ERROR MECHANISMS IN SPEECH INTELLIGIBILITY MEASUREMENTS

Peter Mapp Peter Mapp Associates, Colchester Essex, UK

(Petermapp@btinternet.com)

INTRODUCTION

The last decade and in particular the last 5 years has seen a significant increase of interest in Speech intelligibility testing. Furthermore, such testing is now no longer almost exclusively confined to the specialist test laboratory but is carried out in almost any public or commercial building or venue that has a PA / VA system. Furthermore, since the introduction of BB93, testing the intelligibility conditions in classrooms and other teaching areas has opened up yet further interest. The upsurge in interest and use of intelligibility testing now means that many of those tasked with the measurements are no longer specialists with experience of the potential errors that are inherent in all the current procedures. (In fact, in the author's experience remarkably few specialists are aware of many of the potential errors they may be invoking). The object of this paper is to briefly review some of the most common problems and current practical limitations. Although, a wide range of metrics will be discussed due to its increasing dominance, particular attention is paid to the Speech Transmission Index (STI) and its derivatives, RaSTI & STIPa.

TESTING METHODS

Intelligibility testing can be may take one of two forms, either Subject Based or Acoustic / Electroacoustic based. Subject base testing is the traditional method and has been referred to as the "Gold Standard" – though in practice it is often far from being that. Acoustic / Electroacoustic based methods are more recent and are all based on indirect methods, whereby a particular acoustic parameters or set of parameters that have a reasonable degree of correlation with intelligibility are measured. In addition to the objective methods that are the primary focus of this paper, subjective rating methods also exist (eg mean opinion scores). Whereas these methods enable a general view of a situation to b obtained, correlating the scores with objective measures can often be difficult as many other factors frequently also influence the judgement.

Word Scores & Subject Based Testing

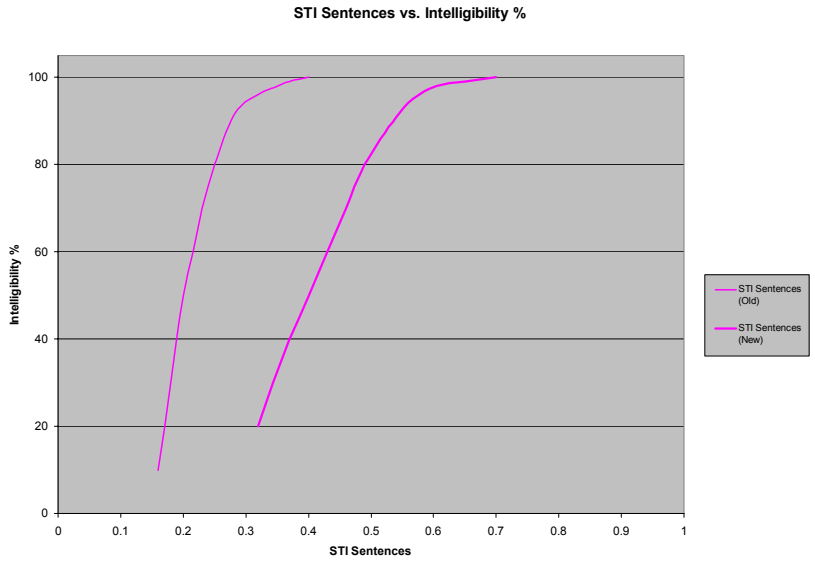
Subject based tests (often incorrectly referred to as subjective tests) such as Word Scores and Sentence Recognition, rely on a panel of listeners or jury to identify a series of words, sentences or nonsense syllables either spoken directly in the space or transmission channel under test or maybe indirectly broadcast (eg by the electronic injection of recordings). There are many variations on this basic approach, both in terns of how the test material is presented and the form of the test words / syllables / logatoms themselves. Although word score testing is by far the oldest form

of measuring intelligibility and has been well standardised, surprisingly significant variations in the score results between testers are regularly encountered. The main problems can be categorised as follows :

1 Too small a listening sample
2 Too few talkers
3 Incorrect or overemphasis of test words
4 Incorrect word rate and sequencing
5 Lack of carrier phrase
6 Poor jury training and preparation
7 Over familiarity of jury with test words /materials
8 Unintentional cues within the test material
9 Poor / incorrect measurement of SNRs

An example of the magnitude of the errors that can occur is shown in Figure 1 which plots the relationship between STI and sentences under a range of conditions as published by two leading and highly respected research institutions. Clearly one of them is in error – but which ? Equally, if well respected institutions can make such an error, then what happens with less experienced testers ?

