SUBJECTIVE SPEECH INTELLIGIBILITY MEASUREMENTS AND 3D AUDIO IMPLEMENTATION

S. E. Mercy QinetiQ, Cody Technology Park, Ively Road, Farnborough, Hants, GU14 0LX
M. J. Aitchison QinetiQ, Cody Technology Park, Ively Road, Farnborough, Hants, GU14 0LX

1 INTRODUCTION

Speech intelligibility measurements can be made by a number of different objective and subjective methods[1]. Objective methods may examine the frequency spectrum of the speech signal, the frequency spectrum of the ambient noise and other associated factors. Speech Inteference Level (SIL)[2], Articulation Index[3] or more recently the Speech Intelligibility Index[4] and Speech Transmission Index (STI)[5] are examples of objective methods. Subjective methods require trained speakers to read lists of words and listeners to respond with what they thought they heard.

The UK MoD sets out the requirements for speech intelligibility of communications systems in DEF-STAN 00-25 Part 16[6]. Whilst objective methods are described, paragraph 6.1.3.11 notes that these methods are not sufficiently precise for assessment of military communications under marginal conditions. It goes on to state that certain assumptions about the nature of the speech implicit in the Articulation Index and Speech Transmission Index methods do not apply to vocoded speech, synthetic speech and possibly for speech transmitted by single sideband radio. These assumptions concern the important acoustic features that make speech understandable. The distribution and range of levels within certain frequency bands are weighted according to the amount of speech information that they contain. In marginal conditions, for example where the Articulation Index is below 0.3, and for vocoded or synthetic speech, these assumptions become invalid. As such, intelligibility tests using panels of talkers and listeners provide a more robust way to assess speech intelligibility.

The US Department of Defense in Military Standard MIL STD 1472F[7] states that Articulation Index and Speech Transmission Index should be used to estimate system performance during the concept and design stage but not as a substitute for intelligibility testing when a production system is available. Therefore, communications equipment that is to be used in a military context should undergo subjective speech intelligibility assessment.

2 SUBJECTIVE SPEECH INTELLIGIBILITY TEST TECHNIQUES

2.1 Introduction

Listeners in subjective speech intelligibility experiments are presented with voiced speech items, ranging from simple combinations of phonemes such as CVC (consonant, vowel, consonant) to sentences or short conversations. The test response may be a choice from a discrete number of alternatives (closed-set), or the listener may respond with any item that they believe was presented (open-set).

The speech items presented within a subjective assessment must have been chosen to provide a representative sample of critical speech sounds. The context within which the speech items are presented is also important, since listener expectations and familiarity with the subject matter can influence the results. Three well established speech intelligibility test methods are: the Phonetically Balanced word lists, the Diagnostic Rhyme Test and the Modified Rhyme Test. These all provide assessment of specific transmission features of the communications system under test.

2.2 Phonetically Balanced word lists

The Phonetically Balanced (PB) word lists were developed during World War II and were based on earlier work on radio and telephone communications. There are 20 lists of 50 monosyllabic English words. Each list is phonetically balanced so that the phonetic components of the English language as a whole are represented. Before each list presentation the word orders are randomised. Each word is presented within a carrier sentence such as 'Would you write ___ now', with no special emphasis on the word and with the vocal effort measured over the whole phrase. Since the words are written down by the listeners spelling is not crucial and scores should be based upon a phonetically correct response (homophones such as 'coarse' and 'course' will score the same).

The embedding of the test word in a carrier phrase enables the talker to control his or her vocal effort and to get the attention of the listener at each utterance. Certain types of communication system also require speech in order for any given channel to remain open and so the carrier phrase can ensure that the test word is not truncated or otherwise affected by the normal operation of the system.

Large scale open format tests, such as PB testing, require a long training time for the listeners. This may be up to 12 hours in order to present all possible 1000 words and ensure that the listeners have reached a 'plateau' in their scores[8]. A recording of all 20 PB word lists will produce approximately 70 minutes worth of source material.

2.3 The Rhyme Tests

The initial Rhyme Test developed by Fairbanks in 1958 presents the listener with a test word and instructs the listener to complete the spelling of the word on their response sheet (e.g. '_ot', '_ay')[9]. Although this limits the listener responses to rhymes of the test word, it is not a closed response test.

Developments of the Rhyme Test have led to the Modified Rhyme Test, where the listener chooses between six rhyming alternatives, including the target word. The choices may differ in their initial consonant or in their final consonant, e.g. kick, lick, sick, pick, wick and tick or kick, king, kid, kit, kin and kill[10].

