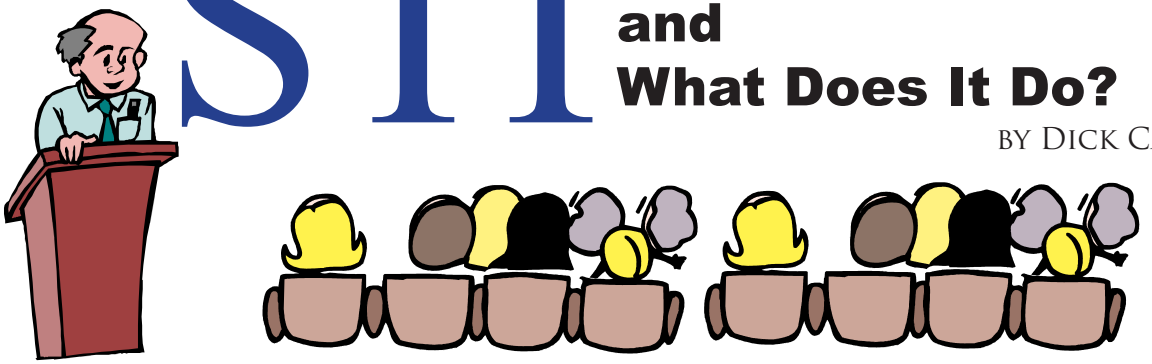
STI Where Did It Come From and What Does It Do?

BY DICK CAMPBELL



Introduction

"Thank you for inviting me to this lecture on languages. I am going to discuss a new language recently heard at the Convention Center during an ALcons tribal conference. The following sample seems to be a form of counting: 'On tu ee aw ei eh eh eh ie eh' although this has not been confirmed. It must have been an important message because I expected the ALcons to run out because they looked confused, but they stayed".

There are many places where accurately-heard counting can be important, like aircraft fight decks control towers, ambulance radios, wingmen, GPS audible guides, drive-up windows, evacuation announcements, but by no means least 'The Word of God'. This last item in my abbreviated list is an area where many Syn-Aud-Con members ply their trade.

Worship is a special place where the icons, fancy robes and meaningful gestures fade into the background when compared with having the congregation understand what the preacher is saying (the second biggest item is having the music "sound like it's supposed to sound!"). After a worship for example you might greet the minister on the front steps and complain about the "horrible acoustics…I could not understand what you said!" You are talking directly to the boss and hopefully something will be done about it.

The World Before Computers

We've been observing problems in speech perception (in the written record) since about 350 B.C. From Exodus Book 26 to Aristotle to Lucretius to the Roman architect Vitruvius ca. 50 B.C. who left 10 books on architecture, in one of which he provided an excellent description of speech transmission in theatres [1]. Fast-forward 1500 years to find further developments in acoustics as a "natural science".

Eventually the "science" of speech transmission surged after the telephone came into being with Bell Labs fostering research and development resulting in the Articulation Index (AI) ca. 1940. Then came WW2 with crews of big noisy airplanes talking to each other without eyeball contact resulting in a Harvard University contract that produced a refined set of AI bands and the intelligibility test word corpus called "phonetically-balanced." From this evolved the computer algorithms we use today including STI, RASTI and SII. AI (and offspring) is a measure of the "hardware" on a scale from 0 to 1. It characterizes the entire path between the speaker and the listener so we can call it "pathware."

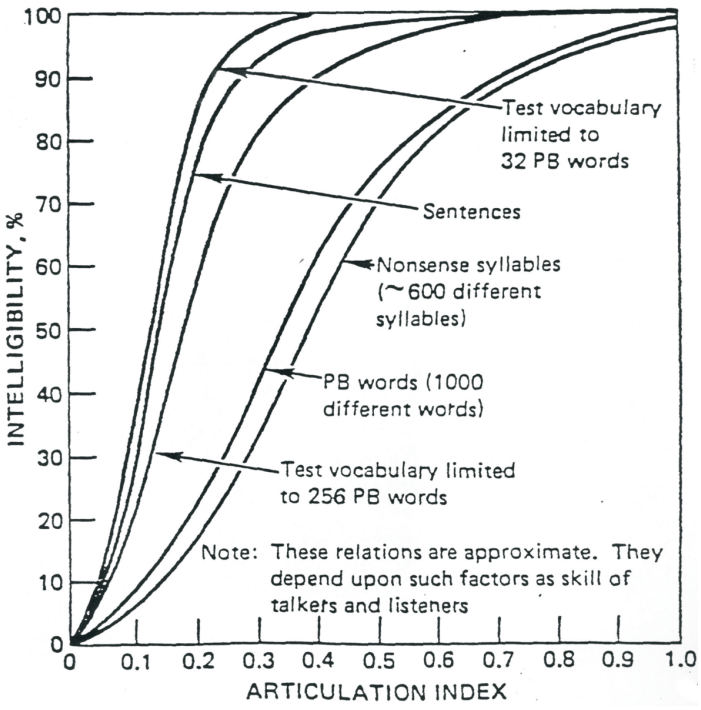


Figure 1 - The effect of message type on speech intelligibility. Note the dramatic difference between single utterances and sentences.

