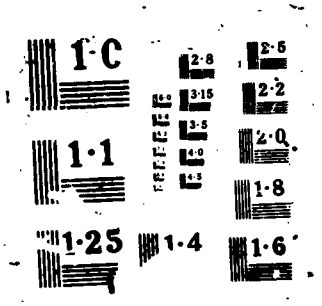
AD-A193 653 ASSESSING THE INTELLIGIBILITY AND ACCEPTABILITY OF VOICE COMMUNICATIONS S..(U) ROYAL SIGNALS AND RADAR ESTABLISHMENT MALVERN (ENGLAND) R L PRATT ET AL. 1/1

UNCLASSIFIED JUN 87 RSRE-87003 DRIC-BR-105064 F/G 25/4 NL

END
DATE
FILMED
8 88


1·C
1·1
1·25
2·8
3·15
3·5
4·0
1·4
2·5
2·2
2·0
1·8
1·6

AD-A193 653

UNLIMITED

ROYAL SIGNALS AND RADAR ESTABLISHMENT

REPORT No 87003

TITLE: ASSESSING THE INTELLIGIBILITY AND ACCEPTABILITY OF VOICE COMMUNICATION SYSTEMS

AUTHORS: R L Pratt, I H Flindell* and A J Belyavin**

DATE: June 1987

ABSTRACT

A facility for quantifying the speech intelligibility of voice communication systems using the Diagnostic Rhyme Test has operated continuously at the Acoustics Laboratory of the Royal Signals and Radar Establishment since February 1985.

User acceptability trials that enable Service personnel to operate, and then assess, voice communication systems under simulated operational conditions have also been conducted.

This report describes the procedures used to assess both intelligibility and acceptability, and presents the results of studies investigating the use of digital vocoders in high noise environments.

An Executive Summary is provided to give project offices and others responsible for designing, specifying or procuring voice communications systems (and components) an indication of the services that are available.

* Institute of Sound and Vibration Research, Southampton University

** Royal Air Force Institute of Aviation Medicine, Farnborough

Copyright
C
Controller HMSO London
1987

UNLIMITED

RSRE REPORT 87003

ASSESSING THE INTELLIGIBILITY AND ACCEPTABILITY OF VOICE COMMUNICATION SYSTEMS

R L Pratt, I H Flindell and A J Belyavin

CONTENTS

1. Executive Summary

2. Introduction

3. The Diagnostic Rhyme Test

4. Characterising the Communication System

5. Characterising the Acoustic Environment

6. The Measurement of Speech-to-Noise Ratio

7. Creating the Speech List Recordings

8. Administration of the Diagnostic Rhyme Test

9. The Statistical Analysis of Diagnostic Rhyme Test Data

10. An Example Analysis of Variance

11. Strategies for Conducting Diagnostic Rhyme Tests

12. Comparison with US Results

13. A Guide to the Conduct of the Diagnostic Rhyme Test

14. Acceptability Assessment Experiments

15. Acceptability Assessment Results

16. The Relationship between Intelligibility and Acceptability

17. Summary

18. Recommendations

19. Acknowledgements

20. References

21. Figures

22. Annex

23. Appendix to Annex

QUALITY INSPECTED 4

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

FIGURES

1. The Diagnostic Rhyme Test Vocabulary.

2. Listening subjects participating in Diagnostic Rhyme Tests.

3. Principal components of a generalised speech transmission link.

4. One-third Octave analyses of in-flight recordings taken from the mask microphone of a fast jet pilot. For frequencies below 1.5kHz the "noise only" analysis is typically 20dB below that of the "speech and noise" case. Thus the contribution made by the noise to the overall level is negligible, with all the energy present being due entirely to the speech signal. In the 2-4kHz range it is the noise that dominates and the speech-to-noise ratio is negative in all but one of the bands.

5. Speech-to-noise ratios produced by fast jet aircrew derived from in-flight recordings. Individual aircrew are an important source of variability. Pilot 1 shows an 8dB range in speech-to-noise ratio (from 12-20dB) for the same flight condition. Navigator 3, however, achieves a speech-to-noise ratio that is independent of flight conditions.

6. Analysis of Variance and basic statistics for a single Diagnostic Rhyme Test.

7. Analysis of Variance results for an experiment to assess the influence of live and mixed speech material on Diagnostic Rhyme Test scores.

8. Analysis of Variance interactions.

   a) The influence of individual talker on Diagnostic Rhyme Test Moves for the three voice coders.

   b) Live and Mixed speech recordings analysed by individual talker.

   c) The influence of live and mixed speech material on Diagnostic Rhyme Test scores for the three voice coders.

9. Summary of Diagnostic Rhyme Test Scores. When Listeners only are treated as a random effect a difference in scores between codes of 1.1% is significant ($p < 0.05$). When both talkers and listeners are treated as random effects this value increases to 3.3%. When making comparisons between live and mixed speech the figures are 0.9% and 6.5% respectively. The greater loss in resolution for the live vs mixed comparison is due to the larger T x S interaction term in Figure 7.

10. The questionnaire used for the user acceptability experiment. The solid triangle represent the average of 12 responses (see section 15).

11. Pilot and Fighter Controller participating in user acceptability experiments.

12. The relationship between DRT scores and categories of voice quality (taken from reference 10).

1. EXECUTIVE SUMMARY

The current proliferation of advanced signal processing techniques in voice communication systems has generated a requirement for sophisticated methods of performance evaluation. The Diagnostic Rhyme Test (DRT) speech intelligibility test facility at the RSRE Malvern meets this need. The test is robust and capable of fine discrimination between competing systems. The test provides useful information at the system design stage and is an extremely valuable supplement to operational trials.

Speech intelligibility tests, such as the DRT, enable quantitative assessments of voice communication systems to be conducted. The degree of intelligibility required for voice communications to be judged as "acceptable" for any given application will depend on a variety of factors such as the personnel using the system, message set employed and workload.

The DRT is not a substitute for user acceptability assessments that enable Service personnel to operate, and then assess, voice communication systems under (simulated) operational conditions (see section 14). The DRT is most useful when comparing the performance of several different systems, or system components such as microphones or earphones. Once an acceptable intelligibility level has been established, DRT scores may then be used for bench-marking purposes in order to perform a quantitative check on the intelligibility of a candidate system. This process must be done with due care however, due to the variability of both the talkers and listeners who participate in the tests. The methods used to produce the speech material also influence the final score (section 7).

The recommended procedure for a voice communication system evaluation comprises six stages:

1.	Define communication system parameters.	Section 4
2.	Define talker and listener acoustic environments.	Section 6
3.	Conduct comparative DRT evaluations.	Section 8
4.	Conduct acceptability rating trials.	Section 14
5.	Determine bench-mark criterion.	Section 16
6.	Evaluate proposed systems against criterion.	Section 13

The test facility at Malvern is operated by contractors, currently the Institute of Sound and Vibration Research at the University of Southampton. A contract managed by DA/Radio covers stages 3 and 6 and can be extended to cover stage 2 if required. Stages 1 and 5 are clearly the responsibility of the project office, but the RSRE and the contractors can offer advice on the basis of previous experience. The RSRE has conducted acceptability experiments (stage 4) and can advise as necessary.

A minimum DRT score may also be set on the basis of user acceptability experiments. Any future system for consideration in that application could be bench-marked against the criterion. A minimum experiment of 10 Talkers x 2 Replications should be used for this purpose.

When tests are conducted in order to compare equipments (rather than to bench-mark them against absolute score) each DRT condition may be tested using either 5 Talkers and 2 Replications or 10 Talkers and 1 Replication, costing £300 at current rates. Typically 7 or 8 conditions are tested each week, but other arrangements are possible, subject to negotiation. To take an example, consider a comparative evaluation of 4 voice coders from different manufacturers. It would be sensible to investigate performance over a range of talker and listener acoustic operating conditions by using three talker speech to noise ratios and three levels of background noise. The complete investigation would comprise 36 (4 x 3 x 3) DRT conditions at a cost of £10,800 and would take approximately 5 weeks under existing arrangements. Additional work required at stage 2 above would be charged appropriately. The resulting statistical analysis would enable

detailed comparisons to be made between different equipments under a precisely controlled range of operating conditions.

As a result of the experience gained from over 2 years of continuous operation of this facility, it has become clear that the interpretation of DRT scores must be exercised with extreme care. There are a number of reasons for this. It has already been noted that the talkers and listeners used for testing purposes introduce variability that needs to be handled using appropriate statistical techniques. The method used to create the talker word lists also influences the resultant DRT score in a non-trivial way. But, most important of all, when the DRT scores for a particular evaluation are produced they need to be translated into terms that are meaningful to the user for whom a particular voice communication system is intended. Without such a translation the scores, by themselves, are virtually meaningless.

The method advanced in this report to effect such a translation is described in sections 14-16. First the acoustic environment is characterised. In an earlier experiment, the use of 2.4 kbits/s Linear Predictive Coders (LPC-10) to communicate to fast jet aircraft was considered. By analysing speech recordings taken in-flight it was found that aircrew achieved speech-to-noise ratios in the range 12-26dB. A series of DRT experiments was then performed using test material created with speech-to-noise ratios of 10dB, 20dB and in quiet, with no noise present. The corresponding DRT scores for the air-to-ground test, where listeners wore ground controller's headsets and were subjected to a representative background noise, were 59%, 73% and 78% respectively.

A separate user acceptaibility experiment was then performed. This involves Service personnel communicating using the proposed communication system and rating subjectively various aspects of system performance including acceptability. In order to introduce a degree of realism for this particular application, a flight simulator computer program running on a BBC home computer was coupled to a fighter controller program running on a second computer. The pilot subject, wearing the RAF Mk IV helmet and oxygen mask, was placed in one of the Acoustic Laboratory's high noise chambers with the fast jet's acoustic cockpit noise environment accurately reproduced. The fighter controller subject was placed in a separate room and wore an Astrolite ground controller's headset. In order to complete a simulated air intercept mission pilot and controller needed to exchange accurate and timely information, as is the case in real life. Such missions were "flown" with the cockpit noise adjusted such that pilot speech-to-noise ratios covered the same range as those taken from the operational inflight recordings, and subjects were subsequently debriefed to obtain their views on the acceptability of the communications system.

Only a limited number of such user acceptability trials have been conducted to date, but the results show that an aircrew speech-to-noise ratio of 15dB or greater is likely to be rated as acceptable or better, for air-to-ground communications. On this basis an interpolated DRT score of 65% has been chosen as the acceptability criterion for the use of LPC-10 in airborne communication systems. This figure is of course only valid for the particular talker word lists used in this investigation. These same lists must therefore be specified if any future tests are to be compared directly with this result.

Even this figure of 65% cannot be applied universally. Linear Predictive Coders represent just one category of degraded speech. For example, additive noise can be added to undistorted speech word lists until the same resultant DRT of 65% is produced. Although of equal intelligibility to the previous case, the subjective response evoked may not be the same. Acceptability experiments must always be conducted with the appropriate type of degradation that will occur under operational conditions.

