Proceedings of the Institute of Acoustics

SOME RESULTS AND EXPERIENCES FROM SUBJECTIVE SPEECH INTELLIGIBILITY TESTS

P.W. Barnett AMS Acoustics, Rayleigh House, Bush Hill Park, EN1 2QB.

1. INTRODUCTION

Over the past 15 years AMS Acoustics has been involved in the science of speech intelligibility and in particular that associated with electro-acoustic systems. Often circumstances dictate that objective measures such as RASTI or STI are not applicable or would not accord with a subjective impression. This being the case either some adjustment has to be made or it is necessary to return to the more traditional methods of subjective testing.

Borne out of necessity we have developed and refined our techniques to improve the 'consistency' and precision of this type of measurement.

With the publication in 1998 of BSEN60849 and the introduction of the Common Intelligibility Scale (see fig. 1), the door is open to use a number of methods to measure intelligibility including several subjective based methods.

This Paper provides an insight to some of the basic issues and pitfalls of subjective testing.

2. SCENARIO 1 - THE LEARNING CURVE

An essential part of subjective testing is the training and familiarisation process given to the listeners or jury. Training comprises familiarisation with the task at hand, the words and the test methods. In spite of training, hands-on experience with the tests is important and to this end jurors and listeners cannot be used for the main task until they have completed some control tests and their performance monitored.

The control tests are tests which we have used before and to which we know the answer. Control testing normally lasts 1 x day during which we get through around 30 x control tests. Most listeners/jurors peak within this period, some need a little longer.

Fig. 2 shows the tests for various listeners/jurors together with data after Egan.

It can be seen that, firstly in spite of a formal training programme, the listeners/jurors do take time to come up to speed and hence, tests involving taking persons straight off the street should be viewed with caution. It is also worth noting that even persons who have extensive experience still require some practice even after a relatively short lay off.

3. PANIC ATTACK

Clearly the act of participating in and listening to word tests requires some concentration and not surprising if this concentration is interrupted then the results might suffer. This scenario involves an experienced juror who for a short time lost concentration. The juror, a female in her forties, was experienced and was classified as an A/B.

Note: In order to maintyain consistency, we categorise and continually monitor listeners'/juirors' performance. Our scale is as follows:

Category	Consistent Score Relative to the average
A	>+2
B	+1 to +2
C	-1 to +1
D	-2 to -1
E	<-3

Fig. 3 shows the scores of the juror in question over the day. It can be seen that her score relative to the average took a decided dip at 15.00 hrs. Our control procedure picked this up and we discovered that as a result that the tests were running late, she was concerned that she would be late to retrieve her children from school. This issue has two-fold implications, firstly the obvious - potential errors in word score testing but secondly and perhaps more important, might this not reflect real life? A person might be walking through a shopping centre trying to decide what to have for tea, would they hear an emergency announcement and comprehend the message in the same way as a person whose mind was less occupied?

4. SCENARIO 3 - A LACK OF CONCENTRATION

The act of understanding a communication from another party is an extremely complex affair. It requires that the recipient recognises the words and this might involve disseminating the information in the presence of noise or reverberation and it also might require a degree of concentration or mental effort.

This effort can be inferred from fig. 4.

In this figure we can see the effect of plotting word score results against an objective measure, say STI. The centre portion of the graph which we might describe as the proportional region which is self-explanatory. The low transition region can be thought of as those conditions interfering with speech are so extreme that the recipient is unable to determine what is being said. The upper region probably results from concentration or mental effort. We can imagine that in this transition region as the tests become easier the required concentration reduces.

During one of our testing sessions, a new juror, not previously known to us was trained in the usual way. It quickly become apparent that this juror fell into the E category since she always scored less than her fellow jurors. More than this, as the tests became more difficult, her scores were proportionately worse. Fig. 5 shows the results from one series of tests.

Upon further investigation we discovered that the juror, although keen and motivated to the task, was a poor achiever at school and did suffer a degree of learning difficulty which, according to a relative, was due to a reduced attention span and the inability to concentrate.