Dr. Leo Beranek sent the following email message [edited by the author]:

Starting in November 1940 and ending September 1945, I was Director of Harvard's Electro-Acoustic Laboratory, which was funded by the US Office of Scientific Research and Development OSRD....Our lab worked on voice communication in combat vehicles. Our most important work was making voice communication possible at aircraft flying at high altitudes in unpressurized aircraft, with pilots using oxygen masks.

.... Harvey Fletcher, director of acoustics research at Bell Labs, was a member of our advisory committee. He had French and Steinberg send me some of their work on AI. We used that and built on it to develop a simple and effective means for determining AI...I published our whole story in the Proceedings of the IRE, 35, 880-860 (Sept 1947). There are some differences. I developed an AI chart (Fig. 10 on p. 886) which makes it easy to determine AI. This chart was later put into an ANSI standard. Their teachings lead to a somewhat different result. I then show that this method came close to the results of actual articulation tests. I also showed for the first time how very high speech levels (possible by turning up the amplifier gain) resulted in reduced intelligibility...

We measure our ability to understand speech on a scale labeled 'intelligibility' that ranges from 0% to 100% of correctly received messages. Messages come to our ears in many flavors and they face numerous obstacles en route. The classification of the type of message is a vital ingredient in framing the problem numerically. We can call intelligibility "brainware".

So, we have pathware and brainware measures but they are as mathematicians would say "orthogonal" meaning that they live on axes that are 90-degrees apart. The only way to connect them is through a curve or function drawn between the axes. The curve shape is discovered only by using human talkers and listeners in enough repeated tests to be statistically significant -- hundreds of tests per curve, as seen in Fig. 1.

But why is there more than one curve? This is where the type of message is accounted for. When the AI is poor, familiar messages can be understood easier than strange or unexpected ones. If we do all of our testing using American English, what happens when a non-native listener is a test subject? The curve changes shape. We are still learning what this new shape will be for any other language. Meantime we have to guess at it – a subject of lively discussion right now.