One final point on the interpretation of DRT scores concerns the comparison with other published results notably from the US. A small number of word lists have been exchanged with Dynastat Inc., who conduct tests on behalf of the US DoD. The results

of this exchange have shown that, when tested with the same speech material, both UK and US listener panels produce very similar DRT scores. A simple comparison of UK and US results is not generally possible as so many variables (talker, microphone, background noise) will be different.

From the above it is clear that the introduction of criteria enabling voice communication system performance to be quantified based on DRT score requires an understanding of the many factors that influence speech intelligibility. The remainder of this report describes in detail the assessment methods used and their soures of variability, thereby permitting the inclusion of sensible performance criteria in future procurement specifications.

Speech intelligibility and acceptability assessments are, however, now mature techniques suitable for providing a quantitative measure of voice communication system performance. Their incorporation into procurement specifications is strongly urged; the RSRE is available to discuss particular applications and advise on appropriate criteria.

2. INTRODUCTION

Voice communications over distance, or in high noise levels, employ transducers to convert acoustic speech signals into a form suitable for transmission and subsequent reproduction at the listener's ear. In addition to the effects of noise and reverberation, both at the talker and the listener position, there will always be physical limitations present on transducer and transmission channel performance. These performance limitations and noise effects degrade the voice communication achievable in the ideal case of a direct conversation in a quiet, acoustically dead environment. The voice communication system designer has to select transducer and transmission channel parameters to optimise system effectiveness against cost, reliability, size, weight, convenience and other constraints. This requires a precise and accurate means for evaluating system effectiveness both in the design stages and in any eventual procurement process.

Voice communication system information transfer can broadly be classified under intelligibility, talker identification and emotional content. Whereas a perfect system will transfer the maximum amount of information under each classification above, any practical system will introduce some degradation of one or more of these types of information. It is often possible to maximise message intelligibility at the expense of other factors such as talker identification and emotional content, or vice versa. It is therefore important to have a clear understanding of the performance priorities for any given application. Military and aerospace communication systems generally have message intelligibility as a priority and will often be required to operate at the limits of available technology with a restricted set of possible messages. This requires system effectiveness to be evaluated in terms of message intelligibility in such a way that is not sensitive to particular message context.

There are a number of physical measurements which can be made on any voice communication system in order to describe its performance. These include frequency response linearity and bandwidth, speech-to-noise ratios as a function of frequency and various linear and non-linear waveform distortions. Unfortunately no combination of these types of measurements has yet been discovered which will adequately predict message intelligibility, except for very simple systems. This is because of the inherent redundancy in speech and the ability of the average listener to extract intelligibility information throughout the range of a typical speech spectrum. Faced with this difficulty, the system designer is forced to use actual voice communication for system evaluation, often on an ad hoc basis during system development and later in the form of a standardised test for system comparison. The standardised tests necessarily involve human talkers and listeners (as opposed to signal generators and frequency analysers) but are made as repeatable as possible through careful design and rigorous experimental control. The international telecommunications community has standardised on the use of subjective listener effort scales for discrimination between voice communication systems with high message intelligibility offering more subtle possible degradations, whereas military and aerospace communication systems require a quantitative behavioural performance test such as the Diagnostic Rhyme Test described in this report.

3. THE DIAGNOSTIC RHYME TEST

The purpose of speech intelligibility tests is to quantify the performance of voice communication systems and their associated components. The basic principle underlying these tests is the creation of speech material (usually monosyllabic words) for assessment by listening subjects who record their interpretations of what was said. Tests of this nature are quantitative since the number of words correctly identified can be counted, yet they are to a large degree subjective, in that the measurement system requires the participation of human subjects. Objective tests exist which seek to eliminate the human element by examining the degradation of artificial test signals [1]. Such tests are subject to serious limitations particularly when applied to narrow-band speech coders and consequently have not achieved widespread acceptance.

The Diagnostic Rhyme Test is used extensively for assessing the intelligibility of military communication systems and has become an accepted NATO standard for testing Linear Predictive Coders [2]. A discussion of the acoustic and phonetic research on which the test is based is beyond the scope of this report, consequently only a brief outline of the test is given here. An account of the origins and development of the DRT is given by Voiers [3].

The DRT vocabulary comprises ninety-six minimally contrasting rhyming word pairs (Figure 1), the initial consonants of which differ only by a single acoustic feature, or attribute. There are six such attributes; Voicing, Nasality, Sustention, Sibilation, Graveness and Compactness. As an example, the attribute voicing is present when the vocal cords are excited; in the word pair "veal-feel", the consonant "v" is voiced, but the consonant "f" is unvoiced. The acoustic basis for the remaining attributes is discussed in reference 3.

The DRT is implemented using the following procedure. The 192 word vocabulary is first recorded by a set number of talkers, usually 5 or 10. A total of thirty different preordained word orders are avialable to eliminate possible learning effects by listening subjects. The microphone used to make the recordings and the ambient acoustic environment are selected according to the proposed operational deployment of the communication system under test. The recordings are then processed by passing them through the actual communication system (or a laboratory simulation if this is not possible) and re-recorded (see section 7). The processed word lists are then presented aurally to a panel of listeners via the appropriate earphone transducer which is usually contained in a helmet, headset or handset. At the same time the listeners are also presented the appropriate word pair visually on a VDU. They then select the word they thought they heard by pressing one of two buttons. This arrangement is shown in Figure 2. The intelligibility score is expressed as a proportion of the number of words correctly identified by the listeners, with the scores adjusted for chance so that a subject guessing will score 0% rather than 50%.

To ensure adequate stability of the results, tests are conducted using not less than five talkers and eight listeners. Each of the talkers utters the 192 word vocabulary at a rate of one word every 1.33 seconds; a five talker test therefore lasts approximately twenty-five minutes. Male and female subjects are recruited from the local population and paid for their participation as listeners in the tests which are normally conducted three mornings a week. The implementation is very close to that employed by US testing agencies [3], thereby allowing the exchange of speech material between the US and the UK for comparative assessment.

The Acoustics Laboratories at Malvern include two high noise chambers where tests can be performed with the listening panel exposed to ambient noise levels representative of a variety of military platforms.

4. CHARACTERISING THE COMMUNICATION SYSTEM

Each test evaluates the performance of a real or simulated voice communication system composed of five elements. These are:

1. Talker's microphone.
2. Acoustic noise at the talker.
3. Transmission channel.
4. Listener's telephone (earphone, loudspeaker, etc).
5. Acoustic noise at the listener.

Any particular combination of these five elements defines a DRT condition (see Figure 3). A database of tape recordings [4] has been assembled covering a wide range of military and civilian voice communication systems. The recordings comprise master talker lists produced either in quiet conditions, or in the presence of operational background noise. In the former case noise can subsequently be added electronically to simulate a particular acoustic environment without having to re-record the talker list. Material produced using this method will be referred to as "mixed", to differentiate it from "live" recordings where the background noise is actually present when the talker lists are created. Each method carries both advantages and disadvantages and a full description of how the material is prepared may be found in section 7. It is possible at any time to process the wordlists through any new types of transmission channel for comparison with previous data. Test condition elements 4 and 5 above are selected at the actual time of test and thus test conditions can readily be repeated with a different type of telephone or headset, or in a different listener acoustic noise field.

A given combination of these DRT elements is deliberately chosen to provide an accurate simulation of a given operational application. There is usually no difficulty (barring equipment availability) conducting tests using the correct types of electro-acoustic transducers, although it is necessary to ensure that the sample transducers are in fact representative of operational stock. Certain military transducers of long standing design can vary in sensitivity and bandwidth between individual units.

Simulating the channel degradation imposed by radio systems may be achieved by connecting a transmitter and receiver back to back on a laboratory bench. Attenuation may be inserted between them to simulate propagation loss. Word list recordings could, in principle, be transmitted through an operational channel bearing in mind that in the case of moving vehicles particularly, transmission parameters are likely to fluctuate.

Often the greatest source of DRT score variability is the typical operational speech-to-noise ratios, both at the talker and listener positions. This can be difficult to measure directly as discussed in section 6 below. Noise discriminating boom microphones and mask microphones improve the speech-to-noise ratio in high ambient noise levels but talker vocal effort is also important. Talker vocal effort is dependent upon sidetone level and verbal feedback (eg "say again") and these parameters can be difficult to control.

Operational personnel will invariably be provided with some means of controlling listener level to achieve adequate volume without discomfort. A free choice of listener level is undesirable for DRT administration as the scores are highly dependent upon the speech-to-noise ratio for the listener when in high noise levels. This means that operational listener speech-to-noise ratios must be determined, and that an appropriate and repeatable listener level must be selected for the listener crew. There is, of course, no difficulty when all relevant operational parameters are well quantified and understood, but the introduction of a new voice communication system could cause operational personnel to alter their preferred parameter settings from those used with older systems. The present procedures for selecting listener levels are discussed in section 8 below.

There is an additional legal constraint on permissible noise levels imposed by the Health and Safety Executive. Consequently there is a listener crew hearing conservation noise exposure limit of 85 dB(A) L_{eq} per 8 hours as measured at the ear using miniature electret microphones. The numbers and durations of high noise level tests have to be limited to comply with this criterion even allowing for noise attenuating characteristics of the headsets or helmets.

5. CHARACTERISING THE ACOUSTIC ENVIRONMENT

Having defined the system components, and selected the platform in which they are to be deployed (ie tank, ops room, fast jet etc), the acoustic environment must be characterised before the Diagnostic Rhyme Tests can be conducted.

When communicating from a noisy environment, the quality of the speech signal (as defined by the speech-to-noise ratio) is dependent on not only the amount of background noise picked up by the talker's microphone, but also his vocal effort. Consequently it is not sufficient to measure simply the ambient noise level; the characteristics of the microphone (ie mask microphone, noise cancelling boom microphone) and the behaviour of the operator must be taken into account. In order to characterise the acoustic environment properly it is therefore necessary to record speech samples under operational conditions, in addition to ambient noise levels.

Ambient noise levels are then subjected to a frequency analysis, usually in one-third octave bands. The procedure for measuring speech-to-noise ratio is not so straightforward, and is described in the following section.

6. SPEECH-TO-NOISE RATIO

Where there is no speech waveform distortion, the major determinant of message intelligibility is the speech-to-noise ratio as measured in each one third octave frequency band. It is therefore vitally important that voice communication systems are tested at operationally representative talker and listener speech-to-noise ratios for the test data to be relevant.