The Diagnostic Rhyme Test further reduces the closed response set to a choice of two rhyming words that differ only in their initial consonant[11]. The DRT has been designed to allow further analysis of the responses to determine whether phonemic features are correctly understood by the listener. These features are: voicing (e.g. veal-feel), sibilation (e.g. sing-thing), nasality (e.g. moan-bone), sustention (e.g. sheet-cheat), graveness (e.g. weed-reed) and compactness (e.g. yield-wield).

Since both the MRT and the DRT present alternatives to the listener for each utterance, the results must be corrected for guessing. The formula to achieve this is:

$$P_c = \frac{P_r - (P_w / n - 1)}{T} 100$$

where:
Pc = the percentage correct, adjusted for guessing

Pr = the actual number of right responses
Pw = the actual number of wrong responses
n = the number of choices
T = the total number of test items

The test word in the Rhyme Tests is presented alone, not in a carrier sentence. This requires that the talker has been sufficiently trained so as to maintain a consistent vocal effort. These closed-set tests only require the listeners to be trained for approximately 5 minutes[8]. In practice, the listeners are trained in the test techniques and by listening to the test material in the clear, which takes longer than the recommended 5 minutes.

3 IMPLEMENTATION OF SUBJECTIVE TESTING

Ideally, speech intelligibility testing should be performed in the environment where the communications system is to be used, using talkers and listeners representative of the final operation. However, this can be time consuming and costly, and access to the facility housing the communications system may be restricted. In these circumstances, the use of recordings of talkers made over the communications system and reproduced in a representative environment can be the most practical test method. Similarly, the use of live talkers that have undergone sufficient training for speech intelligibility testing is clearly a limiting factor on the assessment. By using high quality recordings of talkers, the assessments can be performed repeatedly over the communications system under a number of different conditions, as required by the assessment.

Together with the Institute of Sound and Vibration (ISVR), QinetiQ has developed a library of talker recordings. Typically, the talker was located in a quiet room, such as an anechoic chamber, and the recordings made using Brüel and Kjær (B&K) instrumentation microphones. The talker used a vocal effort meter to maintain a constant vocal effort. Different microphones, such as mask microphones, have also been used in recordings.

Figure 1 QinetiQ’s Speech Intelligibility Facility

These recordings can then be passed over a communications system, using a B&K Head and Torso Simulator where appropriate, and recorded at the listener position. The noise environment at the talker location must also be representative of the final use of the communications system. This re-recorded talker material can then be played to the listeners. The listeners should also be seated in a noise environment as close as possible to that expected in day-to-day use of the communications system and, if relevant, wearing the appropriate headset or helmet.

QinetiQ has developed a facility whereby up to 12 listeners can be presented with the same speech intelligibility test material at the same time (Figure 1). The facility incorporates a noise reproduction system in order to generate the relevant background noise. This may include noise environments from speech babble, as may be encountered in a call centre, to helicopter or fast-jet noise.

Since the DRT requires a choice between two alternatives, the listeners use a response-box with two switches with which to indicate their answer. The MRT requires a choice between six alternatives and for this test the listeners use a mouse to ‘click’ on their chosen answer. If a free response is allowed, such as a PB test, then the listeners are provided with a keyboard and enter their answers into a text box. The QinetiQ facility is preconfigured for immediate use.

4 BRIEF DESCRIPTION OF 3D AUDIO

In recent years technology has been developed capable of presenting audio signals such that they appear to emanate from unique positions in auditory space. This technology is commonly known as 3D audio. One use of 3D audio is to present multiple audio signals over a stereo headset simultaneously each appearing to come from a distinct a direction, external to the head, i.e. the sound is located in three-dimensional (3D) space.

Humans can process directionality of audio signals because we have two ears. The signal reaching each ear is slightly different, allowing the brain to determine the direction of the originating source. For an audio source to the left of a listener, the signal will reach the left ear before it reaches the right ear. Further more the signal reaching the right ear will be attenuated due to acoustic shadowing by the listeners head. These two factors are known as the Inter-aural Time Difference (ITD) and Inter-aural Intensity Difference (IID). Acoustic signals also undergo spectral shaping, dependent on the direction of the source relative to the listener.

In order to make an acoustic signal appear to emanate from a given direction, these factors must be emulated. This is done using digital signal processing and is based around a filter set known as the Head Related Transfer Function (HRTF). Essentially this consists of a ‘left ear’ filter and a ‘right ear’ filter for each position that a sound may be presented from. The signal processor simply finds the two relevant filters and presents the ‘right ear’ filtered version to the right ear and the ‘left ear’ version to the left ear. Headphones are used to keep the left and right ear signals distinct (avoiding cross-talk).

Each individual listener is distinct, having different head, ear and shoulder shapes and sizes. Consequently, the best effect is achieved using individual HRTFs measured for each specific listener. A good effect can still be achieved with a generic HRTF and, given the cost of measuring individual HRTFs, generic HRTFs generally provide a satisfactory solution.