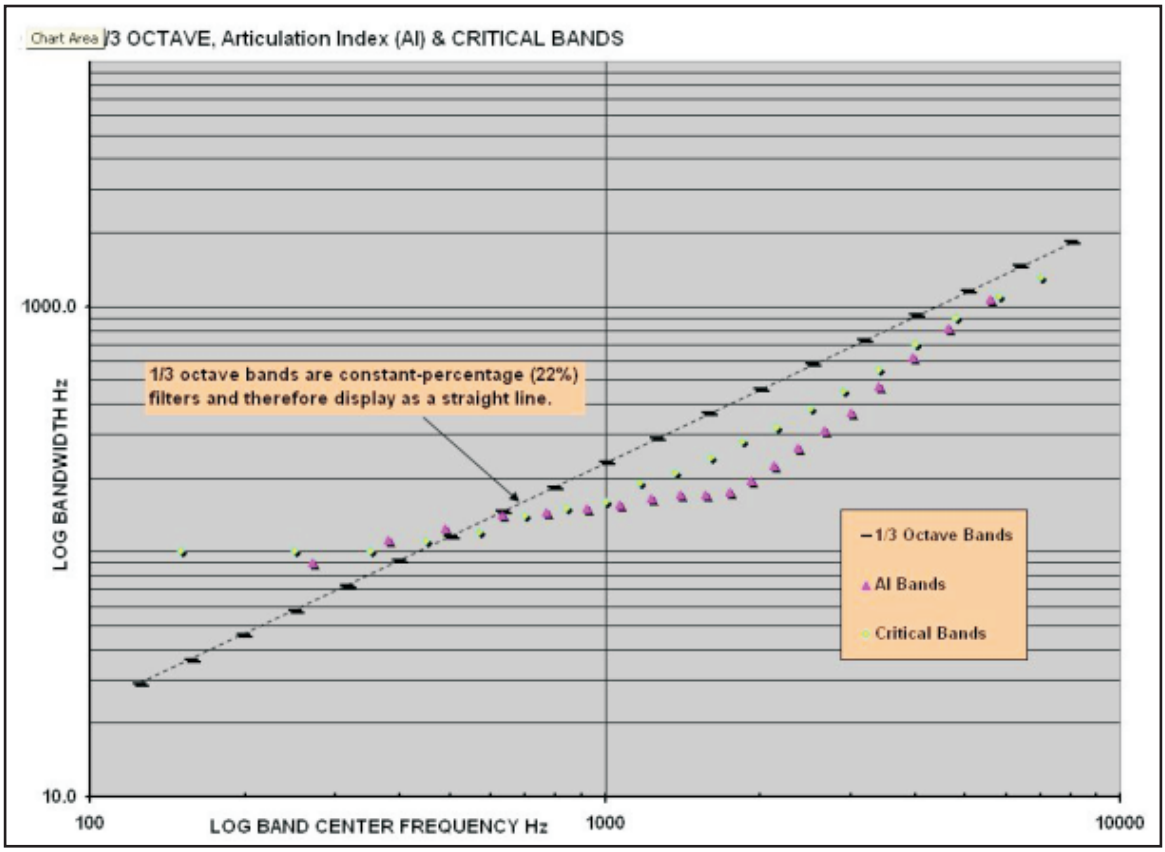


Figure 2 -Log-log plot of bandwidth and center frequency for three types of analysis bands related to speech intelligibility research.

There is a frequent mix-up between AI (pathware) and intelligibility (brainware). Please do not say that "AI is a measure of intelligibility." It's not. AI is only a measurement of the path between the talker and the listener -- as influenced by frequency response, distortion, reverberation and noise. The actual intelligibility of a message is a function of how X talks and Y listens through such a path.

Since AI was developed in the USA the data that provided the curves connecting AI with intelligibility were taken using American English and a phonetically balanced word set. It was discovered early on that these words had to be embedded into a sentence to sync the listener to their utterance. Thus "You will write [word] on the line" gives a heads-up to turn on your brain and get the word. Since the test words themselves have no meaning within the sentence, they are sometimes referred to as "nonsense syllables."

The basic concept of AI is simple: broadband speech has an amplitude modulation of about 30dB between envelope peaks. Continuous speech displayed on an oscilloscope with a slow time base shows the envelope and will reveal its peak-to-peak behavior. The RMS value statistically is about 12dB below the upper peak and about 18dB above the lower peak. The developers of AI reasoned that any interfering noise that masks the envelope peak-peak range will be detrimental to intelligibility. The entire speech envelope has to get through the path unaltered for the highest possible intelligibility.

Note that I used the word "statistically." The various procedures for intelligibility calculation depend upon statistical observations – the aggregate speaking-listening tests between a wide variety of people repeated hundreds of times. Any one individual may be at the edge of the band of variability either very good or very bad. The procedures for calculating intelligibility represent Mary, Dick, Alice, Sam, Bob, June and Pat -- all taken together in one statistical heap.

The speech envelope of course contains within it all of the speech frequencies. It has long been shown that consonant sounds are more important than vowel sounds. But how to slice up the speech spectrum? The result of many hours of AI research was to create 20 filter bands spanning 230Hz to 6166Hz. The bands in the ~1300Hz – 3200Hz region have small bandwidth and are therefore crowded together on a log frequency scale. Therein lay the important consonant sounds. At the extreme upper and at lower frequencies the bands are wider; hence there are fewer of them because they have lesser contribution to intelligibility. The bottom band centered at 270Hz (230Hz to 320Hz) has a ~33% bandwidth. The two most important bands at 1740Hz and 1920Hz have only a 10.1% bandwidth. We can label this collection of bands "brainware bands."

In the AI calculation system each band has equal "importance" and the changing percent bandwidth accounts for their respective contribution. There is a problem here – if the interfering noise is given in 1/3-octave bands where each "measurement" band has the same ~22% bandwidth. If we use these data for an AI calculation rather than the AI bands, then the "importance" of each 1/3-octave band is adjusted by changing its effectiveness for interference. These are called "importance functions." More on this later.

Fast forward 20 years from the AI research and we find researchers who study masking making significant progress. This study primarily involves how noise can render a pure tone inaudible if the noise were high enough. Beginning with French & Steinberg (1947) and later Zwicker (1961) in the Netherlands, the masking capability of interfering noise was completely characterized. We have a remarkable ability to switch on a narrow-band filter to enhance the detection of tone signals. The result was a series of "masking bands" called "critical bands". This feature is a component of "brainware" -- relevant but not essential to this discussion.