In order to define a speech-to-noise ratio, separate estimates of the speech and the noise energies are required. However, to obtain realistic recordings with appropriate background noise and vocal effort, the speech samples must be obtained under operational conditions and will generally be contaminated by noise. Consequently the only recordings that can be made are; (i) those of the speech and noise inextricably mixed and (ii), the noise on its own. An estimate of the speech level may then be obtained by conducting a one-third octave frequency analysis on these data and subtracting (in linear units, not dB) the noise from the speech plus noise, for each one-third octave band; thus

$$S_i = 10.\mathrm{Log}_{10}\left[10^{SN_i/10} - 10^{N_i/10} \right] \qquad i = 14,42 \qquad (1)$$

$$\mathrm{SNR(Lin)} = 10.\mathrm{Log}_{10} \sum_{i=14}^{42} 10^{S_i} - 10.\mathrm{Log}_{10} \sum_{i=14}^{42} 10^{N_i} \qquad (2)$$

where: i is the one-third octave band number (25 - 16,000Hz).

SN_i is the Sound Pressure Level (in dB) of the combined speech plus noise signal in the ith one third octave band.

N_i is the Sound Pressure Level (in dB) of the noise signal in the ith one third octave band.

S_i is the Sound Pressure Level (in dB) of the speech signal in the ith one third octave band.

The A-weighting function may be applied to the one-third octave levels (for measurement purposes only), in which case the resultant overall speech-to-noise ratio is referred to as SNR(A). This procedure for measuring speech-to-noise ratio will now be illustrated by considering a practical example.

An earlier RSRE Memorandum [5] examined the proposed deployment of Linear Predictive Coders (LPC-10) in fast jet aircraft. A flight trial was conducted during which in-flight recordings were made using man-mounted miniature tape recorders sampling both the ambient cockpit noise and the speech signal from aircrew mask microphones.

In order to obtain the correct speech level, the "speech plus noise" recording must be analysed during a passage where there is continuous speech. However, the voice traffic during a typical sortie may be only occasional and very abrupt. In order to provide a continuous passage of not less than sixteen seconds for analysis purposes, aircrew were requested to recite the NATO alphabet (alpha, bravo, charlie etc) at several stages in the sortie.

An estimate of the "noise only" condition is not straightforward when an oxygen mask microphone is used and may not simply be made on a segment of the recording where no speech is present. There are two reasons for this; (i) the action of the oxygen supply regulator system may introduce a noise component during inhalation only (which is consequently not present when speaking) and (ii), the act of exhalation opens an expiritory valve causing a change in the acoustic attenuation of the mask. This alters both spectrum and level of the background noise picked up by the mask microphone. Therefore the conditions for obtaining a valid estimate of the noise pickup alone are only satisfied when the speech is present! In order to circumvent this difficulty the "noise only" analyses were performed on a series of one-half second exhalation periods. This arrangement is closest to the desired condition as there will be no regulator noise, and some opening of the expiratory valve. A typical example of such an analysis is given in Figure 4.

A similar problem is sometimes encountered with close-talking noise cancelling boom microphones where the airflow produced by speech (or indeed normal breathing) can be the source of aerodynamic noise around the microphone casing. Operators will tend to re-position the microphone to minimise the effect and a careful scrutiny of the operational recordings is required before a technical judgement can be made on the appropriate analysis technique.

A number of speech recordings from fast jet aircrew were analysed to yield overall speech-to-noise ratios. It was found that individual aircrew have a more profound effect on speech-to-noise ratios than the flight conditions (see Figure 5). This means that considerable care must be exercised when measuring talker speech-to-noise ratio to ensure that an adequate sample from the intended user population is taken. Similar recordings taken at the ear must be made to establish corresponding listener speech-to-noise ratios.

7. CREATING THE SPEECH LIST RECORDINGS

The DRT talker lists can be recorded in one of two different ways. One method requires talkers to be immersed in the appropriate background noise and wear the corresponding headgear incorporating the necessary transducers. Their speaking level is monitored by the operator conducting the recordings and also presented to the talker visually. The operator then sets a target speech level (derived from the operational recordings) for the talker to attain, which corresponds to the particular overall level and speech-to-noise ratio at which the recording is to be made. In practice it is only possible using this method to control the speech-to-noise ratio within approximately +/-2dB, a range which is sufficiently large to affect appreciably the DRT score. This method therefore inextricably confounds the effects of individual talker and speech-to-noise ratio. Recordings made in this way are referred to as "live" noise recordings and represent the most accurate possible laboratory simulation of operational conditions.

In order to eliminate this confounding, it is necessary to record the lists in quiet conditions (but with speakers still using the appropriate operational vocal effort) and subsequently mix in the background noise component electrically. This is the second method for creating speech material, and the talker lists prepared in this way are referred to as "mixed" noise recordings. Only by using this latter technique is it possible to ensure that the effect of noise is the sole independent variable. These two methods do not always produce the same DRT score and a comparison of live versus mixed recordings is discussed at section 10 below. Reference [6] contains a more detailed account of the recording process.

A set of seven word lists has been created using a range of techniques that degrade the speech to yield DRT scores in the range 50-90%. These lists are known as the "probe" lists, and they are presented to the listening panel at regular intervals in order to check individual subject consistency. The lists are also used to assess potential recruits to the listening panel.

8. ADMINISTRATION OF THE DIAGNOSTIC RHYME TEST

The listening subjects are locally recruited male and female paid volunteers. All are otologically and audiometrically normal, with hearing loss of less than 20 dB as defined by ISO R389 [7]. Listeners are selected on the basis of their ability to concentrate for extended periods and on the consistency of scores achieved with reference wordlists (called the probe lists) during a two day training course. Listeners normally attend 3 mornings per week with a 30 minute test session separated by 20 minute rest periods.

For the results to be statistically useful, it has been found that complete sets of data (ie every listener participates in every test condition) are required from not less than eight listeners. This requirement is currently met by recruiting a crew of ten listeners for each test series, which caters for some degree of absenteeism.

There is often some form of replication of each test in order to check for individual listener consistency and possible errors, or otherwise unnoticed equipment faults. The replication can either be an exact replication on a later date, or an interpolated replication where a sufficient number of levels of an independent variable are under test. The overall DRT score is averaged only across those listeners that have a complete set of data for at least one replication and will thus normally be averaged across up to the full ten listeners participating in each test condition. Missing values from either the first or second replication are dealt with by the statistical Analysis of Variance software provided by the RAF Institute of Aviation Medicine.

The listener crew selection procedures have recently been changed to include a larger pool of trained listeners from which to select groups of ten for each particular test series. The earlier practice of maintaining the same crew for all tests led to motivation and fatigue problems for some individuals who now benefit from regular resting periods. In addition, the knowledge that reselection for future tests depends not only on availability but also on previous test performance has improved listener consistency. Regular training sessions using the probe lists are conducted, interspersed with the normal test programme in order to confirm a high level of listener application to the task. It is also necessary to conduct regular checks of headset and helmet attenuation where high noise level testing occurs in order to ensure compliance with the hearing conservation noise exposure limit. Regular audiometric checks are performed on the entire listener pool together with a cyclic programme of individual audiometric checks after high noise level tests to probe for temporary threshold shift which would be a contra-indication for future high noise level tests for that listener.

The listener crew are not operational military or aerospace personnel. They tend to be mainly housewives or men not in regular employment prepared to supplement their income by participating in tests as and when required. It is possible that operational personnel might score differently if tested in exactly the same way. However, it should be remembered that the speech intelligibility facility described in this report is a measurement system. It is not the intention of these experiments to predict the DRT score that operational personnel might achieve if they participated as listening subjects. It is to attain a reliable and repeatable metric that may be used to quantify system performance and thus infer criteria of acceptability based on feedback from operational personnel carrying out tasks under simulated operational conditions. Such acceptability assessments are described in Sections 15 and 16. There does appear to be a difference, however, in the selection of volume level settings when listening in high noise levels. The volunteer crew, when given a free choice of listening level often select a speech level that minimises their noise exposure at the ear, at the expense of intelligibility. Operational personnel are unlikely to do this. Recent testing has shown that the DRT scores in high noise levels continue to increase when the speech-to-noise ratio at the listener is increased above the listener preferred level. This problem has been overcome by using levels derived from measurements of the operational environment. This procedure must be invoked with great care, as sometimes operators have been known to select levels so high

that not only is the risk to their hearing considerable, but additionally their hearing system has been so overloaded that the intelligibility of the incoming signal has actually been impaired.

Most listeners when first recruited improve their scores rapidly within the first few days of training or actual testing. A few listeners continue to improve very slightly over a period of several months but the scores for the majority of listeners then reduce at a very slow rate after several months of continuous testing. This is the reason for the current listener rotation procedures. There appears to be no detectable word list learning effects even when the same randomisation sequence is presented repeatedly over several weeks testing. However, different word list randomisations are used wherever possible as a precaution against possible residual word list learning effects.

9. THE STATISTICAL ANALYSIS OF DIAGNOSTIC RHYME TEST DATA

At the conclusion of a test, the data are processed by a program that computes a two-way Analysis of Variance and provides basic statistics. An example of such an analysis is given in Figure 6.

These results shown are from an experiment which examined the effect of additive broad-band noise with a speech-like spectrum on intelligibility. The five elements making up the condition are given, together with details of the talkers and listeners. The list numbers refer to a prescribed ordering of the DRT vocabulary. This is followed by an Analysis of Variance, which tests for a difference in scores amongst the samples of the independent variables (talkers and listeners) and the existence of a Talker x Listener interaction. A Newman-Keuls test may then be used to determine the significance of the rank order of the mean scores for the talkers and listeners. A group of listeners or talkers that are underlined (see Figure 6) indicates that there is no significant difference between the scores of the members of that group.

Thus in the example printout shown there is no significance in the rank ordering of the bottom 9 or top 9 listeners. However, Listener 4 scored significantly lower than the remaining subjects, similarly Listener 9 scored significantly higher than the remaining subjects. For the talkers, there is no significance in the rank ordering between talkers in the two groups T3-T6-T5-T2, and T1-T2. The absence of a line beneath Talker 4 means that his scores are significantly lower than all the other talkers.

The Diagnostic Rhyme Test comprises six attributes and therefore it is possible to compute the attribute scores, averaged over all the talkers. This is done both for when the attribute is present, and when it is absent. Thus the score for "voicing present" (87.5%) refers to the number of voiced consonants correctly identified. The "voicing absent" score (64.6%) refers to the number of unvoiced consonants correctly identified. The mean score is the average of both cases, and the number in the column SE gives the appropriate Standard Error of the mean.

This information is then followed by basic statistics giving the condition scores averaged over talkers and listeners, with corresponding estimates for Standard Deviation and Standard Error. It is normal when presenting Diagnostic Rhyme Test results to treat listeners as a random effect and talkers as a fixed effect, which results in a set of numbers for each listener which is averaged over talkers. It is the Mean and Standard Error of these numbers that appear at end of the analysis.

