Proceedings of the Institute of Acoustics

THE USE OF BINAURAL RECORDING TECHNIQUES IN THE ASSESSMENT OF SPEECH INTELLIGIBILITY USING WORD SCORES ON LONDON UNDERGROUND

P.W. Barnett

AMS Acoustics, 52 Chase Side, Southgate, London, N14 5PA.

INTRODUCTION

This Paper draws on a small fraction of a speech intelligibility and acoustic investigation into problems encountered in the London Underground System. This Paper is concerned with the assessment of speech intelligibility using Word Scores binaurally recorded.

Most of the work was concentrated on Phonetically-Balanced (PB) Word Scores, although Modified Rhyme Tests (MRT) and sentences were also employed.

PREAMBLE

There were three primary reasons for recording the Word Scores:

1. Firstly London Underground have strict safety rules regarding contractors and the like on platforms. Each person permitted to work on a platform has to undergo a half-day Station Awareness Course.

2. Secondly, some 800 tests were required and since conditions on the platform are not ideal we felt that fatigue and other environmental factors may influence the results.

3. Thirdly, we were interested in the effect of noise and hence, by recording the scores, it would be possible to post-process the data with the addition of station occupancy noise.

It may therefore be readily understood that conducting the tests 'live' for all conditions would present a formidable and lengthy task.

Recordings and analysis by jury seemed to be the best option.

We used two recording techniques:
(i) Stereo
(ii) Binaural.
Both committed the results to DAT tape.

THE SITE

The investigation was carried out at the Aldwych Station. This station has two advantages:

(i) It is not used between the hours of 10 am and 4 pm.

(ii) It was mid-way through a renovation programme when the work was halted. This resulted in half the platform which has tiles and is reverberant and the other half has had the tiles removed and is acoustically dead.

The acoustics of this platform proved extremely interesting. The live end had a mid-frequency reverberation time of around 2.5 sec. and the dead end 1.2 sec.

MEASUREMENTS

The investigation involved both Word Scores and STI/RASTI measurements, in both the **live** and **dead** end of the platform ON and OFF axis of the loudspeakers.

Two Word Score lists were read for each position and for each condition.

At each position both stereo and binaural recordings were made. The STI measurements were also recorded (monaurally) for post-processing.

The word lists were scored live under noiseless conditions by a listening crew comprising 6 persons.

These results were subsequently compared with the jury results as a control.

POST-PROCESSING

The Word Score recordings were subsequently mixed with actual recorded station noise at varying signal-to-noise ratios.

Both the station noise and the Word Scores were variable in level and therefore an average signal-to-noise ratio was taken. Each recording was also output to a paper chart to verify the signal-to-noise ratio and also to allow subsequent analysis.

WORD SCORES

Three types of tests were used:

Phonetically-Balanced (PB) Open Sets
Each test comprises 25 words drawn at random from a population of 300.

Modified Rhyme (MRT) Closed Sets

Each test comprised 25 words grouped into nests of 5 from a total population of 300.

Sentences

Two tests were involved using sentences:
(a) Sentences containing numbers, where the listener was asked to write down the number heard.
(b) Sentences where the listener was asked to score on a scale of 1-5 the ease by which the content was understood.

ENUNCIATOR, LISTENING CREW AND LISTENING JURY

The tests were carried out and calibrated in general accordance with ISO TR4870 and ANSI S3, 2-1989.

Each person was given a simple audiometry test prepared by ourselves. The test did not comprise a clinical test and would not meet any of the required standards. It was designed simply to ensure that the listeners had no major hearing defect and that their hearing extended to at least 8kHz.

Enunciator

The enunciator had undergone speech training and is a professional actress. Specific training was given regarding the method of delivery, timing and the carrier sentences. She was also responsible for the provision and construction of the sentences.

Listening Crew

The listening crew comprised the engineers and administration staff (six in all) assigned to this project. All the listening crew were given instruction, most of the crew had previous experience of these type of tests.

Listening Jury

The listening jury comprised an amateur theatrical group. They had no previous experience of these type of tests and instruction was given including live tests by our enunciator. The jury members were in the age group 21-47 and had a good split male and female.

Most of the jury members native language was English although two were bilingual (Greek/English).

All jury members were considered competent at the conclusion of the training. all audiometry tests were satisfactory. None of the jury members attended a discotheque the night before nor did they listen to a personal stereo prior to the tests.

Binaural Recording System

The binaural recording system was developed in West Germany for Mercedes Benz the motor car manufacturer.

The 'Head' is used for jury panel tests regarding internal car noise. Considerable care was taken to ensure the perception is close to that of a human.

This equipment was flown in specially from the USA with a skilled operator for the purpose of these tests.

