Proceedings of The Institute of Acoustics

THE ASSESSMENT OF SPEECH INTELLIGIBILITY AT RSRE

R.L. Pratt

Royal Signals and Radar Establishment, Malvern

INTRODUCTION

The Royal Signals and Radar Establishment has a particular interest in quantifying the performance of voice communications systems, and is developing a permanent facility for conducting speech intelligibility tests using the Diagnostic Rhyme Test (DRT). Research is also being conducted on the correlation between DRT score and user acceptability.

ACOUSTIC FACILITIES

In order to carry out acoustics research a special-purpose building was opened in 1979 comprising an anechoic chamber, a pair of high noise rooms, an audiometric room and a conversational laboratory.

The anechoic chamber is a 5m cube with wedges 1.5m long and can operate down to about 70Hz. The background noise has been kept to a very low level by careful isolation of the chamber from the main building. The chamber is used mainly for measuring the near (10mm) and far (1m) field response of noise-cancelling microphones.

The high noise rooms are hard-walled, irregularly shaped rooms 50 cubic metres in volume, designed to produce a diffuse sound field in which sound pressure levels of up to 120 dB may be produced. The rooms are used to prepare speech material and conduct listening experiments in noise environments representing a variety of different military platforms, such as tanks, fast jets and rotary-wing aircraft.

The audiometric room is used to monitor the hearing of subjects who participate in such experiments.

The conversational laboratory comprises two acoustically isolated booths whereby a pair of subjects may conduct a dialogue using a simulation of the communications channel under study, and then rate the performance of the channel using a questionnaire.

THE ASSESSMENT OF SPEECH INTELLIGIBILITY AT RSRE

INTELLIGIBILITY TESTING

The assessment of speech intelligibility has its origins in the early part of this century with the use of sentence material for investigating the performance of telephone circuits. However, the ability of subjects to exploit contextual information led to the development of isolated word intelligibility tests. This work culminated in the Phonetically Balanced (PB) word lists published by Egan [1]. Such tests, where subjects actually write down the words they thought they heard, are time-consuming to administer and score; moreover they also require a high level of subject training.

To overcome these limitations, interest was focused on techniques using rhyming speech material. In these tests subjects chose the word they thought they heard from a closed set of rhyming alternatives. Of this latter type, the test that has proved the most useful for military applications is the Diagnostic Rhyme Test, of Voiers [2].

The DRT vocabulary consists of 116 rhyming word pairs whose initial consonant only is varied. The complete vocabulary of 232 words is recorded by a number of talkers using microphones and background noises that are appropriate to a particular operational environment. The word lists are subsequently replayed to a panel of listening subjects via the appropriate communication link. At the same time a micro-computer presents the word pairs to the subject visually, and logs his response.

The six features, voicing, nasality, sustension, sibilence, graveness and compactness are tested and scored independantly, thus providing diagnostic information which can be related to engineering design. For example, the effectiveness of pitch extraction algorithms may be tested by examining scores for the voicing feature. The overall result of a test is expressed as the percentage of the vocabulary correctly identified by the listeners.

The use of a standard set of both talkers and listeners enables comparisons to be made between a wide variety of speech channels.

ASSESSING ACCEPTABILITY

In order to place the results of such intelligibility tests in context, the acoustic environment of a military platform may be created, and operational users invited to participate in experiments which include an attempt to simulate their normal workload, thus enabling them to rate acceptability under

realistic conditions. By conducting DRT experiments under the same acoustic conditions but using our standard talker and listener subjects, the relationship between intelligibility scores and acceptability may be investigated for different classes of user.

This relationship is unlikely to be a simple one, since the user's concept of "acceptability" is generally multi-dimensional. However, in the military context, where users are working in high background noise levels, the dimension of intelligibility tends to dominate this judgement. Thus one would expect good, but not perfect, correlation between DRT score and the user's acceptability rating in these circumstances. Studies of this nature are essential in order that the results from intelligibility experiments may be sensibly interpreted.

RESULTS

The DRT has been used by the US Dept. of Defence (DoD) for nearly 20 years. As a result of numerous studies, a set of categories for various DRT scores has been established for a "typical" user, and this is given in Figure 1.

For aircraft communications in particular, a recent report by Tierney and Schecter [3] showed that a channel yielding a DRT score of 71% was given a rating of 68 on the scale Unacceptable(0)-Acceptable(50)-Excellent(100) by operational aircrew, suggesting that the borderline for acceptability is somewhat below the figure of 70% quoted in Figure 1. It should be noted that this study did not attempt to simulate workload.

The results of an intelligibility experiment that we have conducted are given in Figure 2. The DRT lists were made by 4 talkers using an aircrew oxygen mask microphone processed by a 2.4 kbit Linear Predictive Coder (LPC-10). The broad band noise source was adjusted to give Speech-to-Noise (SNR) ratios of 10 and 20 dB, and a third recording in quiet conditions was also included. The figures shown are the average of 5 listeners.

Thus for a given DRT criterion (say 70%) the equivalent SNR may be estimated, and by recording aircrew speech in-flight, the proportion of exchanges above and below this value may be determined.

REFERENCES

[1] J.P.Egan, 'Articulation testing methods', Laryngoscope, Vol 58, 955-991, (1948).

THE ASSESSMENT OF SPEECH INTELLIGIBILITY AT RSRE

[2] W.D.Voiers, 'Diagnostic evaluation of speech intelligibility', Benchmark papers in acoustics, Vol 11, Speech intelligibility and speaker recognition (M. Hawley, ed) Dowden, Hutchinson and Ross, Stroudsburg, (1977).

[3] J.Tierney and H.Schecter, 'The Lincoln Laboratory-Aerospace Medical Research Laboratory digital speech test facility', Massachusetts Institute of Technology Lincoln Laboratory, Technical Report Number 683, (1984).

THE ASSESSMENT OF SPEECH INTELLIGIBILITY AT RSRE

DRT Score	Category	Examples	Qualifiers for these examples
100	Excellent	Unfiltered speech Speech low-pass filtered at 4 kHz	Speech from a quiet environment; no significant distortions; high-quality microphone
96	Very Good	CVSD at 32 K bps CVSD at 16 K bps	Error rate less than 1%; speech from a quiet environment
91	Good	Typical commercial telephony within continental USA APC Processor at 9600 bps LPC-10 Vocoder at 2400 bps, no bit errors	Speech from a quiet environment
87	Moderate	LPC-10 Vocoder with bit error protection, at 2400 bps with 2% random bit errors	Speech from a quiet environment
83	Fair	LPC-10 Vocoder without bit error protection, at 2400 bps with 2% random bit errors	Speech from a quiet environment
79	Poor	LPC-10 Vocoder with bit error protection, at 2400 bps with 5% random bit errors	Speech from a quiet environment
75	Very Poor	Experimental 800 bps voice processor with no bit errors	Speech from a quiet environment
70	Unacceptable	LPC-10 Vocoder at 2400 bps	Speech from a helicopter noise environment

Figure 1. The relationship between DRT scores and categories of voice quality.

SPEECH-TO-NOISE RATIO (dB) - MASK MICROPHONE

	10	20	QUIET	LISTENING CONDITIONS
TEST 1	55	72	77	Quiet
TEST 2 (Replication)	59	73	78	Quiet

Figure 2. DRT scores (%) based on 4 talkers and 5 listeners.