These results could then be compared with those from any other condition using the student t-test, which will test the hypothesis that the two data sets came from the same population. In practice however most investigations examine a range of conditions which would consequently require a large number of paired comparisons. Under these circumstances it is more appropriate to conduct an Analysis of Variance. This technique allows the total variance of the DRT scores to be decomposed into a series of factors and their interactions.

There are three basic factors represented in a DRT; Talker, Condition and Listener. The Condition may itself be further decomposed as described by Figure 2. Factors may be classified either as fixed effects, for which all possible levels of the relevant factors have been samples, or as random effects, where the levels of the factors in the experiment represent a random sample form an infinite population. If a factor is treated as a fixed effect, any inferences from the data are only valid for the particular samples of that factor used in experiment, whilst if it is treated as a random effect, inference may then be extended to the parent population.

Of the three factors that are present in a DRT, clearly Condition, with its deliberately selected combination of elements, is treated as a fixed effect. Listeners should be treated as a random effect, since the results of a DRT must be sufficiently robust to accommodate changes in the composition of the listening crew. Talkers have customarily been treated as a fixed effect since the source material, once recorded, has fixed characteristics. This approach has important consequences that are best illustrated by the example given in the next section. A more detailed treatment of Analysis of Variance may be found in the Annex.

10. AN EXAMPLE OF ANALYSIS OF VARIANCE

An experiment was designed to examine primarily any difference in DRT scores between live and mixed recordings (as defined in section 6). The range of factors studied make this experiment a suitable vehicle for conducting a detailed analysis of all the factors that contribute to speech intelligibility as measured by the DRT.

The effect of using speech material recorded using the live and mixed techniques was examined using three coders, 2.4 k.bits/s Linear Predictive Coder (LPC-10), 9.6 k.bits/s Residually Excited Linear Prediction (RELP) and 16 k.bits/s Continuously Variable Slope Delta-modulation (CVSD). The five talkers participating each recorded two lists (i.e. two sets of "utterances") under both live and mixed conditions. The entire test was subsequently replicated.

The design and subsequent analysis of the experiment allows the following hypotheses to be tested; namely, that there is no difference between the scores of the following.

1. Coders
2. Talkers
3. Listeners
4. Live and mixed speech material
5. Utterances
6. Replications

Normally it is only the performance of the components of the voice communication system that are of interest (in this case the coders). However, the results of an Analysis of Variance show that there is a number of important interactions between the experimental factors and it is the precise nature of these interactions that exerts a considerable influence on the design and conduct of the DRT.

In order to examine the influence of all six factors, an Analysis of Variance was performed treating all factors as fixed effects. With the exception of Replication, all main effects are highly significant, with Talkers exerting the biggest influence on the total variance, followed by Utterances, Coders, Live/Mixed speech, and finally Listeners. Thus the null hypotheses 1-5 above are rejected ($p<0.0001$). The effect of the listeners although highly significant was found to be very small in magnitude.

Clearly while it is important to examine such listener effects in order to quantify their contribution to the overall variance, the analysis may now be simplified by treating Listeners as a random effect and re-analysing the data (Figure 7). Again, with the exception of Replications, all the main effects were found to be very highly significant. Replication only becomes statistically significant for interactions involving four factors and at that level there are no important consequences for the conduct of the experiment. This demonstrates that the DRT is a repeatable test which gives consistent results.

In order to gain an appreciation of the main trends of this experiment, consider the following first order interactions T x C, T x S and C x S, which are illustrated in Figure 8. A graphical representation is useful for demonstrating the cause of an interaction. The existence of a T x C interaction is extremely important and is shown in Figure 8a. It shows that the performance of vocoders is talker dependent. The 2.4 k.bits/s and 9.6 k.bits/s systems use the same basic principle (Linear Predictive Coding) and the performance of the five talkers is very uniform, in other words the lines joining the talker scores for these two coders are nearly parallel. The scores for the CVSD system do not show the same profile as a function of talker, the most marked difference being the "cross-over" for Talker 2. The effect of such a systematic bias may be compensated for by re-analysing the data treating talkers as a random effect, but it does indicate the

dangers of making comparisons between coders using only a small talker sample. Coders not included in this study (ie channel vocoders) may be subject to even greater talker variability.

The T x S interaction (Figure 8b) is a combination of two effects; (i) the success with which noise can be added in the correct proportion to match the corresponding live list and (ii), any characteristics of the talker's voice (or indeed the particular recording) that might introduce systematic bias. There may be further possible explanations for the observed interaction. The former effect would be an artifact of the recording process and might affect the results of all three coders, whereas only the 2.4 kbit/s scores were substantially altered. The characteristics of the talkers voice, the use of a sound level indicator to obtain the correct vocal effort and the way in which mixed noise alter the nature of the word list recordings all require further study before a satisfactory explanation can be offered.

The C x S interaction (Figure 8c) suggests that the difference between the use of live and mixed speech is only significant for the 2.4 k.bits/s coder. This raised the question of whether this effect was common to all 2.4 k.bits/s systems, or peculiar to the particular device used in the test. It was decided to re-test this device together with a different LPC-10 2.4 k.bits/s coder conforming to the same interoperability specification [1] and the result is shown in Figure 9. The re-test scores for the original coder were very consistent and the new device showed a difference between live and mixed material of only 3.3% compared with 7.3% (7.2% on the first test). This suggests that the effect is, at least partially, implementation dependent.

The object of the DRT is to quantify system performance, but an appreciation of the importance of all the factors is necessary in order to conduct sensible experiments and perform appropriate analyses. Using the Newman-Keuls test it is possible to test the significance of the differences in mean scores for the three coders.

The test was applied to the data for the following two cases; (i) treating listener as a random effect and (ii), treating both listener and talker as random effects. The former test, whose results may be inferred to the general population of listeners but are only applicable to the five talkers tested, showed that a difference in DRT score betweeen coders of greater than 1.1% was significant. The corresponding figure when comparing the scores for live and mixed speech is 0.9%. When both talkers and listeners are treated as random effects, these figures rise to 3.3% and 6.5% respectively. The comparatively large increase in variability (0.9% to 6.5%) when talkers are treated as a random effect is due to the magnitude of the T x S interaction (see Figures 7 and 8b).

There are two main conclusions to be drawn from this experiment. Firstly, the performance of the three coders depends to some degree on the individual talker. An increase in the number of talkers used in the tests would therefore help to improve both robustness and discrimination of the tests. Secondly, there would appear to be a genuine difference in DRT score between live and mixed speech material when low bit rate (2.4 k.bits/s) vocoders are tested. This difference for the two implementations tested varied from 7.2% to 3.3%, with the live recordings producing the higher score in each case. The correseponding scores for the 9.6 and 16 k.bits/s systems are not significantly different. This result should be noted with specifying DRT conditions for performance bench-mark testing purposes. A discussion on the use of live versus mixed speech may be found in section 11.

11. STRATEGIES FOR CONDUCTING THE DRT

The results presented in the previous section have important consequences for the conduct of the DRT. The difference in scores between the two sets of Utterances was found to be 3.8% when averaged over all other factors and is due to the difficulty of recording live talker lists to a specified SNR. As a consequence, comparison between equipments should only be made when using a common set of talker recordings. This is not a serious imposition, as even the repeated presentation of the same DRT word list randomisation to the listening crew has not been found to induce any detectable learning effects.

For the coders operating at 9.6 and 16 k.bits/s there would appear to be no material difference as to whether live or mixed speech is used, but for the two systems running at 2.4 k.bits/s that were examined, the live lists scored somewhat higher than the corresponding mixed set. As it is the live recordings that more closely resemble the operational acoustic environment it is desirable to use such recordings, particularly when testing low bit rate vocoders. However the difficulty in controlling the experimental variables, such as the speech-to-noise ratio, favours the use of mixed material. The choice of whether to use live or mixed word lists will depend on prevailing circumstances, but the method used must be stated.

In the early stages of the test, two different experimental designs were employed. Initially one of the Acoustic Laboratory high noise rooms was fitted with eight subject stations and a permanent listening crew of twelve people was recruited and trained. Experiments were then conducted using a selection of eight listeners (from the total of twelve) participating in any one test.

In September 1985 a second noise room was commissioned and an alternative design, using all twelve listeners (six in each room), was implemented. Based on these experiences, three possible strategies for conducting the tests were considered:

1. Conduct a five talker experiment using a single room with eight listeners and a second replication ensuring all twelve subjects complete each test.

2. Conduct a five talker experiment using both rooms with all twelve listeners participating in a single replication.

3. Conduct a ten talker experiment using both rooms with all twelve listeners participating in a single replication.

The statistical basis of the three strategies is examined in the Annex and the conclusions are as follows. Strategy 2 leads to the smallest experiment, whilst Strategies 1 and 3 take twice as long to administer, with Strategy 3 requiring extra preparation time for the recording and processing of the additional five talker list. In practice it was found that, taking Strategy 2 as the baseline, the statistical resolution is improved by 5% for Strategy 1 and by 26% for Strategy 3. It should be noted that Strategy 1 does contain an element of replication and thus provides useful information concerning test consistency and possible errors or otherwise undiscovered equipment faults.

As a result of this experience it was decided to equip one room to accommodate 10 listening subjects and perform a replication for each test to confirm the validity of the results obtained. This procedure has enabled any anomalies to be readily spotted and provides adequate statistical discrimination for the majority of investigations. It is intended to conduct a proportion of future tests with a 10 talker sample using only one presentation to minimise the influence that individual talkers have on the results, without increasing the experimental time.

12. COMPARISONS WITH US DRT RESULTS

The DRT has been conducted in the US for many years and a considerable corpus of results exists for a wide range of communication systems. Results of experiments using LPC-10 equipment naturally invite comparison with those conducted in the UK, but the validity of such comparisons still remains to be established.

A limited exchange of speech material has enabled some preliminary experiments to be conducted. The use of "probe" lists as described in section 7 to select listeners and maintain checks on their consistency follows the procedures devised by Dr W D Voiers, President of Dynastat Inc. The RSRE has exchanged probe lists with Dynastat for assessment by each others listening crews, and the discrepancy in DRT score ranges from 0-5%.

One particular comparison of interest concerns the use of LPC-10 in F15 noise. The DRT figure quoted by a US Report [8] is 70.5% (S.E. = 1.95%). Copies of these tape were obtained and the test replicated at the RSRE and a corresponding figure of 69.6% (S.E. = 1.4%). Insufficient comparative data exists to draw any firm conclusions but the differences noted to date are minor. However, it is hoped to conduct further exchanges of material at some future date.

13. A GUIDE TO THE CONDUCT OF THE DIAGNOSTIC RHYME TEST

The preceding sections have described in detail the administration and analysis of the Diagnostic Rhyme Test. This section gives guidance for the conduct of such tests. The DRT can be used for two distinct purposes and the approach used needs to be adapted accordingly.

When used as a comparative test the results should be subjected to an Analysis of Variance followed by Newman-Keuls tests to examine the significance of the differences in scores between the independent variables. This is the procedure followed for the analysis of the experiment described in section 10 and it allowed comparisons to be made between the independent variables such as coders, replications and the use of live and mixed speech.

