Whilst the method has generally fallen out of use, it is still very useful for determining the potential speech privacy between offices or within open plan areas. In this case the Privacy Index (PI = 1- AI) is generally used. When studying open plan and low height partition systems, it is essential that the sound source correctly mimics the directivity and frequency response of the human talker. As has been shown in an earlier paper to this conference [2], failure to do so can lead to significant error. (see also [3]).

2.2 % ALcons

The % Alcons technique,, derived from a simple prediction formula for classroom intelligibility, is essentially a Direct to Reverberant measure. It was conceived in 1986 for the TEF Time Domain Spectrometry analyser. [4] The initial version required considerable operator skill in both setting up the measurement and selecting the analysis cursor / delineation time windows. Later versions incorporated software that automatically controlled the initial settings. However, the overall measurement still requires skill in interpretation and is open to potential operator adjustment error. The technique's main shortfall however is that it is based solely on a measurement made over a narrow frequency band centred on 2kHz. The method therefore cannot deal with sound systems or acoustic conditions that vary appreciably at other frequencies, which in practice most Voice Alarm and Soundfield systems do ! [5,6]

2.3 C50 & C35 Early to Late Sound Ratios

Whilst the C50 and to a lesser extent C35 scales are well recognised in the field of Auditorium Acoustics as being useful indicators of potential speech intelligibility, no formalised scale has been developed (Although, as a rule, a C50 value of at least 0-2 dB is required for good intelligibility – though in the author's experience this is reverberation time dependent). Traditionally the measurements are made purely within the 1 kHz octave band. The methods do not take background noise nor noise masking effects into account and find limited application with respect to sound system assessment. [5]. It is interesting to note that whilst acousticians working in the field of auditorium acoustics accept 50 mS as being a suitable delineator between useful and detrimental sound, the audio fraternity suggest that this limit should be shortened to around 20 mS or at most 35 mS.

2.4 U50 & U80 Useful to Detrimental Ratios

This concept was suggested by Bradley [7] and combines both the Direct (early) to Reverberant (late) sound energy ratios with background noise.

$$U_{50} = 10 \log [D/ D\text{-}1 + n/s]$$

Where D is the early energy fraction and
n/s is the signal to noise ratio in energy terms

Interestingly, Bradley found the best correlations with measured speech intelligibility scores when he used an integration time of 80 mS for the useful sound component, which contrasts with the C50 and C35 measures noted above.

2.5 STI

The STI scale and technique was initially developed during the early 1970s by Houtgast & Steeneken [8]. The method not only uses a speech like stimulus** but also assesses the system or room over the complete speech band by making octave band measurements over the range 125 Hz – 8 kHz. The technique replicates the natural low frequency modulations in speech from 0.63 to 12.5 Hz and measures the reduction in these modulations in each of the seven carrier bands of 125 to 8 kHz. This produces a modulation transfer function matrix of some 98 measurement points. Steeneken & Houtgast found that the reduction of the modulation depth, which can be caused by either

or noise or reverberation or a combination of the two) correlated well with perceived speech intelligibility and word scores. However, in the 1970s and 1980s the processing power required to measure the mtfs and compute the complete modulation matrix was severely restricted. Therefore, whilst the technique was shown to be acoustically viable, its practical implementation was not achievable at the time.

{ ** Note, that whilst the original STI concept employed a modulated speech like signal, following Shroeder's analysis showing that the mtf could be derived from the impulse response, later implementations of STI measurement systems adopted this latter approach. Therefore, depending on the stimulus used to generate or compute the impulse response (eg ML sequences, sine sweeps/chirps or direct impulses) very different signal formats with a diverse range of characteristics and crest factors are currently employed. The way in which each of these signals may potentially interact with a sound system can be quite different and none directly replicate the modulation signal concept of STI }.