Presentation to the jury members was over high quality Senheisser Type HD 56011 headphones.

METHOD - WORD LISTS

The enunciator read each word list into the main system microphone. The listening crew and the recordings were made simultaneously and conducted in both live and dead ends of the station, for both ON and OFF loudspeaker axis positions. Two word lists of each type at each position were used.

The binaural recording system was positioned at a normal traveller position facing away from the loudspeakers. The listening crew were positioned at various positions ON and OFF axis, the crew changed positions at each alternate test.

METHOD - STATION NOISE

Material was selected from the monaural recordings made during our station surveys and was re-recorded binaurally.

The material used was selected to be representative, fairly steady and should not contain any items which would be likely to distract the jurors i.e. passages of identifiable speech were rejected. Train noise was not included.

METHOD - RECORDING

The recordings were made with the equipment set-up as shown in fig.1. The actual signal-to-noise level was deduced from the chart recordings made with the DAT recordings. Samples of these charts are shown in fig. 2.

The recordings were made with due regard to the ensuing jury tests.

No juror would hear the same word set more than once and they would all be subjected to both ON and OFF axis material for varying amounts of noise. Ten tapes were made in all.

METHOD - REPLAY TO JURORS

The Eureka Theatre Company supplied a room which was laid out as shown in fig. 3. Each juror position was screened from the next. The jurors listened to the recorded material through headphones.

The noise level in the room was kept low and there was little noise intrusion from the outside.

The twelve jurors were arranged into groups X, Y and Z. Once they were assigned to a group they remained with it for the duration of the exercise. Each was assigned an identification number within the group i.e. X1, Y3, Z2 etc., and they kept this number also.

Jury panels were rotated and no panel completed more than four lists (listening approximately 9 minutes) without a rest.

Regular general rests were also had to avoid any fatigue which might affect the results. Jurors were not advised of the score to remove any element of competition which might have affected the results. The tests were conducted over a two-day period.

DATA ANALYSIS

The Word Score sheets were marked to determine the number of correct answers$^{(R)}$, the number of incorrect answers$^{(W)}$ (including words missed) and the total test population.

The total population was determined from

$N = R + W$

where: N = total population
R = number of correct answers
W = number of incorrect answers.

Items clearly identified by the jurors as a distraction were removed from the population.

The %score was calculated, as prescribed by ISO TR 4870, from:

$$\%INT = \frac{100}{T}\left(R - \frac{W}{(N-1)}\right)$$

where: %INT = % Intelligibility
T = set population
R = correct answers
W = incorrect answers
N = number of prescribed alternatives.

The quotient $\frac{W}{(N-1)}$ is a correction factor to remove the element of chance.

RESULTS

Station Tests

A cursory examination of the test measurements indicate that one of the listening crew consistently returned a low score, this member of the crew was the American who may have consistently scored low for a number of reasons, vis:

(a) Failure to appreciate different pronunciation.

(b) Preoccupation with binaural recordings.

(c) Jet lag (he only arrived in the UK 24 hrs. prior to the test).

It was therefore decided to remove the scores of this listener from the test results.

The following tables give the summarised modified results of both the MRT and PB Word Scores.

Test Type MRT	Scores	
Condition	%INT.	SD σ
ON -Axis Live End	96.6	6.4
OFF-Axis Live End	98.6	2.3
ON-Axis Dead End	97.2	4.8
OFF-Axis Dead End	97.6	3.6

Test Type PB	Scores	
Condition	%INT.	SD σ
ON -Axis Dead End	92.8	4.1
OFF-Axis Dead End	88.8	7.3
ON-Axis Live End	83.8	7.9
OFF-Axis Live End	83.4	6.6

Clearly, there is a marked difference between the MRT and PB results.

Comparing ON and OFF-axis MRT results shows little difference and we believe this is due to the apparent insensitivity of the tests in this range (see fig. 4).

The PB Word Scores however do appear to be more sensitive and hence probably a more reliable measure.

Comparing ON and OFF-axis measurements shows an improvement (ON-axis) of between 2 and 4%.

Validation of the Jury Method

Since, as indicated earlier, it was not practicable to carry out the full set of tests on the platform, the jury method was proposed. It is necessary to validate this method. this may be done by comparing the station results (no noise) with the jury results (no noise).