4.1 3D Audio Applications

There are many audio tasks that can be aided with the use of 3D audio, but they all fall into one of two main configurations, non-head tracked and head tracked.

The non-head tracked configuration is the simplest implementation of 3D audio. A number of audio signals are presented at spatially distinct positions, fixed relative to the listener’s head. For example, an audio signal presented to the listeners left will always emanate from the left headphone, regardless of how the listener moves their head. The main application of non-head tracked 3D audio is often referred to as “Comms Splitting”. Speech intelligibility testing such as that discussed in this paper has shown that comms splitting can significantly improve intelligibility of multiple communications signals presented simultaneously[12, 13, 14, 15].

A head-tracked configuration extends the localisation capability, using a head tracking system to monitor the movements of the listener's head and to fix the presentation positions of the audio signals in space. Consequently, a listener can move their head to turn and face each audio signal, unlike the non-head tracked configuration. The ability to fix audio signals in space allows inherent positional information to be conveyed to the listener. Consequently, audio signals can be presented at positions relative to their meaning. Head tracking can also be used to augment the comms splitting application, presenting each comms channel from the direction of the source. For example, in the military cockpit comms sources may come from air traffic control, ground forces, other aircraft and so on, each presented from their respective 'real world' direction.

4.2 Comms splitting using 3D audio

Comms splitting uses the natural human ability to distinguish between voices originating from different directions relative to the listener. We generally use this ability without being aware of it. One popular example is referred to as the 'Cocktail Party Effect'. When in a room with many people talking concurrently we are able to focus on an individual because their voice emanates from a distinct point in space. Most of us have experienced our attention being diverted to someone else's conversation when we hear our name mentioned. This shift in focus does not require the listener to physically turn or move, their brain uses the fact that the voices come from different directions.

Currently most communications systems are based around monaural presentation systems. This means that the same signal is presented to both ears. As such, signals all appear to emanate from the same point in space. Further more, that point is perceived as being inside the listener's head. By using 3D audio to present each channel of communications a different direction, the listener can distinguish between each one more effectively. With this ability should come increased intelligibility.

5 TESTING SPEECH INTELLIGIBILITY OF COMMS SPLITTING

The intelligibility test method was based on the PB word list test as defined in ANSI S3.2–1989. This was adapted to assess the effect of 3D audio presentation on speech intelligibility. The emphasis was on determining intelligibility improvements as opposed to specifying the system to a standard. As such, the test method used a reduced number of lists, fewer talkers and less training time than specified in the standard. Lists six, nine, ten and eleven were used as they contained the fewest Americanisms. These adaptations were essentially due to time and cost limitations.

Each word from each of the four PB lists was embedded in a carrier sentence. The carrier sentence included a keyword combination immediately preceding the word that the subjects were required to identify. An example phrase is: "Baron Authentication Goose Over" where the test word is Goose. These carrier sentences were then embedded into a background noise track.

Two independent trials have been conducted using QinetiQ's dedicated speech intelligibility test facility, both using the same adapted PB test method.

5.1 Trial 1

The first trial[15] considered the variations in speech intelligibility when four channels of simultaneous comms were presented using mono, stereo and non-head tracked 3D audio systems. A fourth condition used head-tracked 3D audio, but this test could only be conducted on one subject at a time as only one head-tracker was available. Each of the four channels of comms was presented from a distinct spatial position. For each PB list one channel had the words, embedded in carrier sentences, and the other three channels had a mixture of real life comms recordings from sources such as air traffic control and helicopter sorties. Consequently, for a given word list, the words always appeared from the same direction. The channels were randomised over the whole trial such that the PB word lists were presented from each of the four positions used.

The use of real life recordings on the three background channels allowed the tests to reflect the speech intelligibility in a realistic comms environment. Whilst this provides a measure of the intelligibility in normal use, the result for a given test word is dependent on the number of channels presenting comms traffic at the instant that the word is presented. In this set up there may have been times when as few as one or as many as three of the channels had comms traffic when the test word was presented. Consequently, this method does not give an absolute measure of the intelligibility of four simultaneously presented comms channels.

The aim of the first trial was to prove the concept of advanced audio presentation methods, including comms splitting. Furthermore, the systems tested were not being assessed to meet a specified standard, such as that defined by ANSI S3.2-1989. The results achieved are indicative of the system intelligibility likely to be afforded in day-to-day use, a useful indicator when proving a concept.