2.6 RaSTI

In 1985 a foreshortened adaptation of STI was developed by Houtgast & Steeneken and termed RaSTI (Rapid STI). This reduced the modulation matrix down from 98 to just 9 points and a corresponding order of magnitude reduction in the computational power required. A handheld measurement instrument was also introduced by B&K in that year and for the first time, a portable (indirect) intelligibility measurement system was available. The introduction of the instrument and the ability to readily measure the apparent potential intelligibility of a sound system, soon led to the adoption of Rasti by a number of international standards, Codes of Practice and the CAA for certification of all aircraft PA systems. The RaSTI technique is restricted to just the 500 Hz and 2 kHz octave bands with 4 and 5 modulation frequencies respectively, thereby cutting the mtf matrix down to 9 points. Whereas this reduction of measurement and data analysis enabled a practical measurement system to be achieved, the implications for sound system measurement and assessment were not realised until several years later.

3 HOW ACCURATE IS RASTI ?

Although the introduction of the unifying CIS scale (see later) was intended to provide a choice of measurement techniques and criteria, in the UK at least, RaSTI is still very much the dominant descriptor. (Though the recent adoption by the USA of CIS may influence this). By limiting the frequency bands to just 500 Hz and 2 kHz does not allow a full audit of a sound system to be made. Whereas with natural voice transmission this does not lead to significant error, this is not the case with respect to sound systems, where the response may be far from linear.

Although a wealth of anecdotal evidence suggested that there could be wide discrepancies between STI & RaSTI when measuring sound and VA system performance, no formal study had ever been undertaken or published. Mapp in 2002 [9] however published the results of just such a study. The STI & RaSTI performance values for 81 sound systems were studied and compared. Figure 1 shows a sample of the data. The figure shows a plot of the difference or error between RaSTI and STI for the condition of reverberation only intelligibility degradation (ie noise is not a contributory factor). Whereas the mean error is 0.08, individual cases can generate potential errors well in excess of this, typically ranging from 0.05 to 0.1. The cases investigated cover a wide range of systems and acoustic environments and interestingly show that RaSTI can both underestimate as well as over

estimate the full STI value. Furthermore, examination of the data in more detail shows there to be no obvious trend or condition which causes RaSTI to individually under or over estimate the result. (For further discussion and analysis see reference [9]).

The situation where noise is a contributory factor (either in conjunction with or without reverberation) is rather more clear-cut with RaSTI almost always over estimating the result. (See Mapp [10 & 13] for more details and discussion of other factors). An example of how Rasti over estimates a situation is shown in the example given below. Here three different loudspeakers are compared with and without noise as the intelligibility degradation factor. As can be seen from the scores, not only does background noise play an important part in determining the overall STI or intelligibility but the order of merit of loudspeaker products can change. In the particular case presented, the measurements were made in the passenger cabin of an aircraft, where a criterion of 0.6 Rasti has to be met. The mean results can be summarised as follows :

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

Whereas there is little difference between the STI & RaSTI values under high signal to noise conditions, when the background noise was present, it can seen that the RaSTI values are generally significantly higher. (eg 0.76 Rasti as opposed to 0.60 STI for type A and 0.71 Rasti as opposed to 0.57 STI for type C). These are markedly different results and in the case of loudspeaker type C, shows that RaSTI is producing a sense of false security in estimating 0.71 STI – well within the 0.6 mandatory criterion as opposed to the actual value of 0.57 STI – a fail condition ! (Further details can be found in ref [11]).

There is ever mounting evidence that RaSTI results must be treated with extreme caution and that possibly the scale should be abandoned all together for verification of PA & VA system performance. Indeed, even Steeneken himself does not now support its use for sound system performance measurement ! [12] Indeed, he has now re-named RaSTI, Room Acoustic Speech Transmission Index !

4 CIS & STIPA

The problem is what to replace RaSTI with. The obvious choice is STI, which although with modern computer processing power is no longer a tortuous experience, the need for additional speech spectrum filters still makes the method cumbersome and open to manipulation. Furthermore, the different types of excitation signal can (and do) give rise to different results. [13] The USA has solved the problem by recently adopting CIS as the criterion. This enables a number of techniques to be adopted – though the question of calibration and verification of the verification measurement system is raised. A possible

way round this is the adoption of STIPa. This is similar to Rasti but uses a sparse matrix and encompasses the complete, seven octave band range from 125 Hz to 8 kHz. The stimulus is spectrally shaped, modulated pseudorandom noise. The method thus fulfils the original STI concept, although a reduced mtf matrix is employed. The system is completely portable and can be produced such as to give the user a very simple interface. Measurements need only take 12-15 seconds per location. Steeneken at TNO played a major part in its development. Interestingly, in the USA, some measurement meters will be set to read out CIS rather than STI but STIPa is the underlying measure.