Having established a minimum DRT score as a result of acceptability experiments of the type described in the next section, it may then be desirable to use the test for bench-mark purposes. This procedure could be invoked to check the adequacy of a particular combination of components forming a voice communication system. In this case the absolute value of the DRT score needs to be assessed which would be valid for the given set of talker word lists. If the observed DRT score exceeds the quoted bench-mark by more than twice the Standard Error as defined in section 9, then the confidence level that the criteria has been met is 95%.

For both applications described above decisions concerning the design of the experiment will have to be made; namely, the number of talkers to be sampled, the requirement to replicate the tests and the use of live or mixed speech material. The first two points should be taken together as they both affect the time taken (and hence cost) to complete an investigation. It has been shown in section 10 that doubling the number of talkers used is a more effective way of increasing the discrimination of the test than replicating the original set, with the time taken to conduct the experiment being the same in each case. This strategy also improves the robustness of the test with respect to individual talker differences by reducing the effect of the T x S and T x C interactions. The use of a replication provides a useful check on listener consistency and assists the speedy identification of anomolous results. The confidence gained from running the facility during the last two years has reduced the need for this replication and it is anticipated that a greater proportion of future tests will be run using ten talkers. Although additional master recordings will need to be made their cost, when spread over the life of the anticipated future programme, should prove modest.

Finally a decision must be made concerning the use of live or mixed speech material. Clearly more research is needed in order to gain a better understanding of the differences observed in the scores of the 2.4 k.bits/s systems before a definitive statement can be made. The preparation of live word lists are more closely follows real usage of voice channels and is therefore to be preferred for the bench-mark studies. Mixed material will continue to be used for studies where experimental variables (such as speech-to-noise ratio) need to be varied singly and in a closely controlled manner. Any findings should then be checked using live material before drawing firm conclusions.

14. ACCEPTABILITY ASSESSMENT EXPERIMENTS

The DRT yields intelligibility scores expressed as the percentage of initial consonants correctly identified (adjusted for chance), together with a statistical estimate of confidence derived from an analysis of variance that may be used to discriminate between different systems.

The question then arises of how to interpret these figures when considering the suitability of a candidate voice communication system for a particular application. What is required here is a context-dependent task in which representatives of the intended user population have the opportunity to assess the system in a way that relates as closely as possible to their operational environment, and then complete a questionnaire that seeks their opinions on a number of aspects of system performance. An example of a questionnaire devised to explore the suitability of LPC-10 for Air Defence communications is shown in Figure 10.

For the case of airborne communications, clearly the most appropriate task would be to conduct flight trials in aircraft fitted with the communications system to be evaluated. This solution is expensive and may often prove impossible if the equipment to be tested is only in development form. The next most realistic environment is a full cockpit flight simulator, to which prototype systems may be connected without too much difficulty or expense. Such a trial was conducted using a Phantom cockpit flight simulator during 1986. A much simpler and cheaper approach which has nevertheless proved very valuable, is the use of a flight simulator software package for a personal computer. Some preliminary work on acceptability has already been conducted using a flight simulator package called "Aviator" which runs on the BBC Microcomputer. This program simulates the controls and handling characteristics of the Spitfire aircraft, combined with the addition of radar and the ability to generate hostile aircraft. In order to introduce voice communications into the task, the pilot's radar display is covered and the radar information passed to a second subject acting as a fighter controller, who gives instructions as appropriate to the pilot (see Figure 11). Thus intercepts car. only be made if pilot and controller can communication sucessfully. The results of a trial using this method are given in the next section.

15. ACCEPTABILITY ASSESSMENT RESULTS

Some informal listening tests of LPC-10 in the presence of 108-112 dB of simulated fast jet cockpit noise were carried out by three RAF aircrew during 1983 before the simulators had been commissioned. Their microphone signals were recorded and subsequently analysed showing speech-to-noise ratios in the range 12 - 15 dB. All three subjects rated the performance of LPC-10 under these conditions as acceptable or better.

Since that time the flight simulator program has been used for an assessment trial using three fighter controller assistants, who also acted as pilots for these experiments. Sorties of the type described in the previous section were flown in fast jet ambient noise levels of 104 - 112 dB, with the pilot subjects achieving speech-to-noise ratios of 12 - 16 dB. These conditions, which correspond typically to DRT scores in the range 65 - 71% using mixed speech, were again rated as acceptable to moderate. Average responses are shown as solid trangles on Figure 10. When asked to judge the system as either acceptable or unacceptable in terms of both intelligibility and overall usability, the system was unaminously judged as acceptable on both counts.

Results from the trial using operational personnel support the conclusion that a LPC-10 voice link yielding DRT scores in excess of 65% (mixed speech material) will be rated as acceptable or better by operational aircrew and fighter controllers.

16. THE RELATIONSHIP BETWEEN INTELLIGIBILITY AND ACCEPTABILITY

The US Department of Defense Digital Voice Processor Consortium (DVPC) has been evaluating the intelligibility of voice communication systems for a number of years, and its membership has arrived at a consensus view of how categories of DRT scores relate to eight verbal descriptors ranging from "excellent" to "unacceptable" (Figure 12). The DVPC notes that these categories are not directed at any particular class of user and should only be taken as a guide.

A document describing a NATO standard for LPC-10 [2] calls for a minimum DRT score of 75% when using in the F15 cockpit environment, but there is no discussion of the origin of this criterion.

More recently a paper by Tierney and Schecter [9] has examined the issue of user acceptability assessments for communications between an F15 and an E3A using various LPC algorithms. F15 pilots and airborne fighter controllers (seated in noise chambers to simulate operational conditions, but not carrying out any tasks) conduct improvised exchanges around three basic air-combat scenarios. When using a continuous rating scale described by the adjectives Unacceptable (0) to Excellent (100) with a mid-point of Acceptable (50), the average score for F15 to E3A communications using the US DOD standard LPC-10 was approximately 63%, ie some way above the mid-point of "aceptable". In a report by Singer (reference 8), a DRT score for three talker lists prepared in simulated ambient F15 noise (ie live) was given as 70.5%. This would suggest therefore that the acceptability borderline is somewhere below this figure.

A RSRE Memorandum [5] described tests using mixed speech lists recommended a DRT score of 65% with scores below 60% regarded as unacceptable, but the results of the live verses mixed experiments would suggest this figure be increased by approximately 5% when using live material with 2.4 k.bits/s vocoders. Thus the predicted "recommended" and "acceptable" values for live speech are 70% and 65% respectively and are comparable with the American results.

It should be stressed that a given intelligibility score is a necessary but not sufficient requirement for a communication systems. For example, consider a wide-band, noiseless, distortion-free communications system that has a long transmission delay. The DRT score could be very high, but the link would be rated unacceptable where speed of response is important (ie air defence intercept). It is the responsibility of the procurement agencies to confirm the DRT scores they require by conducting suitable acceptability experiments using operational personnel. When the user acceptability data from the trial has been fully analysed, and DRT experiments conducted under representative acoustic conditions, a base-line acceptable DRT score can be finally established for that application.

17. SUMMARY

This Report has advanced an assessment methodology based on two different types of experiment.

(i) The Diagnostic Rhyme Test, which is a quantitative, context-independent measure of initial consonant intelligibility.

(ii) Simulation techniques, which involve the user in a task that evokes as closely as possible the conditions of his own operational environment and subsequently permit him to assess various aspects of system performance including intelligibility.

By comparing the responses to the questionnaire with the corresponding DRT scores, it has been possible to explore the relationship that exists between the two assessment techniques and give provisional guidelines concerning the DRT scores that will be required for Air Defence sorties.

This relationship has only been validated for the particular application of Air Defence intercepts. It will need to be established for other military contexts (for example communications between tanks) as appropriate.

18. RECOMMENDATIONS

1. Tests should continue to be performed using not less than five talkers and complete sets of data from not less than eight listeners.

2. DRT scores should be subjected to an Analysis of Variance and the Newman-Keuls test used to discriminate between experimental variables.

3. Future tests that are used to bench-mark systems with respect to a target DRT score should preferably be conducted using at least ten talkers with "live" (refer to section 7) speech material. Comparative studies may continue to use five talkers if this provides adequate statistical discrimination and the results are checked for anomalies using live speech.

4. In addition to carrying out standard tests, a research programme should be initiated which will explore in more detail issues including the differences between the use of live and mixed speech material, the effect of utterance on the consistency of DRT scores when using mixed material, and the factors influencing talker variability. This programme should also study the use of physical measurement testing methods, such as the Speech Transmission Index, and conduct a comparison of the capabilities of such methods with corresponding DRT techniques.

5. A collaborative programme of tests be agreed with other international testing agencies to examine national differences between talkers and listening crews.

6. Speech intelligibility and acceptability assessments are now mature techniques suitable for providing a quantitative measure of voice communication system performance. Their incorporation into procurement specification is strongly urged; the RSRE is available to discuss particular applications and advise on appropriate criteria.

19. ACKNOWLEDGEMENTS

The authors would like to thank the RAF personnel for their participation in the flight trials to collect speech and cockpit noise recordings and Dr W D Voiers and his colleagues at Dynastat Inc for many helpful discussions concerning the conduct of Diagnostic Rhyme Tests.

20. REFERENCES

1. Nato Stanag 4198. Parameters and characteristics that must be common to ensure interoperability of 2400 bits/s linear predictive encoded digital speech.

2. Steeneken H. J. M., and Houtgast T. A physical method for measuring speech transmission quality. Journal of the Acoustical Society of America, Volume 66, pp 318-326, 1980.

3. Voiers W. D. Evaluating processed speech using the Diagnostic Rhyme Test. Speech Technology, pp30-39, January/February 1983.

4. Lower M. C. Flindell I. H. and Wheeler P. D. Index to DRT Wordlist Sets produced at ISVR. Institute of Sound and Vibration (Southampton University) Report No AC 575/2, June 1986.

5. Pratt R. L. Royal Signals and Radar Establishment Memorandum 3710, May 1987.

6. Lower M. C. Flindell I. H. and Wheeler P. D. The Diagnostic Rhyme Test Facility at RSRE. Institute of Sound and Vibration (Southampton University) Report No AC 575/1, May 1986.

7. ISO R389. Standard reference zero for the calibration of pure-tone audiometers. International Standards Organisation, 1964.

8. Singer E. A. A comparative study of narrow-band vocoder algorithms in Air Force operational environments using the Diagnostic Rhyme Test. Lincoln Laboratory Technical Report 590, January 1982.

9. Tierney J. and Schecter H. The Lincoln Laboratory-Aerospace Medical Research Laboratory Digital Speech Test Facility. Lincoln Laboratory MIT Technical Report 683, May 1984.