Figure 1 Differences in sentence intelligibility for same STI



Translating laboratory test results into practical reality is also a thorny problem, particularly for example when testing PA systems, where reverberation plays an important, though not yet fully understood role. Furthermore, in reality, few announcers are properly trained in microphone and announcement / speech clarity techniques and large differences between different announcers can occur particularly as the listening acoustic conditions deteriorate.

Determining the effective signal to noise ratio is not as straightforward as may at first be thought. Speech by its very nature is a dynamic signal, with second to second variations of typically 12-20 dB though overall the speech dynamic range is around 30 dB.

Compression and other forms of dynamic processing further complicates the issue and also affects the potential intelligibility in a very non linear way.

Whereas word score testing has been referred to as the “gold standard” of intelligibility testing, in practice it is often far from this – a factor that obviously needs to be borne in mind when assessing the performance of indirect acoustic / electronic intelligibility measures and metrics. Word score testing can however be a very effective means of measuring the relative intelligibility between systems or for monitoring the changes undergone by systems as various parameters are adjusted or altered.

Indirect (Electronic & Acoustic) intelligibility measures

Whereas subject based testing may be a practical option for lab based testing, it does not lend itself well to site testing eg PA and other communication systems – where often multiple testing within a building or facility is required. (The author regularly tests facilities where >100 such tests are required). Whereas it is possible to binaurally record word score test sequences in situ for later test scoring off site, this is still prohibitively time consuming and expensive. Furthermore, the technique also introduces an additional variable – that of the transfer function of the recording and playback chain. The need for an objective but not subject related intelligibility test has been recognized for over forty years. During the intervening period several metrics have been devised. These range from the Articulation Index (AI), devised in the 1960s primarily to test communication channels to energy ratios such as C50. This latter measure, although widely used in the field of auditorium acoustics has never been standardised and a formal scale produced. It is however a useful measure where reverberation is the primary intelligibility degradation factor.

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 a previous paper by the author, failure to do employ the correct directivity can lead to significant error. (see also [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.

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.

STI, RaSTI & STIPa

It is because of its inherent ability to account for both noise and reverberation effects that has allowed STI to become so widely adopted (together with the production of dedicated instrumentation). In a similar manner to AI, STI was initially conceived to measure the performance of communication channels but its uses have expanded way beyond this initial realisation. STI is essentially monaural in nature, which, under some circumstances (as further described later) can lead to an underestimation of the potential intelligibility. Over the past 10 – 15 years, STI has become the defacto standard for measuring the intelligibility of PA and other related voice communication systems. It is enshrined in many national and international standards ranging from emergency sound systems to aircraft PA & communication systems. [2]. The complexity of the STI measurement however inhibited its early adoption as a general practical measure. However, the introduction of a portable measuring device by B & K in 1985 to measure RaSTI enabled the technique to become more widely adopted.

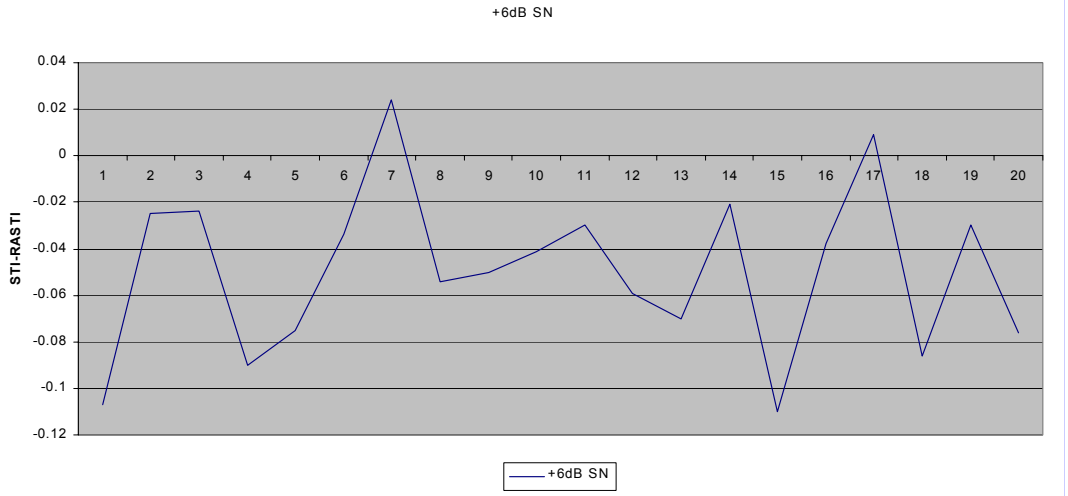
How Accurate is RaSTI ?