Being based on a Pseudo-random signal, STIPa readings can vary. The following table shows a number of typical data sets

Table 1 STIPa measurement variations

Location	Run 1	Run 2	Run 3	Run 4	Run 5	Mean
1	0.54	0.54	**0.50**	0.55	0.54	**0.54**
2	0.71	0.68	0.68	0.69	- -	**0.69**
3	0.63	0.61	0.60	0.61	0.58(NOISE)	**0.61**
4	0.50	0.53	**0.30 ****	0.53	0.55	**0.53**
5	0.61	0.63	0.62	0.61	0.61	**0.62**

As the tables shows, STIPa measurements typically vary by about 0.02-0.03 STI, but occasional discrepancies also occur,(shown in bold type) which means that several readings should be taken in order to ensure an accurate measurement is made.

Although the STIPa modulated signal is very complex, it does not require the signal transmitter and test receiver / analyser to be synchronised. (Unlike MLS signals for example). It can therefore be recorded and played back, either through the system under test, or via a head and torso simulator etc. One method of doing this is to record the signal onto a CD and then play this back. However, it would appear that all CD players are not created equal and significant errors can result. Reproduction systems incorporating data compression such as mini disc or MPEG, can also result in erroneous readings being obtained.

A major failing of STIPa and STI as currently being implemented is that they are not able to correctly take account of irregular or poor sound system frequency responses. This problem is particularly noticeable when dealing with sound systems operating in reverberant but high signal to noise conditions. This effect has been reported elsewhere by the author and is currently under intense investigation.[10,13]

5 SPEECH INTELLIGIBILITY INDEX SII

The speech intelligibility Index (Ansi 3.05 1998) does not provide qualification intervals (eg as per the 5 point STI scale) but instead has just two benchmarks > 0.75 Good and < 0.45 Poor. It is primarily intended for the evaluation of speech communication channels (eg Radio communication) and whilst taking account of Noise, Bandwidth and speech spectrum , it does not cater for reverberation or other temporal distortions

6 STI MEASUREMENT TECHNIQUES

6.1 Occupancy

Ideally, intelligibility measurements should be made during normal operating conditions, i.e. typical occupancy and occupancy noise. However, in practice this is not often possible / acceptable. For this reason, tests are normally conducted when the space is empty or outside normal working hours. This can often mean that the acoustic environment during the tests is not representative of normal conditions. For example, the background noise may be lower but conversely, the reverberation time may be higher. A typical example of this problem would be the situation found when testing the PA system at a football stadium. Unoccupied, the noise levels may be anything from 20 to 60 dBA quieter than when occupied. Conversely however, the reverberation time may decrease by as much as 50%.when occupied. Auditoria and classrooms are other typical cases in point, with reverberation time changes of 20-50 % being typical and variations greater than this not unknown. Clearly such acoustic changes need to be accounted for in an overall assessment. Furthermore, stationary noise (eg from air conditioning or continuous process noise) can usually be readily accounted for if the intelligibility testing is carried out in a low noise condition. However, variable noise such as from spectators, children in class or certain industrial processes and transportation noise is a different matter and although a statistical approach can be taken (eg by use of LA10), it is not always clear how the noise levels really relate to intelligibility masking.

6.2 Acoustic & Direct Injection of Test Signals

When testing Sound Systems it is often essential to include the system microphone within the measurement chain. However, where it can be shown that the microphone's response and acoustic environment will have little or no impact on the measured result, then it may be permissible to directly inject the electronic test signal into the sound system under test.