10. Smith C. P. Narrow-band (LPC-10) vocoder performance under combined effects of random bit errors and jet aircraft cabin noise. Rome Air Development Center, Technical Report No 83-293, December 1983.

VOICING	NASALITY	SUSTENTION
Voiced---Unvoiced	*Nasal---Oral*	*Sustained---Interrupted*
veal---feel	meat---beat	vee---bee
bean---peen	need---deed	sheet---cheat
gin---chin	mitt---bit	vill---bill
dint---tint	nip---dip	thick---tick
zoo---sue	moot---boot	foo---pooh
dune---tune	news---dues	shoes---choose
voal---foal	moan---bone	those---doze
goat---coat	note---dote	though---dough
zed---said	mend---bend	then---den
dense---tense	neck---deck	fence---pence
vast---fast	mad---bad	than---dan
gaff---caff *	nab---dab	shad---chad
vault---fault	moss---boss	thong---tong
daunt---taunt	gnaw---daw	shaw---caw
jock---chock	mom---bomb	von---bon
bond---pond	knock---dock	vox---box

SIBILATION	GRAVENESS	COMPACTNESS
Sibilated---Unsibilated	*Grave---Acute*	*Compact---Diffuse*
zee---thee	weed---reed	yield---wield
cheep---keep	peak---teak	key---tea
jilt---gilt	bid---did	hit---fit
sing---thing	fin---thin	gill---dill
juice---goose	moon---noon	coop---poop
chew---coo	pool---tool	you---rue
joe---go	bowl---dole	ghost---boast
sole---thole	fore---thor	show---so
jest---guest	met---net	keg---peg
chair---care	pent---tent	yen---wren
jab---gab	bank---dank	gat---bat
sank---thank	fad---thad	shag---sag
jaws---gauze	fought---thought	yawl---wall
saw---thaw	bong---dong	caught---taught
jot---got	wad---rod	hop---fop
chop---cop	pot---tot	got---dot

FIGURE 1

The Diagnostic Rhyme Test Vocabulary

* Note the spelling of the original word calf has been changed to caff so that it rhymes with gaff, when pronounced by British speakers.

FIGURE 2

Listening Subjects participating in Diagnostic Rhyme Tests

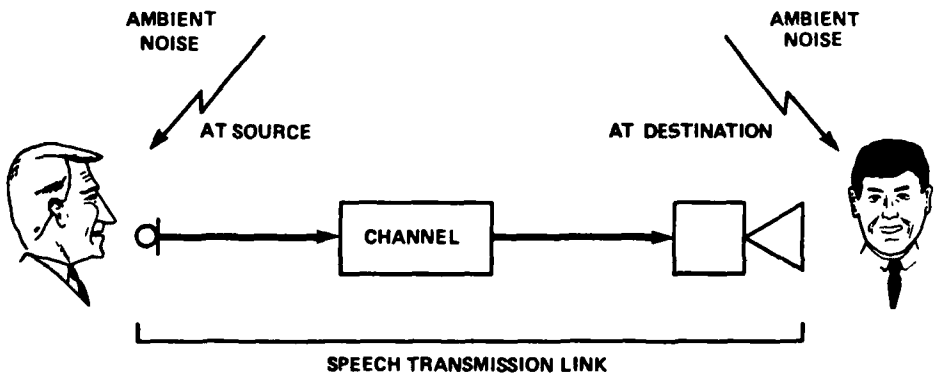


FIGURE 3

Principal Components of a Generalised Speech Transmission Link

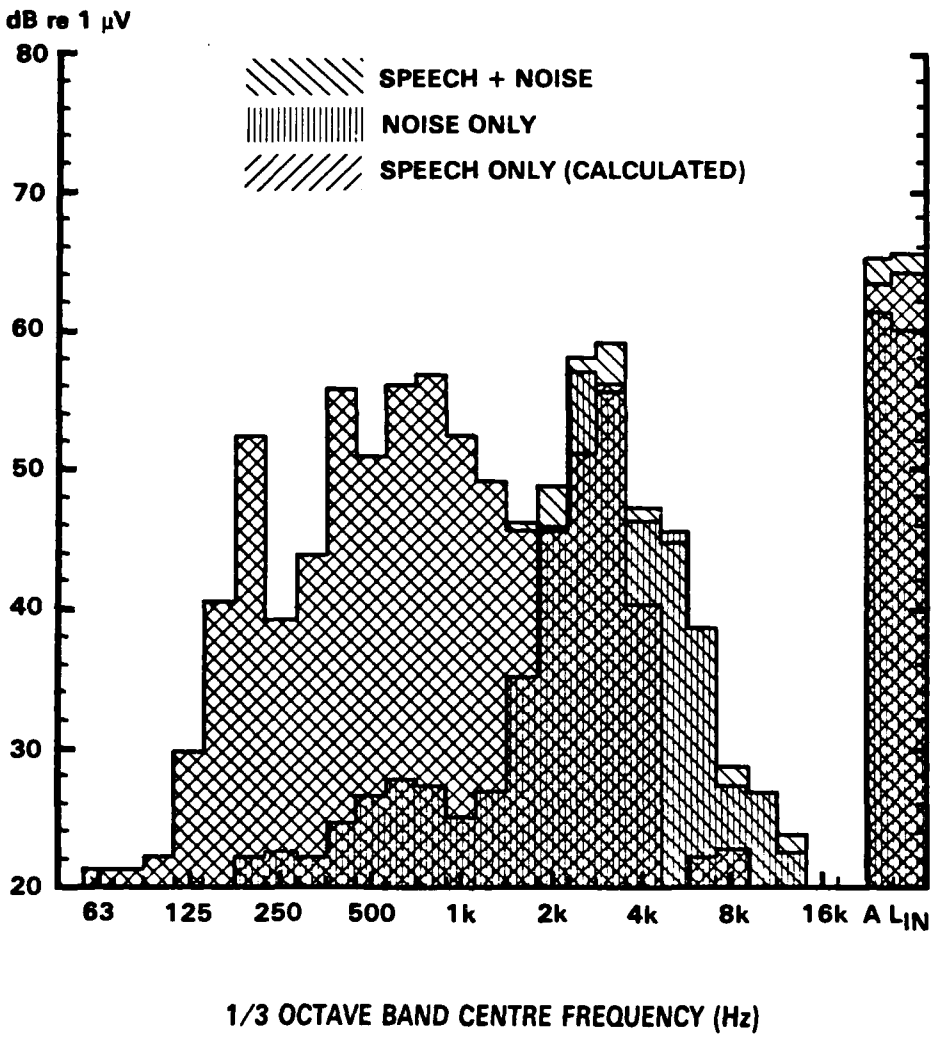


FIGURE 4

One-third octave analyses of in-flight recordings taken from the mask microphone of a fast jet pilot. For frequencies below 1.5kHz the "noise only" condition is typically 20dB below that of the "speech and noise" condition. Thus the contribution made by the noise to the overall level is negligible, with all the energy present being due entirely to the speech signal. In the 2-4kHz range it is the noise that dominates and the speech-to-noise ratio is negative in all but one of the bands.

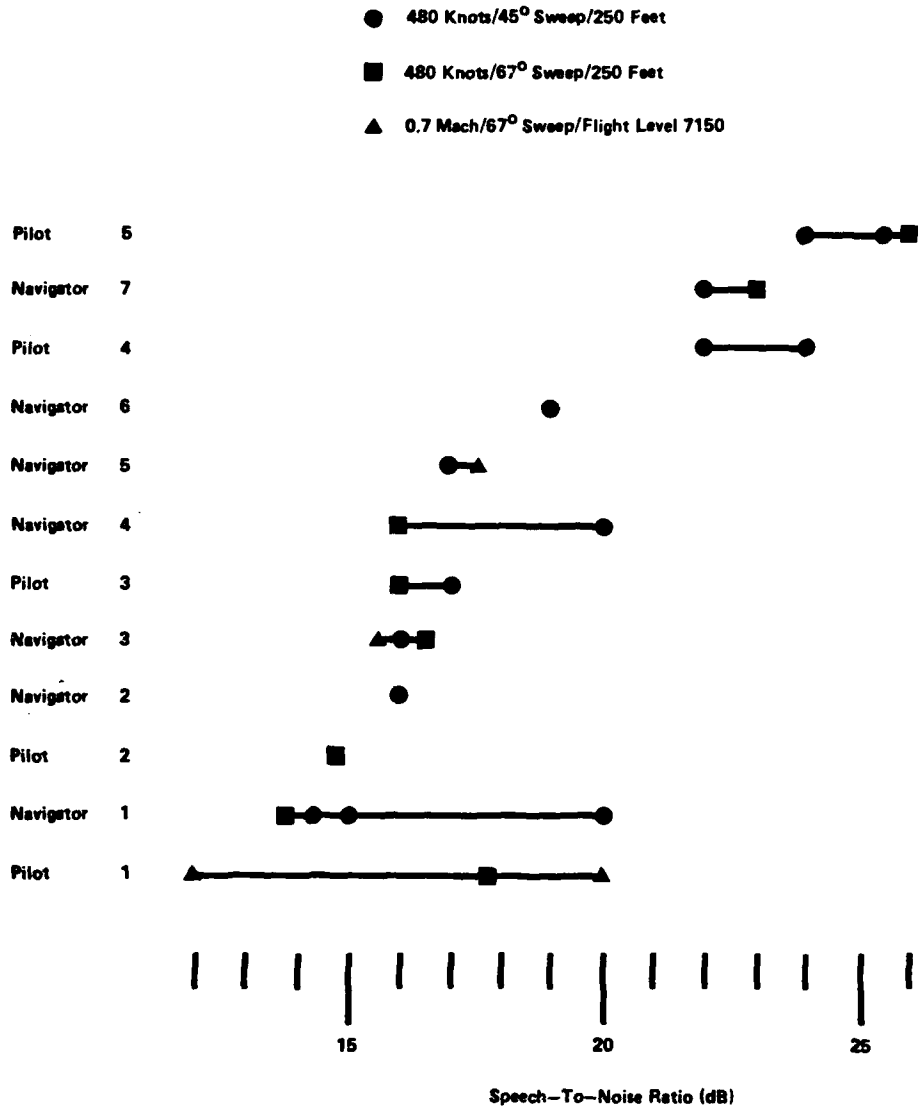


FIGURE 5

Speech-to-noise ratios produced by fast jet aircrew derived from in-flight recordings. Individual aircrew are an important source of variability. Pilot 1 shows an 8dB range in speech-to-noise ratio (from 12 - 20dB) for the same flight condition. Navigator 3, however, achieves a speech-to-noise ratio that is independent of flight condition.