Whether her low scores was due to reduced concentration or not would require a more detailed study and indeed with many more subjects but this hypothesis is to some extent supported by fig. 6 which plots the results against RASTI measurements made at the same time as the lists were recorded.

It can be seen that the characteristic recognition 'knee' is not present in the use of the juror in question. Similar traits were observed from other tests in the same series.

5. SCENARIO 4 - UNTRAINED TALKER

Most of the juror scenarios came about by accident or as a result of controls we introduced. This next experiment was deliberate. For our tests we use trained professional talkers, persons whose diction and articulation tend towards perfection. We decided to determine if reduced diction would produce a significant difference.

For our untrained subject we chose a young lady who is a teacher (not of speech) and who therefore is used to communicating. The young lady is well educated and speaks well with a slight North London accent. In social circles she would be categorised as well spoken.

We gave her training in the task at hand i.e. the rate of delivery and the need to maintain a constant level. We gave no instruction or made no observations in regard of her diction nor the formulation of the words.

She was unaware of the intent of the experiment. Her recordings were subject to the same controls and processing as applied to the professional talkers.

Two sets of recordings were made at two positions in a reverberant space (RT circa 2.0 seconds), broad band noise was added prior to presentation to a jury of listeners.

The results are shown in fig. 7.

Clearly it can be seen that diction has a profound effect on the outcome of word score tests.

6. SCENARIO 6 - T00 HARD

When word lists are presented to listeners they are generally in groups of 3, hence each 'session' takes around 15 mins. This is then followed by a 5 min. rest. Within any group it is possible that sets might be in the range 'easy' (scores 80%) to 'hard' (scores <40%). We have found that if the three sets are hard-hard-hard then the final two sets score worse than they should have, probably caused by fatigue. This is demonstrated in fig. 8. In practice therefore, tests need to be staggered as easy-hard-easy or hard-easy-easy.

7. DISCUSSION AND CONCLUSIONS

From the foregoing it is evidenced that subjective testing is not without its difficulties which could catch out the unaware.

Notwithstanding this, there is ample evidence that objective methods are also liable to error and in this regard the problem is insidious since it is possible to be blissfully unaware that the measurement made is not in accord with reality. We should not forget that both RASTI and STI correlates to intelligibility through word scores, hence both objective and subjective methods should be in accord. If they are not, then the problem might lie in either camp.

Subjective Speech Intelligibility Tests - P.W. Barnett

The issue is that BSEN60849 allows a variety of methods to predict or audit speech intelligibility. Not only should the chosen method be applicable it must be carried out with rigour lest the results, upon which much might depend, could be flawed.

This Paper has focused on some problems with subjective testing. The list is obviously not exhaustive but it is hoped that it might serve to alert those concerned to possible types of errors that might distort the outcome. Neither is it intended that the focus of attention should be directed towards the fragility of subjective testing, objective methods are just as vulnerable.

Perhaps the message should be - whatever method is chosen - apply rigour.