Where a system microphone is to be included within the measurement set up, it is important that a mouth simulator with a directivity corresponding to that of the human voice is employed The simulator should be positioned at the normal talker distance and orientation with respect to the microphone and the level adjusted such that the A-weighted sound pressure level at the capsule is set to be equivalent to either the normal user or to a typical reference level. Eg 66 dBA at 0.5 m or 86 dBA at 50 mm. The frequency response of the simulator in conjunction with the STI test signal must be set to replicate the average long-term spectral response of human speech.

Where the intelligibility due to the natural acoustics of a space eg classroom or auditorium are to be evaluated, then the test source must have the same directional characteristics as the human head & mouth. This is not the same as a typical stand alone mouth simulator. Small head sized loudspeakers are often used for this purpose, but it has been shown by the author in a companion paper at this conference [1] that such practice can lead to significant errors – typically of around 0.1 STI. See figure 2

6.3 Measurement Positions

A sufficient number of measurement positions need to be chosen to appropriately represent the area and to allow an accurate estimate of the standard deviation of the STI within the space to be calculated. Intelligibility contours can sometimes also be plotted to advantage. When testing distributed sound systems, this generally means that a greater number of off axis than on axis positions should be chosen.

Unless specified otherwise, the measuring microphone is normally positioned 1.5m above floor level for standing listeners and 1.2m when the listeners are seated, however, when measuring in class room situations, these heights will need to be adjusted as per the requirements of Bulletin 93.

6.4 Measurement Techniques

When testing sound systems, certain signal processing may need to be disabled or by-passed. This normally includes phase / frequency shifters when using FFT or MLS techniques and compressors or dynamic equalisation when using modulated signals). Other speech enhancers/ processors may or may not register appropriately, so it is usually advisable to disable them.

Ideally, the measuring microphone should be mounted on a tripod or stand to ensure that it remains stationary throughout the measurement period. This prevents temporal changes in the received test signal, due to the motion of the microphone, affecting the results. (Some forms of stimulus are particularly sensitive to this). As far as possible, the test engineer and observers should stand well away from the microphone such that masking of reflections or direct sound from loudspeakers is avoided.

6.5 Simulation of Occupancy Noise

Although occupancy noise can be mathematically corrected for, it is often useful (and usually more accurate) to simulate this and incorporate it directly within the measurement. The level and frequency spectrum of the noise source should be adjusted to equal that of the measured (or expected) occupancy noise at the measurement positions. Correcting for reverberation changes due to occupancy are less easily carried out and in certain cases may require either complex mathematical or computer modeling.

7 CONCLUSIONS

1 It has been shown that RaSTI can introduce a significant degree of error into STI measurements of sound systems. Whereas under some conditions, the errors can be readily predicted, this is frequently not the case.

2 The currently availability of measurement systems capable of measuring STI over the complete speech range, suggests that the time has arrived when RaSTI should be retired for verification purposes and replaced with full STI / CIS/ STIPa measurements. There are significant implications for any such change. These include improved modeling and calculation procedures as well as implications for sound systems design and costs.

3 Whereas it is easy to criticise RaSTI, it should not be forgotten that in the UK at least, it's introduction and adoption is without doubt the single-most important factor responsible for improving Public Address and Voice Alarm system intelligibility and quality ever to have occurred.

4 When measuring the natural intelligibility of a room or space, it is essential that a sound source with similar directivity to a human talker be employed. Failure to do so can lead to significant errors.

5 Testing or verifying the potential Intelligibility of sound systems, is very different to assessing speech performance within a room or auditorium. The implications of non linear processing and system response need to be fully appreciated and understood when undertaking such measurements.