Now we have three different sets of frequency bands to think about:

* AI bands (20) each having equal contribution to speech intelligibility

* Critical bands (21) that define our ability to hear tonal signals in noise (masking)

* One-third octave bands – the ones we measure

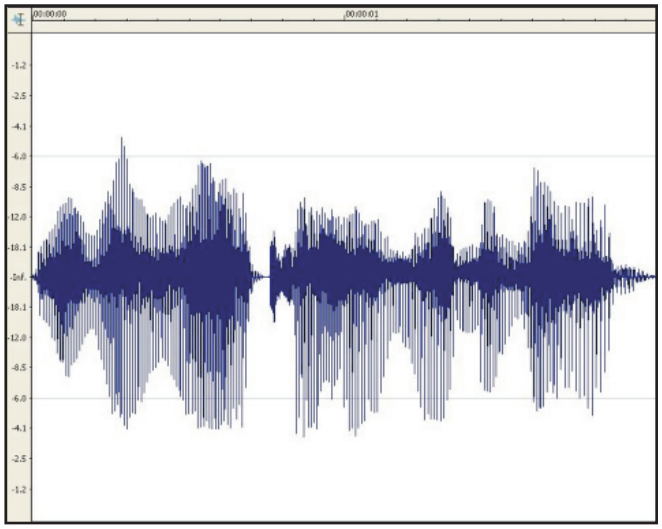


Figure 3 - Speech wave from PB List 2 in a quiet environment (male voice). The utterance is the second word of 50 words: "You will write tang on the line". Each PB word "tang" in this case from the list is embedded in the same 'carrier' sentence.

with an instrument.

Figure 2 shows these three bands plotted together. Notice the significant difference between the AI bands and 1/3-octave bands in the frequency region of speech consonants 1000 to 3000 Hz. The AI bands are narrow and squeezed together to provide the "importance functions."

The dilemma for using 1/3 or 1/1-octave noise in the AI calculation is obvious in the chart. The method for doing this is covered in the ANSI Standard S3.5-1997 that defines yet another title: SII "Speech Intelligibility Index".

What About Reverberation?

Reverberation is a form of transient interfering noise that decays following a speech utterance. If the decay is long, it may be still strong enough to mask the next utterance. Almost everyone has experienced poor speech intelligibility in spaces with long reverberation time.

The AI system takes this into account by making certain assumptions about the room and the source. It assumes one source with known directivity factor and it assumes that the space is a "Sabine" space – e.g. uniform diffuse field with exponential decay [4]. Maybe this was practical in the '40's, but it is not too useful with modern speech reinforcement systems. The reality is that most spaces are not "Sabine," and the directivity factor of a distributed audio system is unquantifiable.

The World After the Computer

The AI computation changed forever with the publication of a seminal paper by Steeneken and Houtgast in 1980 entitled "A Physical Method for Measuring Speech-Transmission Quality" JASA (67). A year later Schroeder published his paper on the modulation transfer function in Acoustica (49). The pieces of the puzzle were now in place to do the whole AI-like calculation with a computer. The name was changed to "Speech Transmission Index" (STI) to distinguish it from other methods.

I mentioned Modulation Transfer Function (MTF). We will not go into the calculation details of the MTF however the application of this parameter needs to be understood before the STI procedure makes sense.

MTF Concepts

Figure 3 shows a typical speech wave in a quiet environment. What follows is the statistical report of this utterance as calculated by the Sound Forge™ editor [3]: Note that the maximum and minimum and RMS values are given as dB below full scale.

1) Minimum sample position (Time)	00:00:00.869
2) Minimum sample value (dB)	-3.559
3) Maximum sample position (Time)	00:00:00.285
4) Maximum sample value (dB)	-4.730
5) RMS level (dB)	-18.484

We can now calculate:
Average of 2 & 4 = -4.15 dB (below FS)
Subtract that from 5 = 14.3 dB peak–to-RMS ratio