Control File: CTL1770

Source: Multi-Listener Master V4
Date: 23 Oct 86

Input Device: NONE
Input Noise: ODB
Link Device: NONE
Output Device: SONY
Output Noise: QUIET

Optional Header Text
SYNTHETIC FOR RLP
V12 C3

Sequence Type: DYNA
Filler Display: Enabled

Words File: DYNAWRD
Sequences File: DYNASEQ

DRTs: 12

! Talker	List
RLP	306A.1
RLP	306A.2
RB	308A.1
RB	308A.2
MCL	315A.1
MCL	315A.2
PRW	312A.1
PRW	312A.2
GAP	314A.1
GAP	314A.2
RDR	303A.1
RDR	303A.2

Listeners: 10

! Listener	Station
KNUTSEN	1
FERRIS	2
TAYLOR	3
PERKINS	4
CHAMBERS	5
PRATT	6
PAWSEY	7
WHEATSTONE	8
ROBERTS	9
BERRY	10

FIGURE 6

Analysis of Variance and basic statistics for a single Diagnostic Rhyme Test.

ANALYSIS OF VARIANCE

SOURCE	SUM OF SQUARES	DEGREE OF FREEDOM	MEAN SQUARE	F-RATIO
LISTENER	2212.63	9	245.85	4.46
TALKERS	4093.31	5	818.66	14.85
LN X TK	1162.69	45	25.84	0.47
ERROR	3307.25	60	55.12	
TOTAL	10775.88	119		

NEWMAN - KEULS TEST

LISTENERS:

4 10 6 5 2 7 3 1 8 9

TALKERS:

4 3 6 5 2 1

ATTRIBUTE SCORES

	PRESENT		ABSENT		MEAN	
	MEAN	SE	MEAN	SE	MEAN	SE
Voicing	87.5	3.9	64.6	8.4	76.0	5.2
Nasality	97.7	0.7	93.1	1.2	95.4	0.8
Sustention	64.0	4.7	64.2	4.7	64.1	3.2
Sibilation	80.2	1.8	93.1	1.1	86.7	1.8
Graveness	69.2	5.4	59.8	5.9	64.5	4.0
Compactness	85.8	1.8	80.8	2.8	83.3	1.7

INDIVIDUAL SCORES

LNR	T1	T2	T3	T4	T5	T6	SD
1	90.6	89.6	80.2	75.0	79.2	81.3	6.2
2	83.3	80.2	75.0	66.7	80.2	78.1	5.9
3	87.5	82.3	81.3	74.0	79.2	78.1	4.5
4	83.3	75.0	71.9	61.5	72.9	74.0	7.0
5	88.5	75.0	78.1	55.2	79.2	68.8	11.3
6	89.6	69.8	70.8	62.5	76.0	74.0	9.0
7	89.6	81.3	76.0	65.6	84.4	81.3	8.2
8	90.6	90.6	79.2	77.1	81.3	82.3	5.8
9	88.5	87.5	82.3	78.1	85.4	88.5	4.1
10	80.2	82.3	69.8	61.5	74.0	75.0	7.5

FIGURE 6 (Cont)

LNR	MEAN OVER TKR
1	82.6
2	77.3
3	80.4
4	73.1
5	74.1
6	73.8
7	79.7
8	83.5
9	85.1
10	73.8

STD. DEVN. = 4.5
STD. ERROR. = 1.4

TKR	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	SD
1	90.6	83.3	87.5	83.3	88.5	89.6	89.6	90.6	88.5	80.2	3.6
2	89.6	80.2	82.3	75.0	75.0	69.8	81.3	90.6	87.5	82.3	6.7
3	80.2	75.0	81.3	71.9	78.1	70.8	76.0	79.2	82.3	69.8	4.5
4	75.0	66.7	74.0	61.5	55.2	62.5	65.6	77.1	78.1	61.5	7.9
5	79.2	80.2	79.2	72.9	79.2	76.0	84.4	81.3	85.4	74.0	4.0
6	81.3	78.1	78.1	74.0	68.8	74.0	81.3	82.3	88.5	75.0	5.6

TKR	MEAN OVER LNR
1	87.2
2	81.4
3	76.5
4	67.7
5	79.2
6	78.1

STD. DEVN. = 6.4
STD. ERROR = 2.6

DRT SCORE = 78.3
STANDARD ERROR = 1.4

FIGURE 6 (Cont)

ANALYSIS OF VARIANCE FOR VARIABLE :DRT

FACTORS

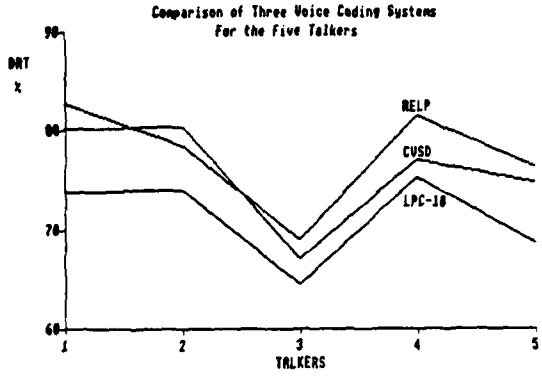
FACTOR	IDENTIFIER	LEVELS
I	INDIVIDUAL	11
T	TALKER	5
R	REPLICATION	2
U	UTTERANCE	2
C	CODER	3
S	LIVE VS MIXED	2

RANDOM EFFECT FACTOR: INDIVIDUAL

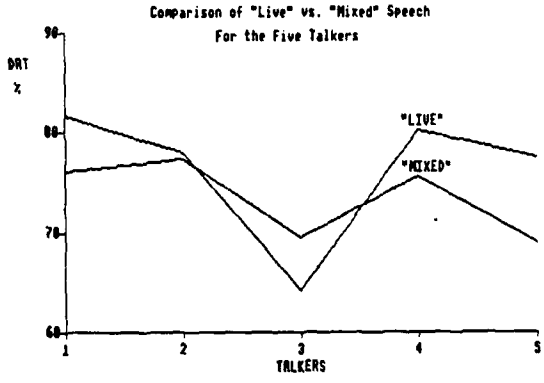
SOURCE OF VAR.	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	ERROR TERM	F VALUE	PROB	
T	26380.7	4	6595.18	TI	112.899	0.0000	***
R	8.52801	1	8.52801	RI	0.228		NS
U	4682.35	1	4682.35	UI	168.395		***
C	9706.63	2	4853.31	CI	100.603	0.0000	***
S	2606.00	1	2606.00	SI	79.655		***
TR	67.9604	4	16.9901	TRI	2.405	0.0651	NS
TU	1526.70	4	381.675	TUI	17.646	0.0000	***
TC	1643.30	8	205.413	TCI	12.038	0.0000	***
TS	7361.60	4	2840.40	TSI	90.206	0.0000	***
RU	0.723371	1	0.723371	RUI	0.055		NS
RC	154.722	2	77.3611	RCI	2.464	0.1080	NS
RS	0.521839E-01	1	0.521839E-01	RSI	0.001		NS
UC	10.5834	2	5.29170	UCI	0.180	0.8373	NS
US	327.704	1	327.704	USI	12.681		**
CS	3206.44	2	1603.22	CSI	96.357	0.0000	***
TRU	64.0995	4	16.0249	TRUI	1.968	0.1175	NS
TRC	171.241	8	21.4051	TRCI	2.640	0.0127	*
TRS	65.5930	4	16.3983	TRSI	1.250	0.3050	NS
TUC	551.458	8	68.9323	TUCI	4.090	0.0004	***
TUS	454.988	4	113.747	TUSI	4.449	0.0045	**
TCS	1077.59	8	134.698	TCSI	7.039	0.0000	***
RUC	10.7611	2	5.38055	RUCI	0.359	0.7016	NS
RUS	12.7443	1	12.7443	RUSI	0.666		NS
RCS	108.542	2	54.2712	RCSI	1.827	0.1835	NS
UCS	26.4013	2	13.2007	UCSI	0.782	0.4677	NS
TRUC	94.8280	8	11.8535	TRUCI	0.809	0.5966	NS
TRUS	40.1148	4	10.0287	TRUSI	0.956	0.4417	NS
TRCS	271.805	8	33.9756	TRCSI	3.351	0.0023	**
TUCS	510.453	8	63.8066	TUCSI	4.237	0.0003	***
RUCS	128.745	2	64.3723	RUCSI	5.390	0.0129	*
TRUCS	137.647	8	17.2059	TRUCSI	1.371	0.2219	NS
TOTAL	97421.2	1319					

FIGURE 7

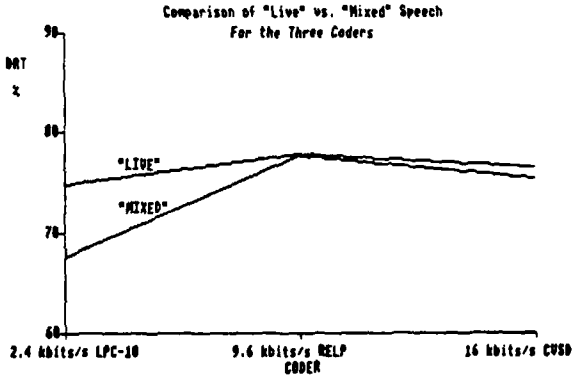
Analysis of Variance results for an experiment to assess the influence of live and mixed speech material on Diagnostic Rhyme Test score.



a) The influence of individual talker on Diagnostic Rhyme Test scores for the three voice coders.



b) Live and Mixed speech recordings analysed by individual talker.



c) The influence of live and mixed speech material on Diagnostic Rhyme Test scores for the three voice coders.

FIGURE 8

Analysis of Variance interactions
DRT SCORE (%)

CODER (k.bits/s)	MIXED	LIVE
2.4(A)	74.8	67.6
9.6	76.5	75.4
16	77.8	77.6
2.4(A)	75.1	67.8
2.4(B)	70.4	67.1

FIGURE 9

Summary of Diagnostic Rhyme Test Scores. When Listeners only are treated as a random effect a difference in scores between coders of 1.1% is significant ($p<0.05$). When both talker and listeners are treated as random effects this value increases to 3.3%. When making comparisons between live and mixed speech the figures are 0.9% and 6.5% respectively. The greater loss in resolution for the live vs mixed comparison is due to the larger T x S interaction term in Figure 7. (A) and (B) are two different implementations of the Nato Standard LPC-10 algorithm.

SPEECH INTELLIGIBILITY QUESTIONNAIRE

Did the task allow a REALISTIC ASSESSMENT of the voice communications system to be made?

Task allowed Realistic Assessment	1	2 ▲	3	4	5	6	7	8	9	10	Realistic Assessment NOT possible

Overall PERFORMANCE of voice communications

Unusable	1	2	3	4	5	6	7	8 ▲	9	10	Excellent

INTELLIGIBILITY of voice communications

Completely Intelligible	1	2	3 ▲	4	5	6	7	8	9	10	Totally Unintelligible

EFFORT required to communicate.

No special Effort required	1	2 ▲	3	4	5	6	7	8	9	10	Extreme Effort required

Did you feel that the amount of effort required COMPROMISED your performance of the task?