8 REFERENCES

1 ANSI standard S3.5 1969 Methods for calculation of the Articulation Index

2 Mapp P – Measuring Speech Intelligibility in classrooms. Proc IOA Vol 25 Pt 7 Sound-Bite conference November 2003

3 Mapp P – The acoustic and Intelligibility Performance of Assistive Listening and Deaf Aid Loop (Afils) Systems. AES 114th Convention Amsterdam, March 2003

4 Davis C, Measurement of % Alcons, JAES vol 34 No 11 1986.

5 Mapp P, The measure of Intelligibility, S&VC Vol 20 No 4

6 Mapp P, Sound Power the Forgotten Loudspeaker Parameter. IOA RS 17 proc IOA vol 23 Pt 8.

7 Bradley JS Predictors of speech Intelligibility in Rooms. JASA Vol 80 1986

8 Houtgast, T, Steeneken, H & Plomp, R. Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. Acustica vol. 46, 1980.

9 Mapp P, Relationships between Speech Intelligibility Measures for Sound Systems. AES 112 Convention Munich 2002.

10 Mapp P Limitations of current sound system intelligibility verification techniques. AES 113th Convention Los Angeles 2002

11 Mapp P, Improving the Intelligibility of Aircraft PA Systems. AES 111th Convention New York. 2001.

12 Steeneken – HM Personal communication & Past, present & future of the Speech Transmission Index, International conference Soesterberg The Netherlands 2002

13 Mapp P Some further thoughts on STI – How accurate are the measurements in practice ? IOA Vol 24 Pt 8, RS 18 Stratford on Avon 2002

9 FIGURES

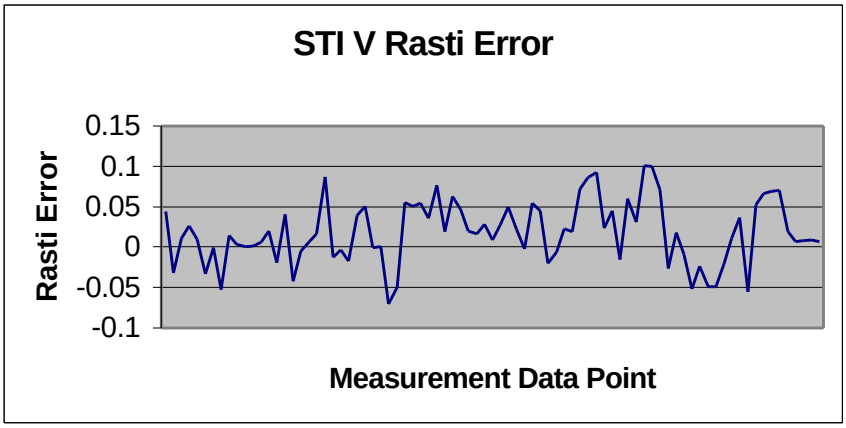

Figure 1 Typical RaSTI – STI Sound System Measurement Errors

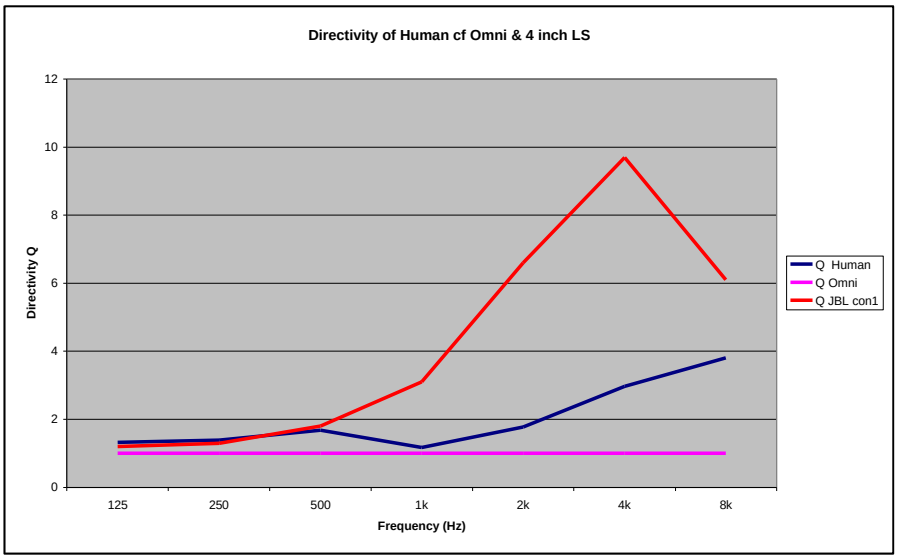

Figure 2 Typical STI / Loudspeaker Source Directivity Errors