Measuring Speech Intelligibility

They say your new twiddle-o-matic makes your audio clearer – but *how much* clearer?

BEFORE AND AFTER. If you've ever upgraded your antenna from the proverbial piece of wet string to a six element beam, the improvement will have been nothing short of staggering. The change would have been easy to quantify using 'before and after' signal strength measurements. In other areas, judging whether a change has been effective is so much harder. If you add a speech processor, for example, your received peak signal strength will remain the same. Similarly, adding a noise filter won't change the strength of incoming signals at all. Yet in both cases, and in many others where you won't see a change in signal strength, you would hope that the effort had not been in vain. Of course, in time, an improvement in your DXCC score might just hint at a beneficial change to your station but surely there's a better way of assessing changes to your line-up. What is needed is an accurate and repeatable method of assessing the intelligibility of speech.

Intuitively, we might expect that it will be blatantly obvious if a change has achieved the desired beneficial effect but experience suggests otherwise. I'm involved in the development of cave radios (see *RadCom* May 2008 and creg.org.uk for more details) that are used by the UK's volunteer rescue groups to transmit through solid rock to establish contact between the surface and an underground rescue team. These radios operate in the LF spectrum where they are plagued by interference from the LORAN-C navigation system that operates on 100kHz but has very wide sidebands. I recently conducted tests on a number of noise filters aimed at the amateur radio market in the hope that they would reduce this interference. All the filters gave the impression that the interference was reduced and that the speech was more readable. It was a big surprise, therefore, when a formalised speech intelligibility test proved that, in this particular application, the intelligibility had actually been reduced! This provided a clear indication that some sort of formal metric is required in assessing the readability of a signal.

COMMON AMATEUR METHODS. Radio amateurs already have a metric for the readability of a signal in the form of the first figure of the RS report but dare I suggest that these reports are not nearly as useful as we might hope? Scanning through your logbook you'll probably find that reports of 59 abound and, ironically, this will often be the case for contacts in which you were asked to repeat your name, QTH and report. So it really only makes sense to rely on reports from those few people who you've specifically asked to give you an honest and critical report. But even then, the R of RS is assessed subjectively, so reports will differ from person to person for an identical signal, and the score of 1-5 provides very little granularity. The SINPO reporting system used by broadcast listeners provides more information by splitting the single figure for readability into four separate scores (for interference, noise, propagation and overall) but in other respects it has exactly the same limitations as the RS report. Clearly we need to look elsewhere.

PHOTO 1: Speech intelligibility measurement allows noise filters like this one from bhi to be evaluated objectively.

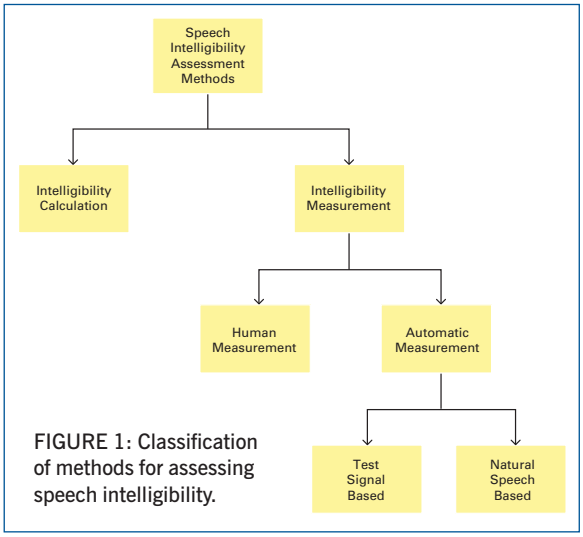

FIGURE 1: Classification of methods for assessing speech intelligibility.

METHODS OVERVIEW. Several methods for assessing the intelligibility of speech transmitted across a communication channel that are either defined in standards or are the subject of ongoing research are shown in **Figure 1**.