Task severely Compromised	1	2	3	4	5	6	7	8	9 ▲	10	Task NOT Compromised at all

Voice QUALITY of incoming communications

Completely Natural	1	2	3	4 ▲	5	6	7	8	9	10	Extremely Degraded

Finally, you are asked to rate the voice communications sytem as either acceptable or unacceptable on the following two counts:

	ACCEPTABLE	UNACCEPTABLE
INTELLIGIBILITY	ALL	
OVERALL USABILITY	ALL	

FIGURE 10

The questionnaire used for the user acceptability experiment. The solid triangle represents the average of 12 responses.

a) Pilot

b) Fighter Controller

FIGURE 11

Pilot and Fighter Controller participating in user acceptability experiments

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 error	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 12

The relationship between DRT scores and categories of voice quality (taken from reference 10)

ANNEX

METHODS OF ANALYSIS OF DRT SCORES

The investigation of differences in DRT scores between N_C conditions tested with N_T talkers and N_L listeners replicated N_R times requires the application of the well known statistical technique Analysis of Variance, in which the total variance of the scores is attributed to a series of pre-defined factors and their interactions. By making suitable statistical assumptions about the nature of the factors and the distribution of the scores to be analysed, tests for the presence or absence of particular effects in the observed data may be constructed. For practical purposes factors may be classified either as fixed effects, for which all possible levels of the relevant factors have been sampled, or as random effects, where the levels of the factors in the experiment represent a random sample from an infinite population. If a factor is treated as a fixed effect, any deductions from the data are only valid for the levels of the factor sampled in the experiment, while, if it is treated as a random effect, inference may then be extended to the parent population.

There are three factors represented in a test of DRT scores: Talker, Listener and Condition. Condition is further decomposed as shown in Figure 2 and may clearly be treated as a fixed effect, while Talker and Listener are best represented as random effects. Results from such an analysis will not be dependent on the particular talkers and listeners used. A representative Analysis of Variance table for a balanced experiment replicated N_R times is given in the Appendix. When all factors involved in an experiment are fixed effects, the test for the presence of a particular effect involves the calculation of the ratio of the mean square due to the effect of interest and the residual mean square

$$F = MS_{Effect}/MS_{Residual}$$

and testing whether this ratio is large using the F distribution. When there are two random effects (T and L), the test of a fixed effect C is provided by the pseudo F ratio

$$F = (MS_C + MS_{CTL})/(MS_{CT} + MS_{CL})$$

where the degrees of freedom for numerator and denominator are derived by Satterthwaite's Method. If a difference between conditions is indicated in the Analysis of Variance, comparison between the individual condition means may then be made using a multiple comparison procedure such as the Newman-Keuls method based on the standard error given in the Annex. Where the different conditions may be labelled by more than one factor, the construction of the appropriate standard error may become more involved, but the principles remain the same.

The standard error for making comparisons between condition means does not represent the absolute accuracy of the estimate of the mean of a specific condition. This standard error involves the variation between talkers and listeners, whereas the former standard error only involves within talker and listener effects. The latter standard error becomes relevant when comparing a set of experimental results with data constructed from a different talker and listener base, and is defined in the Appendix.

STRATEGIES FOR LONG TERM TESTING

During the past three years, two different experimental designs have been employed on different occasions. In the first a panel of eleven listeners has been tested with a complete set of experimental material, while in the second eight listeners out of the eleven have been present at each test session. Since it is simpler to run eight listeners than eleven or twelve, it is worth considering three long term strategies:

(a) Use a single testing room with eight listeners tested on each occasion, five talker tapes and a second replication to ensure that all of a panel of twelve listeners are tested with all the experimental material.

(b) Use two testing rooms with twelve listeners (six per room) tested on each occasion with five talker tapes and a single replication.

(c) Use two testing rooms with twelve listeners (six per room) tested on each occasion with ten talker tapes and a single replication.

Strategy (b) leads to the smallest experiment. Strategies (a) and (c) take the same total amount of time to run as each other, but (c) provides more information than (a). Any decision between these three strategies must be based on the relative effectiveness of the experiments in comparing two different pieces of equipment. The comparison which follows uses the results from the balanced study of eleven listeners conducted during the past twelve months. Three coders were tested with eleven listeners, for two signal-to-noise ratios, two different utterances, five talkers and two replications. The main comparison of interest is that between two different coders. Using the data from the experiment described in section 10 an assessment is made of the effectiveness with which this comparison would be made using the three strategies described earlier.

The first step is to calculate the components of variance which contribute to the standard error for comparing between two coders.

TABLE 1

Estimated components of variance for talker and listener interactions with coder

Component	Value	Estimate
s^2_{CT}	2.140	$(MS_{CT} - MS_{CLT})/N_R\ N_L$
s^2_{CL}	0.779	$(MS_{CL} - MS_{CLT})/N_R\ N_T$
s^2_{CLT}	0.357	$(MS_{CLT} - MS_{Res})/N_R$
s^2	14.210	MS_{Res}

From the components of variance given in Table 1 it is possible using formula A1 to calculate the standard error for the comparison between two coders using the three strategies. Values have to be substituted for N_R, N_L and N_T. For all three strategies $N_L = 12$. For strategy (b) and strategy (c) $N_R = 1$. To a first approximation the value 1.333 may be applied for N_R in strategy (a). For strategies (a) and (b) $N_T = 5$ while for strategy (c) $N_T = 10$.

The values of the standard errors are displayed in Table 2. From these standard errors, it is then possible to calculate the power of a test at the 5% level for prescribed differences in the values of the DRT scores. (The power of a test is the probability that the null hypothesis will be rejected at the given significance level given a prescribed version of the alternative hypothesis). Three differences between DRT scores were considered: 2, 3, 4. Because the degrees of freedom used in a pseudo F or t test are a random variable, these powers were calculated by simulation using 1000 trials, implying an accuracy of approximately 3%.

TABLE 2

Standard errors for the difference between two coders and the powers of a 5% tests.

	Strategy (a)	Strategy (b)	Strategy (c)
Standard Error	1.146	1.213	0.898
Power (Percent)			
D = 2	29.4	26.2	54.6
D = 3	58.0	51.7	88.2
D = 4	82.0	77.4	99.3

It is clear from Table 2 that there is relatively little benefit in following strategy (a) rather than strategy (b), however there is a clear advantage in following strategy (c) relative to either of the other two. The cause of this result is the relatively high component of variance due to the talker x coder interaction (Table 1). If it is necessary to have a better than 50 percent change of detecting a difference between coders as small as 2 units in the DRT scores, the indications of this brief investigation is that ten talker tapes are essential. By examination of the standard errors, it is readily shown that this is by far the most effective method of reducing the standard error to the required level.

APPENDIX TO ANNEX

A notional Analysis of Variance table for a balanced design is displayed in Table 3. The components of variance which contribute to each of the calculated mean squares are displayed in the right hand column of the table. It is assumed that there are N_L Listeners, N_T Talkers, N_R Replications, and N_C Conditions.

TABLE 3

Notional Analysis of Variance Table for a Balanced Design

Source	Degrees of Freedom	Mean Square
Talkers	(N_T-1)	$N_RN_CN_LS^2_T + N_RN_CS^2_{TL} + S^2$
Listeners	(N_L-1)	$N_RN_CN_TS^2_L + N_RN_CS^2_{TL} + S^2$
T X L	$(N_T-1)(N_L-1)$	$N_RN_CS^2_{TL} + S^2$
Conditions	(N_C-1)	$N_RN_LN_TS^2_C + N_RN_LS^2_{CT} + N_RN_TS^2_{CL} + N_RS^2_{CTL} + S^2$
C X T	$(N_C-1)(N_T-1)$	$N_RN_LS^2_{CT} + N_RS^2_{CTL} + S^2$
C X L	$(N_C-1)(N_T-1)(N_L-1)$	$N_RN_TS^2_{CT} + N_RS^2_{CTL} + S^2$
C X T X L	$(N_C-1)(N_T-1)(N_L-1)$	$N_RS^2_{CTL} + S^2$
Residual	$(N_R-1)(N_TN_LN_C-1)$	S^2

The standard error for the difference between two conditions S_{**} is calculated from the following equation:

$$S^2_{**} = 2\ (S^2_{CT}/N_T + S^2_{CL}/N_L + S^2_{CTL}/N_LN_T + S^2/N_LN_TN_R) \qquad \text{(A1)}$$

It is important to note that even if N_R becomes very large, the standard error may not be reduced below a value determined by the number of talkers and the number of listeners. If S^2_{CT} is not small, the number of talkers may become the most important determinant of the size of the standard error. If the number of talkers is five, the standard error of the difference of two conditions is always greater than 0.925, regardless of how many listeners or replications are employed.

From Table 1 it may be seen that an estimate of the right hand side of equation A1 is provided by the form

$$S^2_{**} = 2\ (MS_{TC} + MS_{LC} - MS_{TLC})/N_TN_LN_R$$

where the MS_{XX} refers to the mean square due to the effect XX in the Analysis of Variance. This expression is a composite of mean squares, and the calculation of approximate degrees of freedom should be made using Satterthwaite's method.

The overall standard error of the mean of all conditions, S_* is given by the relation

$$S^2_* = S^2_T/N_T + S^2_L/N_L + S^2_{LT}/N_L N_T + S^2/N_T N_L N_R N_C$$

and this quantity may be estimated from the composite

$$S^2_* = (MS_T + MS_L - MS_{TL})/N_T N_L N_R N_C$$

The overall standard error of a specific condition, S_{***} is itself given by the expression

$$S^2_{***} = S^2_* + \tfrac{1}{2} S^2_{**} (N_C - 1)/N_C$$

DOCUMENT CONTROL SHEET

Overall security classification of sheet .....UNCLASSIFIED.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	REPORT 87003		U/C

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
778400	RSRE, ST ANDREWS ROAD, MALVERN, WORCS WR14 3PS
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

ASSESSING THE INTELLIGIBILITY AND ACCEPTABILITY OF VOICE COMMUNICATION SYSTEMS

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials	9(a) Author 2	9(b) Authors 3,4...	10. Date	pp. ref.
PRATT, R.L.	FLINDELL, I.H.	BELYAVIN, A.J.	1987.06	VP

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

Descriptors (or keywords)

continue on separate piece of paper

Abstract A facility for quantifying the speech intelligibility of voice communication systems using the Diagnostic Rhyme Test has operated continuously at the Acoustics Laboratory of the Royal Signals and Radar Establishment since February 1985.

User acceptability trials that enable Service personnel to operate, and then assess, voice communication systems under simulated operational conditions have also been conducted.

This report describes the procedures used to assess both intelligibility and acceptability, and presents the results of studies investigating the use of digital vocoders in high noise environments.

An Executive Summary is provided to give project offices and others responsible for designing, specifying or procuring voice communications systems (and components) an indication of the services that are available.

S80/48

END

DATE
FILMED

8 88