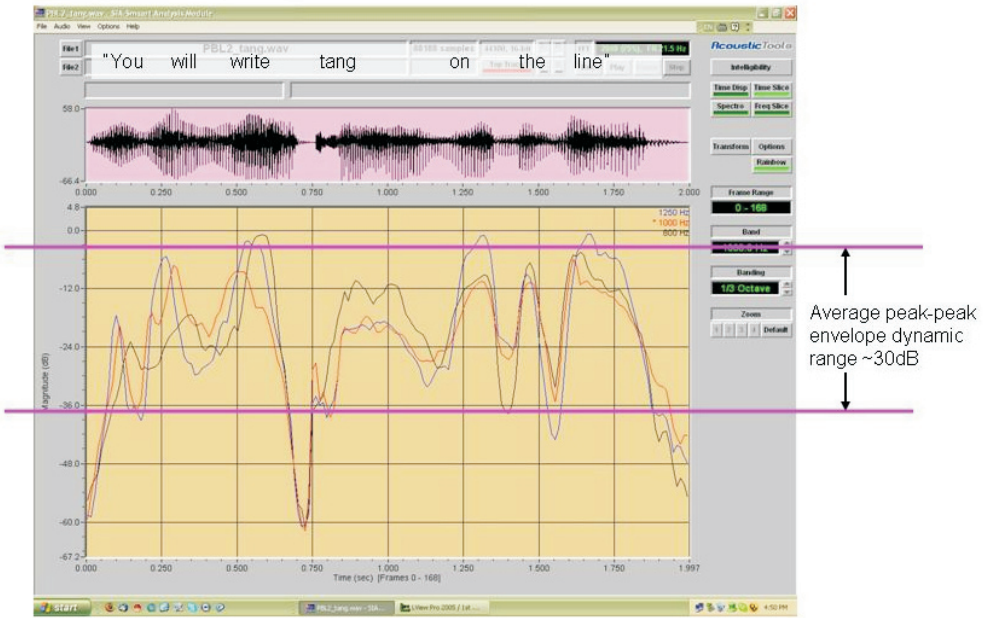


Figure 4 - Three 1/3-octave band pressure level plots of the phrase "You will write tang on the line".

If we load the 'tang' phrase into SIASmaart™ analysis tool we can obtain filtered levels as a function of time (time slice). Figure 3 shows the 800, 1000, and 1250 Hz 1/3-octave bands.

Notice that the envelope of the band pressure level variation is a signal. It is possible to do a frequency analysis on this signal and calculate its frequency content. This will play an important role in the STI calculation. The signal can be described in the frequency domain using 1/3-octave bands from 0.63 to 12.5 Hz. That's about as fast as our body parts can move to affect speech!

If this is a modulation envelope what is being modulated? It is the speech frequency content in the audio band – the stuff that looks like hash in Fig. 2 – that may be described in 1/1-octave or 1/3-octave bands for example from 125Hz to 4000Hz.

In Fig. 3 we see the envelope in three adjacent 1/3-octave bands - 800, 1000, and 1250Hz. Note the significant difference between them.

Next we mix with this speech some pink noise that has the same RMS value – what we might call a '0dB' signal-to-noise ratio (RMS). Figure 5 shows the result as displayed in Sound Forge.

It appears that this is all noise. However, the speech can be heard and is "understandable" after listening to the quiet version a few times (brainware is wonderful). Here are the stats on this trace:

1)	Minimum sample position (Time)	00:00:00.311
2)	Minimum sample value (dB)	-0.517
3)	Maximum sample position (Time)	00:00:00.285
4)	Maximum sample value (dB)	-2.413
5)	RMS level (dB)	-15.413

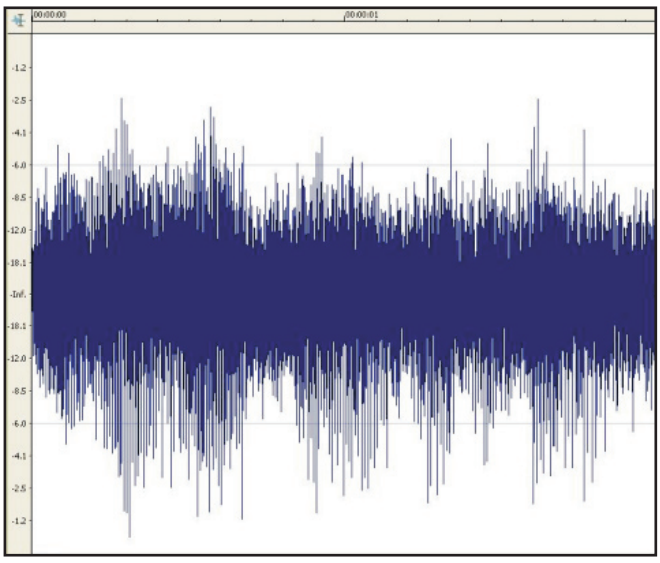


Figure 5 - Equal RMS mix of speech from Fig. 3 and pink noise

As expected the RMS value of the combination increased about 3dB. The peaks seem to be more defined by the noise rather than the speech.

We can now look at both quiet and noisy files in SIA Smaart in Figure 6. The top window shows the quiet file in black and the noisy one in red. In this figure the filtering is the 1000Hz octave band. Notice that the dynamic range has been severely reduced to about 9 dB. The speech envelope below -9 dB in the graph is obliterated by the noise. It is essential to remember that all of the speech dynamic must get through the system for excellent intelligibility (in addition to other requirements).