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 the dominant descriptor. (Though the recent introduction of STIPa and the ready availability of a range of measurement devices is now making this the more widely used measure). Limiting RaSTI's measurement 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 2 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]).

Figure 2 Rasti Error as compared to STI



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, it is interesting that in his last papers Steeneken has re-named RaSTI, Room Acoustic Speech Transmission Index rather than Rapid !

STIPA

STIPa is similar to Rasti in that it uses a dedicated modulated signals, but instead employs a 'sparse matrix' that 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. [12]. 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.

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

Table 2 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 – a major advantage when testing large spaces or buildings.

Agreement between STIPa & STI is generally very good. Table 3 below for example presents a typical comparison made in a reverberant space, using a high density, distributed sound system.

Table 5

Position	1	2	3	4	5	6	7	8	9	10
STIPa	0.37	0.39	0.47	0.45	0.44	0.52	0.48	0.33	0.40	0.48
Mlssa	0.40	0.40	0.46	0.42	0.44	0.51	0.44	0.31	0.42	0.48

It is interesting to note that the mean values for the space, as measured by the two different methods, are identical at 0.44 STI. The standard deviations are also comparable.

Although STI was conceived to use a modulated signal, Schroeder showed that the STI could also be derived from a system's impulse response. The advent of more powerful personal computers and analysis equipment such as MISSA and TEF in the early 1990s enabled STI to reach a wider (though still very specialist) user base. Its wider adoption and user base also began to show up limitations of the techniques, the limited frequency range of RaSTI for example being shown to be a major limitation when testing PA systems. The introduction in 2001 of STIPa, with its 6 – 7 octave and range, whilst overcoming this limitation, still did not resolve other fundamental flaws in the STI technique. Many of these limitations are not widely appreciated – though they have can have considerable contractual implications. At the time of writing, in the author's view the following limitations are still to be resolved.

1 Irregular frequency response & sound system equalisation
2 Binaural effects
3 Echoes & strong discrete reflection effects
4 Compression
5 Distortion (particularly when combined with other effects)
6 Level dependency (particularly under reverb conditions)

7 Software and STIPa equipment variations
8 Poor measurement techniques

Whilst items 7 & 8 cannot be attributed directly to the STI technique itself, none the less, they are important potential error mechanisms that still need addressing.

Frequency Errors & Effects

Previous research & papers by the author [10,11,13,14] have well established the fact that STI is often inaccurate under quiet, reverberant conditions, when the PA or loudspeaker system under test exhibits an irregular frequency response – and particularly if the system contains lower mid and mid frequency response peaks. Figure 3 for example shows the frequency response of a good quality sound system in a reverberant church (2.5 sec RT) before and after equalisation. The pink, post EQ curve gave rise to a significant improvement in word score intelligibility (21%) whilst the measured STI for both conditions remained the same.

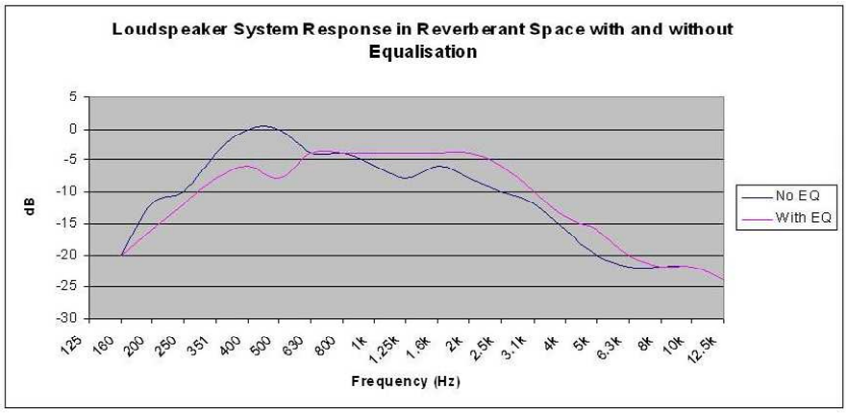


Figure 3

Figure 4 is a more dramatic case and is an example taken from a series of experiments conducted by the author whereby the frequency response of a sound system was adjusted and correlated against word score tests and STI measurements. Despite the large variation in response the STI again remained the same at 0.46 whilst the equivalent word scores varied between 0.46 and 0.30. (ie from reasonably intelligible with careful listening) to extremely poor intelligibility with only the occasional word being correctly deciphered.