Subjective Speech Intelligibility Test - P. W. Barnett

Relationship Between Intelligibility Scales

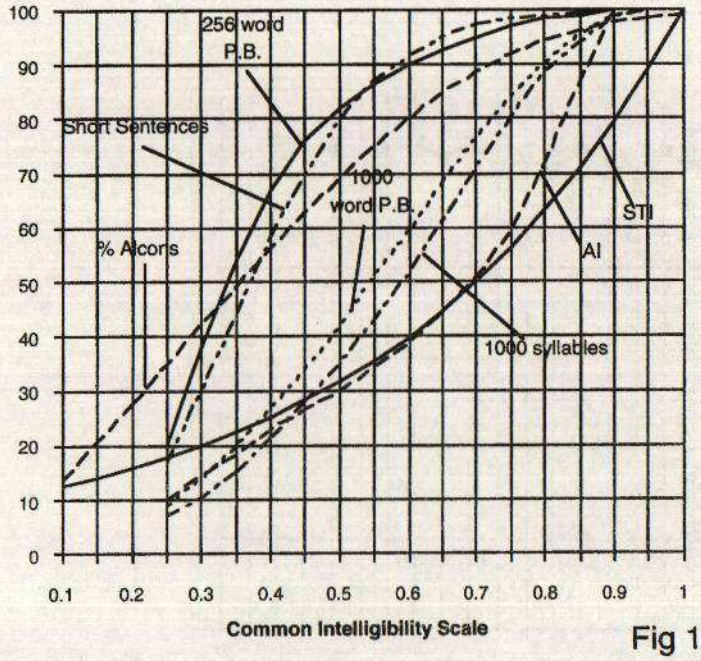

Fig 1

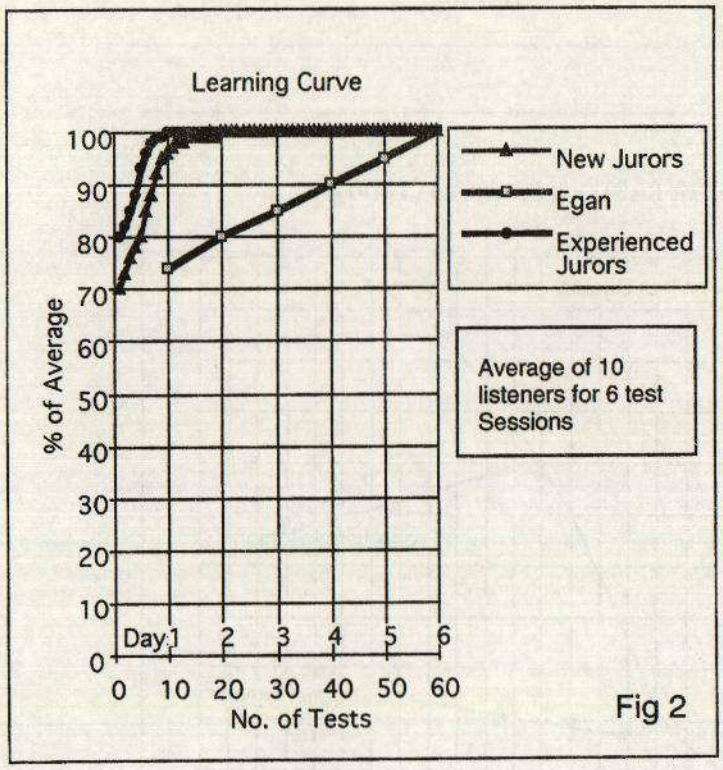

Fig 2

Graph of a Listeners Score as Average of all Listeners

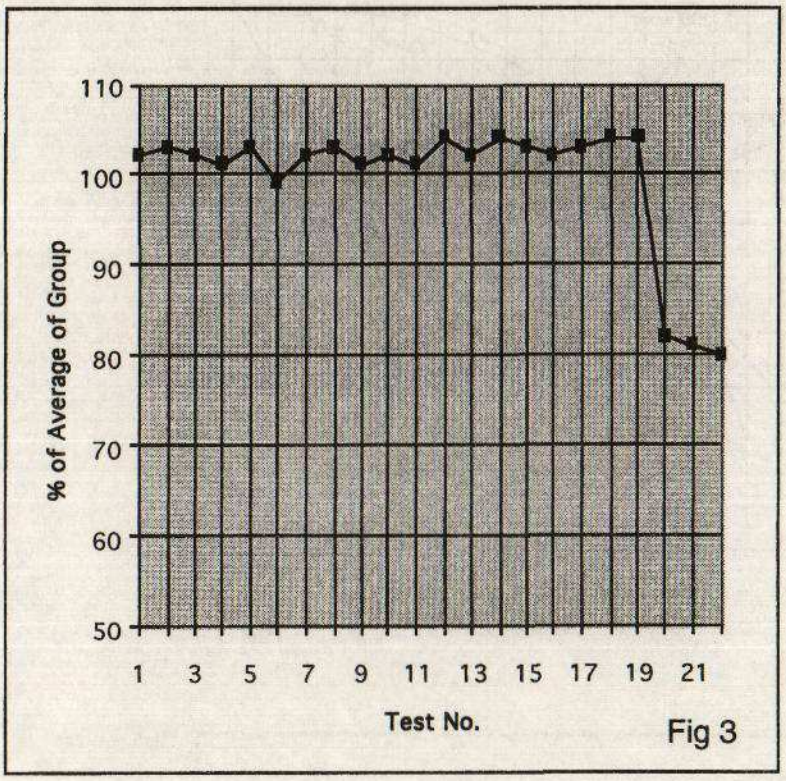