The first division is into those methods that rely on a calculation and those methods that involve practical measurement. ANSI standard S3.5-1997 **[1]** specifies a method for calculating the Speech Intelligibility Index (SII) from the equivalent speech spectrum level, the equivalent noise spectrum level and the equivalent hearing threshold level. SII is said to be "highly correlated with the intelligibility of speech under a variety of adverse listening conditions, such as noise, filtering, and reverberation". The method described in the standard is applicable when the various input variables to the model can be either measured or accurately estimated, which will rarely be true in the highly variable short wave bands. Since the standard also specifies that it is only applicable to systems that are approximately linear and don't include sharply filtered bands of speech or sharply filtered noise, it's safe to suggest that this is not generally applicable to amateur radio use. We shall turn our attention, therefore, to methods for measuring speech intelligibility.

Intelligibility measurement methods further divide into methods that involve some sort of human assessment, and fully automated methods. The automated methods split again into those that use test signals and those that involve the transmission of genuine speech, albeit automatically. The most widely adopted test signal method is defined in the International Standard IEC 60268-16 **[2]**. The method involves making numerous readings of the ratio of the output to the input modulation level for a matrix of octave bands and modulation frequencies. There are a number of drawbacks to the method. Distortions in the system under test may affect the measured speech intelligibility differently from real speech intelligibility. For instance, a recorded voice that is played back at a slightly higher speed is still quite intelligible, but the measured intelligibility may drop significantly.

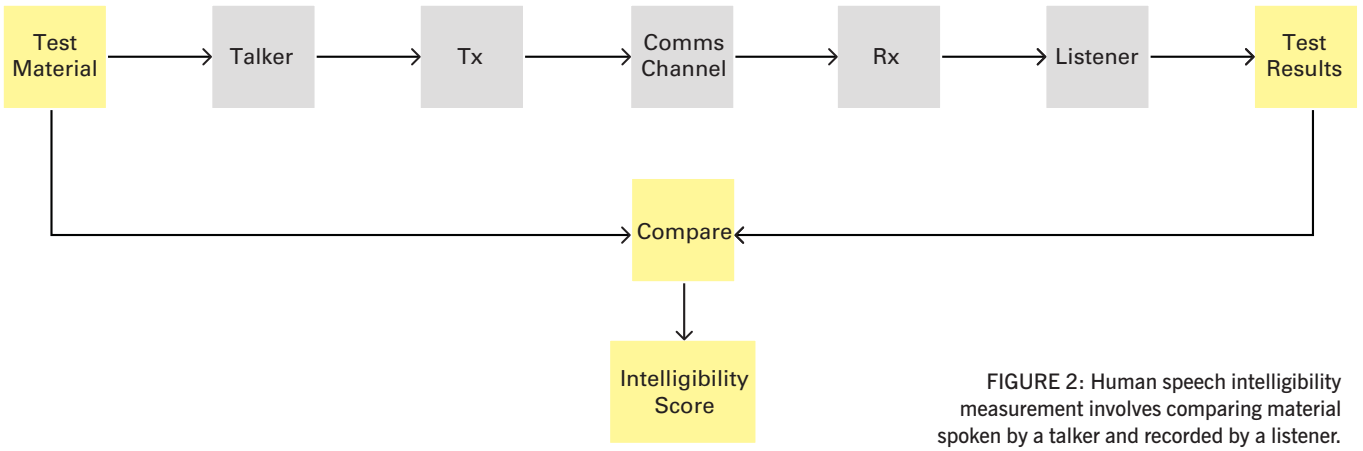

FIGURE 2: Human speech intelligibility measurement involves comparing material spoken by a talker and recorded by a listener.

Similarly centre clipping (cross-over distortion) may affect real speech intelligibility much more severely than the measured value. Because of this the standard specifies that it should not be used for transmission channels that introduce frequency shifts or frequency multiplication, or include vocoders. When we also bear in mind that expensive test equipment is required it is clear that this method does not meet our requirements.

Automated intelligibility measurement using real speech as opposed to test signals is still at the leading edge and not defined in any standard. A recent paper [3] describes work involving a neural network that has shown some potential for calculating intelligibility by analysing continuous natural speech. However, at the time the paper was published the system had only been validated by simulation. The paper also reports only that 'some success' has been achieved in providing speaker-independent measurements. In both these areas, further work is needed but in personal correspondence the author stated that, so long as appropriate samples were used in the training set, the system should be capable of handling compressed speech, something that is a limitation of some other automated methods. Although still at a very early stage, this method would be worth keeping an eye on as it could provide a means, in association with appropriate speech-based beacons, of monitoring bands with a view to providing an alert of conditions favourable to speech communication.