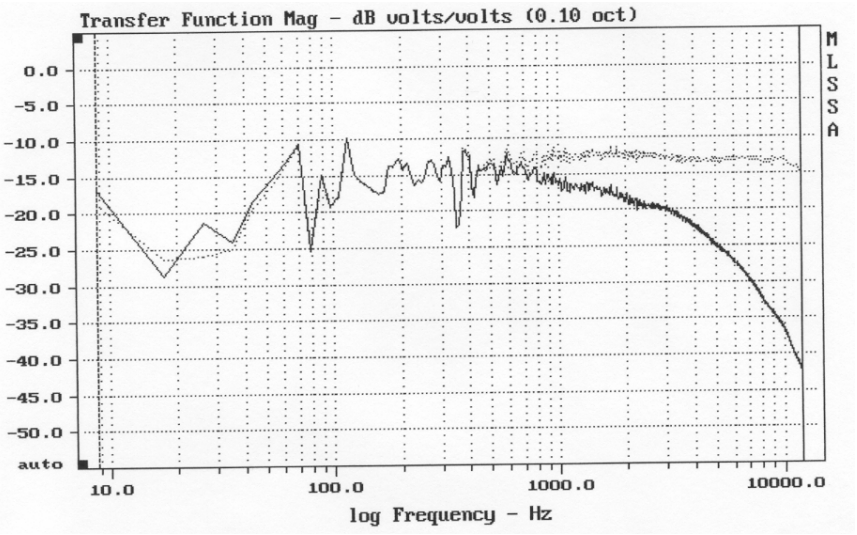


Figure 4 Two frequency responses with different intelligibility scores but same measured STI

Binaural Effects

Whereas we listen with two ears (ie binaurally) STI measurements are made with a single non directional microphone. Whereas under many conditions, this can give a remarkably good correlation, there are situations were the lack of directional information and cross-correlation processing of the two signals that the ear / brain system carries out, results in too simplistic a measure. Typical situations where this

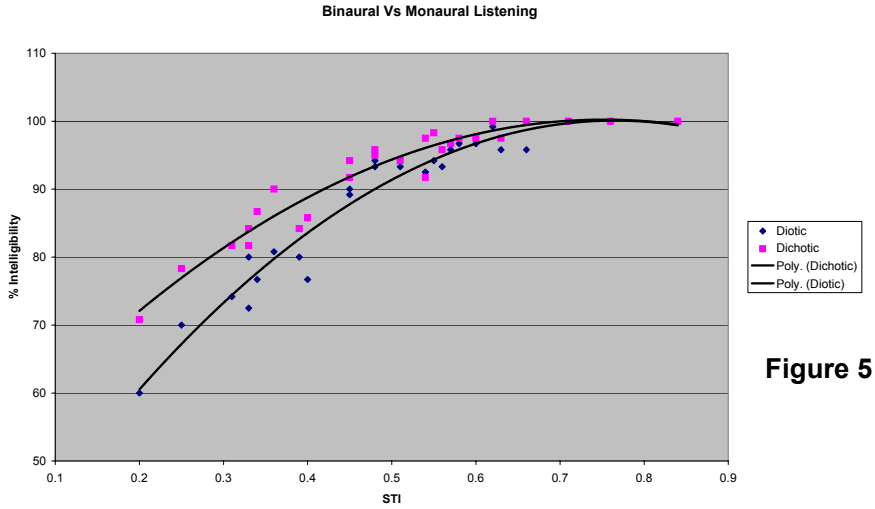


Figure 5

error can particularly occur include (1) where noise and speech arrive at the listener form different directions (2) where strong discrete reflections arrive from a different direction to the speech source (3) discrete sound or multiple sound sources with general reverberation. Figure 5 shows a general relationship identified by the author. It is interesting to note that the binaural advantage reduces as the overall level of intelligibility increases.

Echoes & Discrete Reflections

It has long been observed that discrete reflections can adversely affect the accuracy of STI measurements. This is particularly the case with RaSTI & STIPa. A detailed study of the effects of echoes on STI & Intelligibility has not been made, though there is significant anecdotal evidence relating to the subject. Discrete echoes can be very disturbing but their effects are very dependent on the particular reflection sequence in which they occur and the general reverberation time of the space. These inherent variables make definitive study difficult and suggest that an impulse response measurement should also be taken in any large space subject to an STI audit, so that further analysis of the reflection sequence can be undertaken and the validity of the STI measurement verified.