The transfer of the modulation envelope in Fig. 5 (along with its contents) from talker to listener has been reduced by ~21dB leaving ~9dB for audibility. If we can characterize this loss of modulation mathematically the result may be used for a computerized evaluation of the noisy speech signal.

The MTF Computation

The AI system is an early form of estimating the modulation transfer function using only signal-to-noise data. The work of Steeneken and Houtgast did not assume any particular acoustic behavior because it used an impulse response of the system acquired at the listener's location. An impulse response contains the total information in the path including reverberation and background noise. After appropriate transforms to the frequency domain, further computation can estimate the amount of a modulation envelope remaining for audibility. Thus was born the Speech Transmission Index STI.

Also, it is possible to generate a synthetic signal that contains a series of sine waves in each audio band that are modulated by a low-frequency envelope signal and measure the STI directly with a portable meter. For this purpose it was easier to restrict the octave bands to two: 500Hz and 2000Hz. This required much less computing horsepower, and was called RASTI, originally meaning "Rapid Assessment of Speech Transmission Index." On a modern PC, the full STI table of Table 1 is displayed in less than 1/2-second.

This is beginning to look like a two-dimensional problem in the frequency domain.

Dimension 1: The envelope modulation from 0.63 to 12.5 Hz that is reduced by interfering noise in the path.

Dimension 2: The contents of the envelope from 125Hz to 8000Hz that can be altered by the frequency response of the path.

The topologist's way of solving this problem is to build a matrix – put the envelope bands on one axis and the audio bands on the other. Where each intersects is a

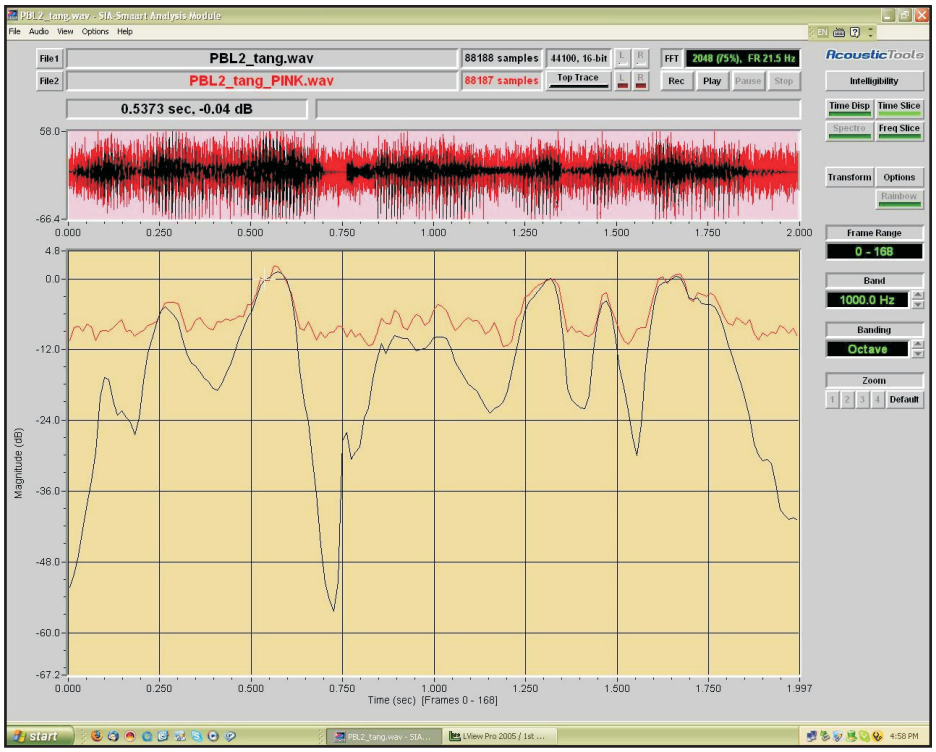


Figure 6 - 1kHz-octave band filtered comparison of the PB phrase "You will write tang on the line" in both quiet (black) and with 0 dB RMS signal-to-noise (red).

data point provided by calculating the particular modulation transfer function.

At any intersection if all of the envelope gets through and if the octave band response meets the criteria for 'flat' speech then the MTF = 1.

It's actually a very clever idea. Most STI matrices are displayed with the envelope 1/3-octave bands on the "Y" axis and the speech frequency octave bands on the "X" axis.