Fig 3

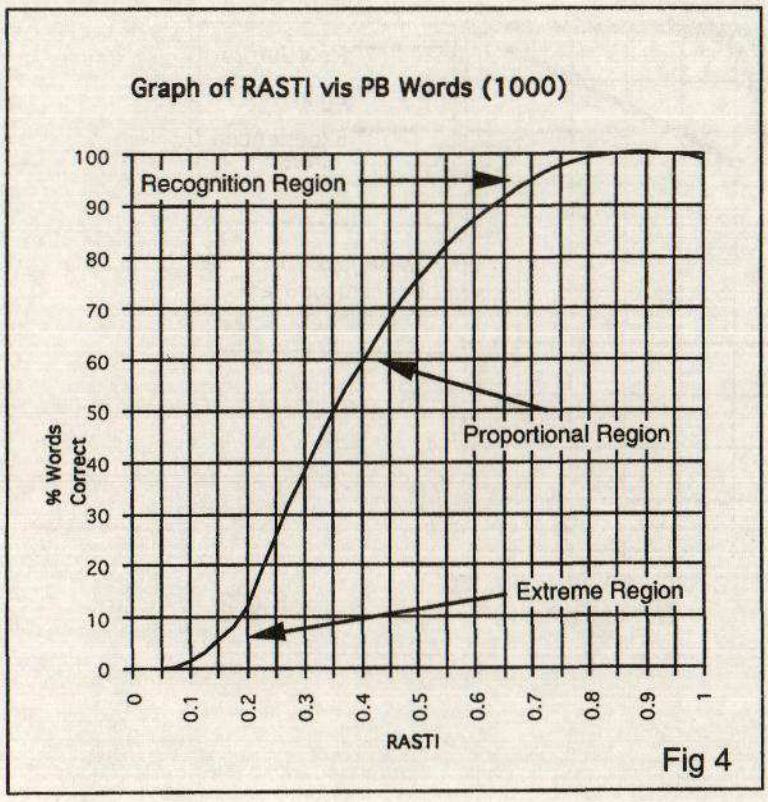

Fig 4

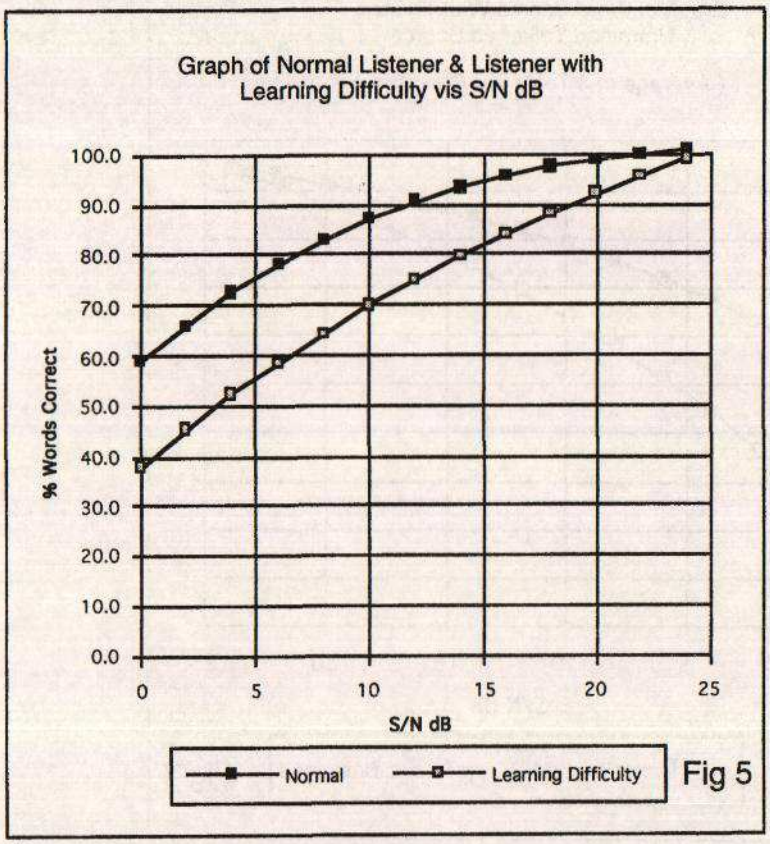
Graph of Normal Listener & Listener with Learning Difficulty vis S/N dB
% Words Correct
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
5
10
15
20
25
S/N dB
Normal
Learning Difficulty