However, for general purpose amateur use, having discounted the alternatives, we are left with measurement methods that involve some form of human intervention. This is covered in the following section.

HUMAN MEASUREMENT. The basic premise of human speech intelligibility measurement, as illustrated in **Figure 2**, is that a talker speaks some test material that is transmitted across a communication channel to a listener who attempts to record the received speech. A measure of intelligibility is calculated by comparing the spoken test material with that recorded by the listener. There are three options, namely open word tests, pseudo-open word tests and closed word tests.

In an open word test, any word can be used and, naively, we might assume that these are preferable because the listener can't learn the words and isn't influenced by being offered suggestions. However, the major drawback is that these tests tend not to include phonetically-balanced words and the degree of difficulty will vary. Accordingly, a large number of words have to be used to get an accurate result. An alternative is the use of nonsense words, otherwise known as logotoms, which are mostly transitions between vowels (V) and consonants (C). Usually a list of VC, CV, VCV or CVC words is used, but longer words, such as CVVC, VCCV, or CCCVCCC, are sometimes used. Test words are usually symmetric, for example aka, iki, uku or kak, kik, kuk. Since references to this method of intelligibility testing are nearly all related to the evaluation of text-to-speech systems it seems reasonable to assume that they offer no real advantages for general purpose speech intelligibility measurement.

Another type of open test is the sentence-level test in which sentences are usually chosen to model the occurrence frequency of words in a particular language. At first sight it might appear that phrases or sentences would provide appropriate test material for our application. However, there are two major reasons why phrases or sentences do not usually lend themselves to general speech intelligibility testing. First, the correct understanding of the words and phrases in sentences is significantly influenced by the knowledge the listener has of the grammar, syntax, and meaning of the ideas involved. Since these factors vary from one listener to another, their contribution in a speech intelligibility test will vary. In other words, there's a significant risk that listeners will be able to guess a word from the context but this ability is not constant between different listeners. Second, it's difficult to create a sufficient number of sentences that are phonetically representative of speech in general and yet of equal difficulty or familiarity to typical listeners. Furthermore, because it's far easier to learn sentences than single words, sentences cannot be repeated with the same listeners.

Tests based on a pseudo-open word list give the listener a free choice but the words are taken from a fixed published list. The benefit of the pseudo-open test lists compared to the open word tests is that it's easier to achieve phonetic and level of difficulty balance. This balance means that valid results can be achieved with far fewer test words than would be possible with an open test. However, to a lesser or greater extent, the listener will become familiar with the words in a pseudo-open list after having conducted a number of tests thereby increasing the chance that the listener will guess a word.

Tests based on a closed word list require the listener to choose a word from a fixed set of possible answers. The reduced number of responses means that the test can be completed quickly. This technique also benefits from the fact that reliable results can be obtained with relatively small subject groups. Furthermore, since the alternatives are always presented, the results won't be affected as the listeners start to remember the words so these tests can be used on multiple occasions with the same listeners. Having carried out an extensive literature search, I would suggest that the closed word list tests are, on balance, the most appropriate for our application.

CLOSED WORD LIST TESTS. American national standard ANSI S3.2-1989 [4] defines one pseudo-open word list test, the Phonetically Balanced Word List test (PB), and two closed word list tests, the Modified Rhyme Test (MRT) and the Diagnostic Rhyme Test (DRT). The DRT provides some indication of why speech ineligibility is impaired by giving measures for phonetic characteristics such as voicing, nasality, sustention, sibilation, graveness and

compactness. However, unless this level of detail is sought, I would recommend the use of the other closed word list test, the MRT, because of its simplicity and ease of use.

The Modified Rhyme Test uses 50 six-word groups of rhyming or similar-sounding monosyllabic English words. Each word is constructed from a consonant-vowel-consonant sound sequence, and the six words in each group differ only in the initial consonant sound (eg went-sent-bent-dent-tent-rent) or final consonant sound (eg pat-pad-pan-path-pack-pass). Listeners are shown a six-word group and then asked to identify which of the six words had been spoken by the talker. Needless to say, the word is chosen at random from each group and it also helps to present the groups in a different random order each time the test is conducted. The standard dictates that a carrier sentence is used. So, for example, the talker would say "Select the word 'went' now." but without stressing the test word.