Compression

Speech signal Compression has regularly been employed as a method of improving the intelligibility of communication systems, hearing aids and PA systems for more than 30 years. It is a deliberate reduction in the speech modulation and so is totally at odds with the STI concept, which is looking for the preservation of speech modulation. Measurement o systems containing compressors can therefore lead to a two fold error effect, as depending on how the compressor ‘attack’ and ‘decay’ and ‘compression ratio’ parameters are set, the STI measurement itself may also be degraded. (ie the observed effect of compression is to increase intelligibility and may therefore be at odds with the STI reading , which in itself may be further reduced due to the non linear behaviour of the compressor. Whereas one can by pass the compression circuitry in most applications and so obtain an accurate STI measurement, the lack of correlation is concerning – particularly in PA and communication systems where the requirement to deliver an given degree of intelligibility can have significant contractual implications.

Distortion

Little research has been undertaken relating to the effects of distortion on intelligibility and STI. Whereas it is generally agreed that the lower the distortion the better, it has also been shown that adding in some forms of distortion at low level can actually enhance clarity and intelligibility ! However, modern sound systems should not exhibit an obvious signs f distortion and if they do, then this is indicating a problem that should be fixed before any attempt to measure the STI (or intelligibility) is made. A circumstance however where distortion may occur and is needed to be included is when testing PA systems at their maximum required operating levels. This is often the case when testing systems designed for emergency purposes that need t operate at high levels in order to overcome adverse background noise conditions. However, a comparison of the duty cycles of the test signal and real speech often shows there to be a discrepancy between the two that should be accounted for when testing in this way.

Measurement Equipment Errors & Differences

The recent introduction of STIPa has radically changed the intelligibility measurement market. Prior to its introduction there was only the B&K Rasti meter available as a dedicated measurement instrument and 3 or 4 software analysis programs deriving STI from measured impulse responses. To the author's knowledge there are currently at least five STIPa hardware measurement platforms currently available on the international market as well as 3 software IR analysis programs). . Interestingly, when tested by the author, 3 of the 5 measurement platforms, had significant flaws in them when initially released !. Due to campaigning by the author the situation has significantly improved, but it is worrying that well known, leading equipment manufacturers could release highly inaccurate measurement equipment.

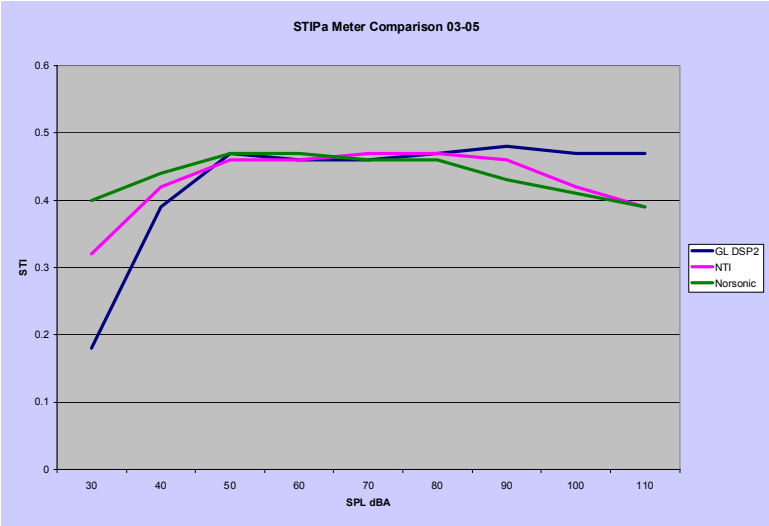


Figure 6 Comparison of Different STIPa meters

Whereas the above errors relate the techniques and equipment implementations, the way in which the measurements or word scores are carried out add yet another dimension to the problem that is beyond the scope of this paper.

CONCLUSIONS

It is clear that the measurement of speech intelligibility is an extremely complex matter, with many pitfalls to catch out the unwary. There is still a disturbing difference between some of the measurement equipment and computer programs currently on the market – though the situation is generally improving. The most significant errors are now tending to be related to the way in which the tests are carried out and the calibration of the equipment or test. Although word scores have been referred to as the 'Gold Standard', in the authors view, this is often far from being the case. Relating intelligibility scores or acoustic measures to the reality of the perceived intelligibility of a PA system working under normal operational conditions still has a long way to go, though the available metrics and techniques currently available do enable repeatable, comparative data to be obtained.

REFERENCES & BIBLIOGRAPHY

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

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

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

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

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 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

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

14 Mapp P, Some Effects of Equalisation on Sound System Intelligibility & STI Measurement ERROR Proc IOA Vol 23 Pt 8