Fig 5

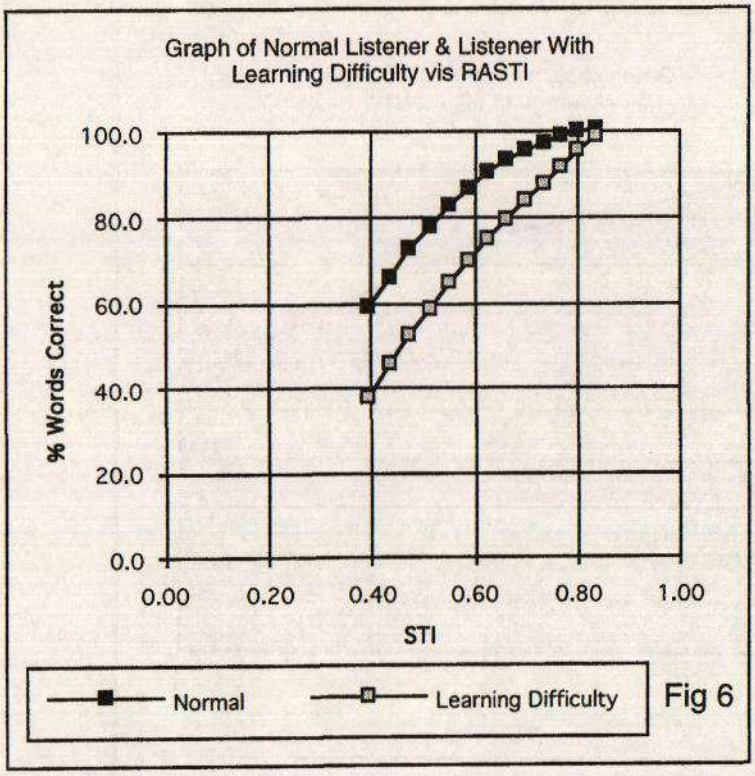
Graph of Normal Listener & Listener With Learning Difficulty vis RASTI
% Words Correct
100.0
80.0
60.0
40.0
20.0
0.0
0.00
0.20
0.40
0.60
0.80
1.00
STI
Normal
Learning Difficulty