The columns are averaged with a weighting to form the octave band modulation transmission index. Then these numbers are averaged across using another weighting (importance function) to finally compute the STI. The octave weightings across (at the bottom of the table) are a function of the particular standard being followed. No one set is adopted universally.

One paper [8] gives an example of an STI computation. One needs to know the modulation frequency (0.63 – 12.5), the EDT and the signal level-minus-noise level (dB) for each octave band. It is a simple formula the authors call the "modulation reduction factor":

$$m(F) = \frac{1}{\sqrt{1+\left[\frac{2\pi FT}{13.8}\right]^2}} * \frac{1}{1+10^{-L_{SN}/10}}$$

Where F = modulation frequency (14 of them, 0.63Hz to 12.5Hz), T = EDT, LSN = Signal – Noise in each octave band (dB), making 98 items to calculate. The reference does not explain where the 13.8 comes from. If T = 0, and LSN is a high value, then the result tends toward unity. If LSN = 0dB then the second term of the equation is 0.5.

Next, an apparent S/N in dB is calculated for each matrix intersection:

$$S/N_{APP} = 10 * LOG(\frac{m}{1-m})$$

The results are clamped to +15dB or -15dB if they exceed these values.

Next, the S/N values within each octave band (14) are averaged to form an octave apparent S/N. Next, the octave values are averaged with a weighting as follows:

Band	125	250	500	1K	2K	4K	8K
STI(S&H)	0.13	0.14	0.11	0.12	0.19	0.17	0.14
AI(F&S)	0.00	0.074	0.149	0.198	0.337	0.241	0.00

Note: S&H is Steeneken and Houtgast, F&S is French and Steinburg

In other words, the 2K band controls 19% of the result in the STI calculation, while AI is assigned 33.7%. These are the importance functions. Notice they add across to unity. There are several different sets of importance functions.

The last step (whew!) is to convert the final S/N in dB into the STI:

MTF Matrix (Uncalibrated)

Frequency-Hz	125	250	500	1000	2000	4000	8000
0.63	0.930	0.943	0.920	0.900	0.916	0.942	0.987
0.80	0.926	0.940	0.917	0.896	0.913	0.939	0.986
1.00	0.914	0.933	0.907	0.886	0.904	0.930	0.984
1.25	0.885	0.915	0.886	0.863	0.883	0.910	0.977
1.60	0.832	0.882	0.845	0.819	0.844	0.872	0.963
2.00	0.773	0.844	0.798	0.769	0.800	0.827	0.945
2.50	0.727	0.808	0.754	0.724	0.763	0.783	0.922
3.15	0.708	0.767	0.705	0.669	0.729	0.739	0.890
4.00	0.721	0.704	0.634	0.588	0.687	0.684	0.847
5.00	0.733	0.632	0.551	0.531	0.642	0.627	0.796
6.30	0.673	0.544	0.494	0.488	0.589	0.564	0.731
8.00	0.587	0.422	0.481	0.432	0.514	0.497	0.648
10.00	0.579	0.317	0.457	0.405	0.431	0.419	0.557
12.50	0.486	0.285	0.404	0.349	0.393	0.352	0.486
octave MTI	0.681	0.670	0.648	0.622	0.658	0.674	0.803

STI value= 0.680 ALcons= 4.3% Rating= GOOD

Table 1 - A typical STI matrix from a MLSSA impulse response. Ozawa Hall, Tanglewood, across stage with seated group, August, 1999, RHC.

$$STI = \frac{S/N_{AVG} + 15}{30}$$

I like the preceding formulation because it's easy to follow. Some of the steps shown above can be combined, but would be more obscure to a general readership. Also, for those who puzzle over the procedures in the SII standard, there are strong similarities to what I have shown here.

Impulse Responses

All of the variables required for the STI calculation shown above can be extracted from an impulse response (IR). The IR can be 1/1-octave band filtered in the time domain -- the signal level is in the initial impulse to arrive, the background noise is in the tail end of the filtered IR, and the EDT can be measured from the filtered Schroeder decay plot. Some computer programs use algorithms that are a variation on this theme. Experts in digital signal processing do much of the analysis in the time domain, as in MLSSA. The MTF can be calculated directly from the squared impulse response, see [10], equation 20, p14, or [9].

At the risk of sounding redundant, an IR from a real location in a real hall with real audience is about as far from a Sabine space as you can get. That's what makes the STI so useful compared with the old AI method.

Applying the STI

Figure 7 is a revealing plot. It shows a contour map of my old lab at WPI with about 30 students seated at specific grid intersections. Each contour line follows the same intelligibility score as obtained using PB words. It is called an "iso-intelligibility map".