The intelligibility score for each of the tests defined in ANSI S3.2-1989 is basically the percentage of words recorded correctly but the standard specifies that, for those tests involving a closed word list, ie DRT and MRT, the score is adjusted for the probability that a certain number of items in each list may be correctly identified by chance or by guessing. The following formula is used to achieve this:

$$I = \frac{100}{T}\left(R - \frac{W}{n-1}\right)$$

where I is the Intelligibility expressed as a percentage, R is the number of words correctly identified by the listener (ie Right), W is the number of words incorrectly identified by the listener (ie Wrong), T is the number of words spoken in the test (50 for MRT), and n is the number of alternative words offered to the listener for each word in the test (6 for MRT).

PRACTICAL CONSIDERATIONS. ANSI standards are expensive to buy but you'll probably be able to find the MRT word lists on-line. At the time of writing I found them at www.meyersound.de/support/papers/speech/mrtlist.htm although there's no guarantee, of course, that they will still be there by the time you read this. **Table 1** shows a portion of the MRT word list.

The standard dictates that at least five talkers and at least five listeners should be used and that there should be at least as many talkers as listeners. The final intelligibility score is an average of the results for all combinations of talker and listener. The talkers and listeners must be representative of the expected user population. In practice, this means that talkers and listeners should be taken from a wide age range and should include both men and women.

We might assume that conducting a test is trivially simple but the standard makes it very clear that it is vitally important to train both talkers and listeners by performing dummy tests so that they become familiar with the test procedure, the word lists and so forth. Scores will vary during the training process but, once the talkers and listeners are trained, should be relatively constant for a constant signal.

The easiest way of conducting a test is for the talker to pick a word at random from each group, marking the selected word on a test sheet as it is spoken. The listeners are given similar test sheets onto which they mark the word they believe to have been spoken. An alternative, which also allows the groups to be presented in a random order, as well as randomly picking the word within each group, is to use software to generate a talker's list, which is just a list of the 50 words to be spoken, and a listener's score sheet that contains 50 sets of six alternatives ordered in the same way as the talker's list. In fact, the procedure can be further automated by having the listeners enter the selected word directly into a software utility that could also calculate the score at the end of the test. This would require a number that seeds the software's random number generator, to be exchanged between talker and listener. And finally, for the ultimate in automation, rather than use a human talker, software could use a collection of audio clips, representing the 300 possible words in the MRT, assembled into the carrier sentence, to generate speech, replacing the human talker entirely.

It's been said that you can't control what you can't measure yet many of today's developments in communication offer benefits that can't be measured in terms of signal strength alone. In this era of digitally-encoded speech, DSP-based speech processors and noise filters, we need a better way of configuring our stations for optimal performance.

Speech intelligibility measurement offers one such method and perhaps deserves more widespread application in amateur radio circles.

Speech intelligibility measurement being used to evaluate noise filters for cave radio.

TABLE 1: A Portion of the word list used for the Modified Rhyme Test

	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
Group 1	went	sent	bent	dent	tent	rent
Group 2	hold	cold	told	fold	sold	gold
Group 3	pat	pad	pan	path	pack	pass
Group 4	lane	lay	late	lake	lace	lame
Group 5	kit	bit	fit	hit	wit	sit
Group 6	must	bust	gust	rust	dust	just
Group 7	teak	team	teal	teach	tear	tease
Group 8	din	dill	dim	dig	dip	did
Group 9	bed	led	fed	red	wed	shed
Group 10	pin	sin	tin	fin	din	win

REFERENCES

[1] ANSI S3.5-1997, American National Standards Institute, "Methods for Calculation of the Speech Intelligibility Index".

[2] IEC 60268-16 (third edition, 2003-5), International Electrotechnical Commission, "Sound system equipment – Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index".

[3] Li, F. F., Cox, T. J. (2003) "Speech Transmission Index from Running Speech: A Neural Network Approach", J. Acoust. Soc. Am. 113 (4), 1999-2008.

[4] ANSI S3.2-1989, American National Standards Institute, "Method for Measuring the Intelligibility of Speech over Communication Systems".