Fig 6

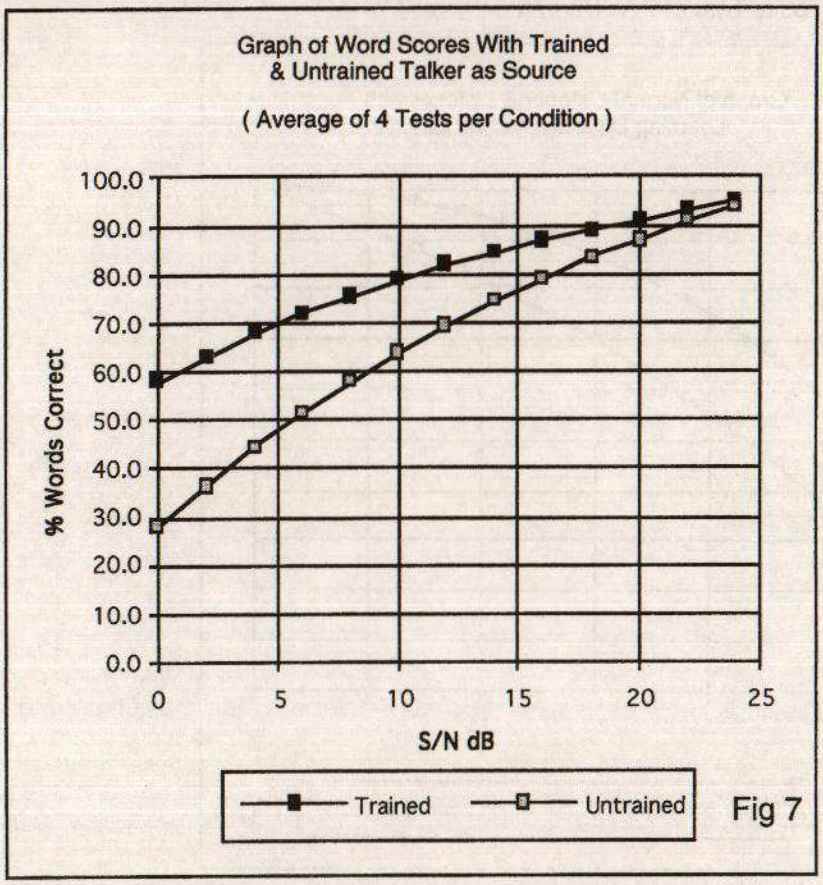
Graph of Word Scores With Trained
& Untrained Talker as Source
(Average of 4 Tests per Condition)
% Words Correct
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
5
10
15
20
25
S/N dB
Trained
Untrained

Fig 7

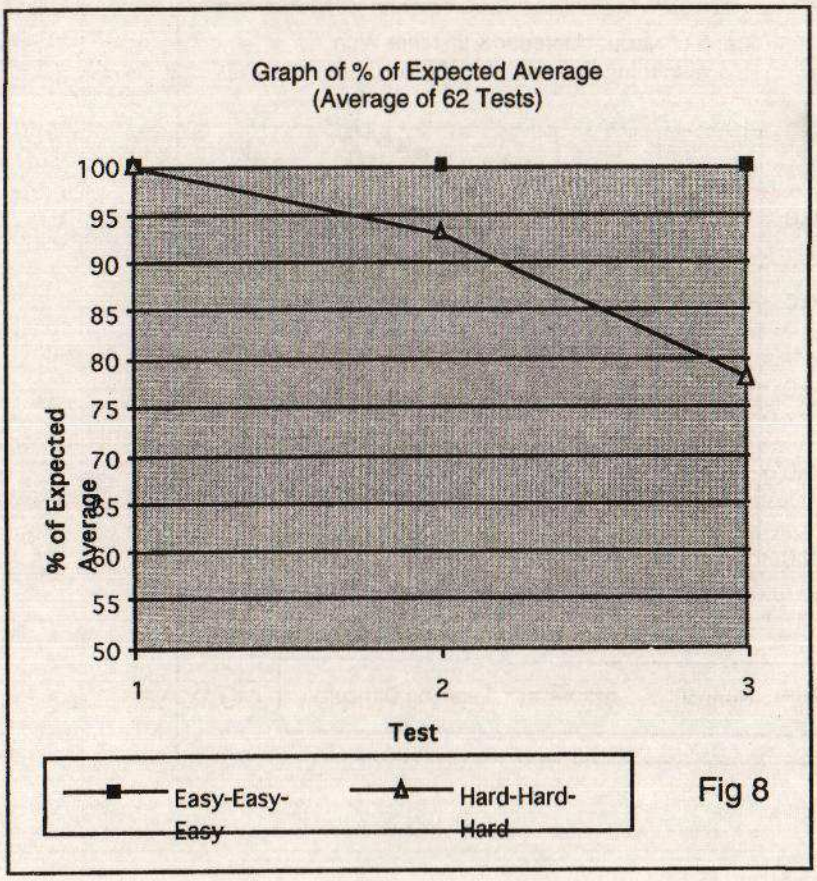
Graph of % of Expected Average
(Average of 62 Tests)
% of Expected Average
100
95
90
85
80
75
70
65
60
55
50
1
2
3
Test
Easy-Easy-Easy
Hard-Hard-Hard

Fig 8