5.2 Trial 2

The second trial[16] investigated the optimum number of simultaneous channels of comms that may be presented using 3D audio. Four, six and eight channels of simultaneous comms were compared using both mono and 3D audio presentation. The test method remained the same as for the first trial, but the various channels of comms consisted of continuous comms speech. Therefore, whenever a test word was presented there was also comms traffic on all of the other channels. The PB words were also randomised across the channels so the words within a single list could appear on any one of the channels.

Having proven the concept of comms splitting in the first trial, this trial aimed to determine an optimum number of channels that could be presented using the technology. Consequently the results needed to reflect the absolute intelligibility for each of four, six and eight channels of simultaneous comms traffic even though it is not truly representative of real life scenarios. Hence, some of the trial runs resulted in an extremely intense comms environment, e.g. when eight channels of comms were presented simultaneously with the test words embedded within any one of the channels.

As with the first trial, the aim was not to assess the system to meet a specification, but to aid in research and development. The results for this trial provide a measure of the speech intelligibility for presentation of four, six and eight channels of simultaneous comms using mono and 3D audio presentation systems. The trial essentially addressed the ‘worst case’ scenario and a more realistic set up, such as that of the first trial, would be expected to yield higher intelligibility levels as a results of there being less simultaneous comms traffic.

6 RESULTS and CONCLUSIONS

In both trials where the modified test method has been employed, the results were conclusive and confirmed experience. Each of the two trials had a different aim and as such used the intelligibility test method to slightly different ends.

In the first case the results showed that both 3D audio and stereo systems provided significant intelligibility improvements over mono presentation. The intelligibility of each presentation method was a measure of the real life performance that could be expected.

The second trial was more intense, with listeners trying to distinguish between up to eight simultaneous comms channels. Again, the results confirmed the experience of the subjects and common sense. As the number of simultaneous channels increased, the intelligibility decreased. This was the case for both mono and 3D audio systems (this trial did not include a stereo system). Furthermore, 3D audio was shown to provide increased intelligibility in all test conditions.

Whilst the ANSI S3.2-1989 standard was the basis for the testing performed, the modifications were as a result of the test aims. The standard provides a unified method for comparing the intelligibility of a speech or communication system. QinetiQ have taken this test method and adapted it to give a comparative assessment of new technology. Used in a research and development context, the test method becomes less about conforming to a specified standard and more about investigating how different systems can be used to improve operator effectiveness.

Obviously, the test method could be employed to assess whether a production system meets a specified speech intelligibility requirement. The test modifications used to aid in the research and development of advanced presentation techniques are similar to those that would be necessary to test a system in accordance with the ANSI standard.

References

1 ‘Ergonomics – Assessment of speech communication’, BS EN ISO 9921:2003, (2003)

2 ‘Ergonomic assessment of speech communication – Part 1: Speech interference level and communication distances for person with normal hearing capacity in direct communication (SIL method)’, BS ISO 9921-1:1996, (1996)

3 ‘Methods for the calculation of the articulation index’, American National Standard ANSI S3.5-1969 (1969)

4 ‘Methods for the Calculation of the Speech Intelligibility Index’, American National Standard ANSI S3.5-1997 (1997)

5 ‘Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index’, BS EN 60268-16:1998 (1998)

6 ‘Human Factors for Designers of Systems - Part 16: Introduction and Manpower Domain’, United Kingdom Ministry of Defence, Defence Standard 00-25, Issue 1, (July 2004)

7 ‘Human Engineering’, United States Department of Defense, Design Criteria Standard, MIL-STD-1472F, (August 1999)

8 ‘Acoustics – The construction and calibration of speech intelligibility tests’, ISO TR 4870 (1991)

9 G. Fairbanks, ’Test of Phonemic Differentiation: The Rhyme Test’, J. Acoust. Soc. Am. 30 596-600 (1958)

10 A. S. House, C. E. Williams, M. H. L. Hecker and K. D. Kryter, ‘Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set’, J. Acoust. Soc. Am. 37 158-166 (1965)

11 W. D. Voiers, Evaluating Processed Speech using the Diagnostic Rhyme Test, Speech Technology, (Jan 1983)

12 D. R. Begault, ‘3D Sound for Virtual Reality and Multimedia’, Academic Press Ltd, London, (1994).

13 W. T. Nelson, R. S. Bolia, M. A. Ericson, R. L. McKinley, ‘Monitoring the simultaneous presentation of multiple spatialized speech signals in the free field’, (March 1998).

14 E. C. Haas, D. C. Wightman, ‘3D audio displays in army helicopter systems’, (March 1998).

15 M Aitchison, ‘Methods for improving speech intelligibility of multiple communications channels’, QinetiQ report, (January 2001).

16 M. Aitchison, ‘Investigating parameters to improve the localisation accuracy of 3D audio signals’, QinetiQ report, (March 2006).