Figure 7 clearly illustrates the need for research in inter-language intelligibility application. The region below 38% on Fig. 7 was where three students from Argentina sat together. These non-native English speakers almost always sat together in classes and would assist each other as needed (which I encouraged), speaking in Spanish. In this test they were "on their own."

The burning question right now concerns the STI required to guarantee acceptable intelligibility to a mixed-language (but "fluent") population. Referring to Fig. 1, to bump a 38% score to, say 50% would require an increase in STI of ~ 0.1. So if 0.5 is acceptable based upon long experience with the English language, this modest test shows that 0.6 might be a better number.

I noticed that PBS has a documentary on the aviation disaster at Tenerife, March 1977, 583 dead. When that happened, I was chairman of a Federal Advisory Commission (RTCA) on performance standards for audio equipment in civil aircraft. A member of my committee was the audio engineer from KLM, and a personal friend of the pilot. When we met shortly after the collision, he was devastated – and over dinner said he had listened to the tapes and concluded that it was mainly a speech intelligibility problem. A Spanish controller speaking to a Dutch pilot in English: two strikes and the fog made three.

To end this on a happier note, those who follow the progress of classroom acoustics are pleased. For de-

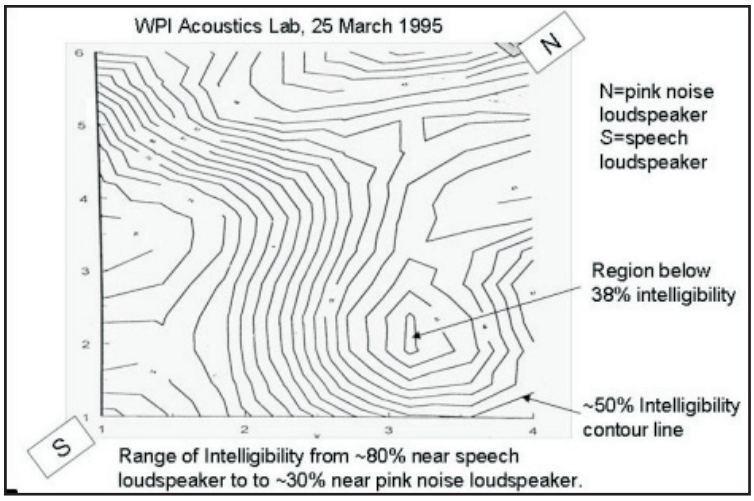


Figure 7 - Equal-intelligibility contour map of ~ 30 students sitting at tables aligned to a grid. The speech source and pink noise source were set to the same RMS sound level in the center of the room. The test material was PB lists 2, 3 and 5 in two runs each (300 monosyllables), male voice. PB list 1 was used twice for training prior to data taking. Contour interval = 2%.

cades classrooms were neglected for acoustic treatment until researchers started to show that bad acoustics led to poor learning. This wave of knowledge began with the speech scientists and passed to the architectural acousticians who knew what to do about it [11]. It's now an ANSI Standard in the USA [12] and the subject can be found in proposed laws and standards throughout the world. Hopefully, the next generation of kids will not be impaired by such faulty pathware.

References

Here are some nice web sites to help sort this out:

http://www.meyersound.com/support/papers/speech/section4.htm
http://www.armchair.com/sci/brunt1.html
http://www.gold-line.com/pdf/articles/p_measure_TNO.pdf

A % ALcons vs. STI calculator:
http://www.sengpielaudio.com/calculator-ALcons-STI.htm

An excellent Power Point presentation from John Erdreich:
http://www.acousticalconsultant.com/articles/speecgintellig.ppt

[1]Hunt "Origins in Acoustics" Yale UP ISBN 0-300-02220
[2] JASA, 19, 1947
[3]http://www.sonymediasoftware.com/products/showproduct.asp?pid=961
[4]See Beranek Acoustics Table 13.6, p415
[5]E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, Berlin: Springer, 1990
[6]http://www.sii.to/RhebergenVersfeld_JASA_2005_.pdf
[7]http://www.freepatentsonline.com/5729658.html
[8]http://www.acoustics.hut.fi/asf/bnam04/webprosari/papers/o27.pdf
[9]Douglas D. Rife, JAES 40(10), 1992 October
[10]ANSI S2.5-1997: Methods for Calculation of the Speech Intelligibility Index
[11]"Classroom Acoustics 1 and 2" booklets available from the ASA: http://asa.aip.org
[12]ANSI S12.60-2002

The author thanks Dave Read for his constructive comments. Dave has a wonderful article on speech intelligibility coming in a future issue of Sound and Communications.