These are presented in the table below:

Test Type MRT	Station		Jury		Difference
Condition	%INT.	SD σ	%INT.	SD σ	%INT.
ON-Axis Dead	96.6	6.4	100	0.0	-3.4
OFF-Axis Dead	98.6	2.3	98.8	2.4	-0.2
ON-Axis Live	97.2	4.8	98.4	2.8	-1.2
OFF-Axis Live	97.6	3.6	98.8	2.4	-1.2

Test Type PB	Station		Jury		Difference
Condition	%INT.	SD σ	%INT.	SD σ	%INT.
ON-Axis Dead	92.8	4.1	94.0	2.0	-1.2
OFF-Axis Dead	88.8	7.3	89.0	5.2	0.2
ON-Axis Live	83.8	7.9	79.9	10.4	3.9
OFF-Axis Live	83.4	6.6	83.3	7.6	0.1

With the exception of the ON-axis live (PB) results, the jurors consistently scored better than the station tests. No specific reason can be proffered however, we are of the opinion that the jurors concentration would have been higher since they were in a more 'classroom' situation.

With regards to the ON-axis live results, it can be seen that the standard deviation was quite large i.e. 10.4. this was caused by one juror who scored particularly badly on this tests. If his results were removed, then the % intelligibility rises to 83.9% which is extremely close to the station results.

It is interesting to note that the live end results, after modifying the jury test ON-axis, are much closer than the dead end results. Once again, no reason can be given.

Validation of the Stereo Recording Method

At the time the binaural recordings were made, stereo recordings were also made.

The table below compares the two methods (noiseless conditions):

Test Type MRT	Station		Jury		Difference
Condition	%INT.	SD σ	%INT.	SD σ	%INT.
ON-Axis Dead	100	0.0	100.0	0.0	0
OFF-Axis Dead	98.8	2.4	100.0	0.0	-1.2
ON-Axis Live	98.4	2.8	100.0	0.0	-1.6
OFF-Axis Live	98.8	2.4	98.8	2.4	0

Test Type PB	Station		Jury		Difference
Condition	%INT.	SD σ	%INT.	SD σ	%INT.
ON-Axis Dead	92.8	2.0	88.0	7.3	-7.0
OFF-Axis Dead	88.8	5.2	85.9	5.1	-4.1
ON-Axis Live	83.8	8.0	87.0	5.1	+3.1
OFF-Axis Live	83.4	7.6	89.5	2.4	+6.2

The difference between the binaural and stereo methods using MRT lists is extremely small and we believe is due to the reduced sensitivity of the MRT method in our situation.

The PB Word Lists however give a significant difference in the dead end in favour of the binaural method and a significant difference in favour of the stereo method in the live end. For comparison purposes the data is presented graphically in fig. 5.

We can only really conjecture regarding the reasons. Firstly the standard deviation of each set is fairly consistent and hence we are not of the opinion that the tests in question contain a spurious result. The only reason that can be offered is that the stereo method benefits the reverberant conditions by not repeating the conditions exactly. The dead end, where the directional properties of the ear are most important, demonstrate that the binaural method gives the better results.

We are therefore of the opinion that whilst the stereo method does give plausible results, the binaural method is to be favoured.

SUMMARY METHODS AND TESTS

At this stage we are able to suggest that MRT tests are not sufficiently sensitive in our range of interest to provide useful data and hence although considerable data was collected and analysed, little further mention of these tests will be made.

We are also able to be reasonably confident that the binaural method is a valid approach.

THE EFFECT OF NOISE ON PB WORD SCORES

As stated earlier, each binaural station recording was adulterated with station-type noise.

PB Word Score Summary Tests Results

S/N Ratio dB	ON/D	OFF/D	ON/Rev.	OFF/Rev.
0	67	57	60	36
6	69	64	65	57
12	85	79	78	77
18	89	88	86	74
>20	94	89	80	83

The data is presented in graphical form in fig. 6.

Firstly, it can be seen that Word Score reduces with reducing signal-to-noise ratio and further that when signal-to-noise ratio is less than 12dB the reduction can be considerable.

The difference between ON and OFF-axis in the dead end is marginal. However, in the live end the difference is considerable. This further demonstrates the importance of using loudspeakers with controlled-dispersion to ensure that most listeners are ON-axis.

RESULTS - SENTENCES

The results of the sentence tests gave one surprising and one extremely useful result.

The surprising result was concerned with our ability to extract a number from a sentence. The numbers were always in context and never exceeded five digits. The results are as follows:

S/N Ratio dB	% Correct
>20	100
12	99.6
0	100

Hence it can be seen that this method is extremely insensitive to noise and therefore we do not consider this a reliable test.

The interesting result is in connection with our sentence tests.

In these tests, the jurors were asked to score on a scale of 1 to 5 the ease by which they were able to understand what was being said. No formal training was given.

The results are presented in the table below:

S/N Ratio dB	Condition			
	Dead		Live	
	ON-axis	OFF-axis	ON-axis	OFF-axis
>20dB	4.8	4.6	3.8	3.8
6dB	3.7	3.1	3.3	3.0

Standard deviations rarely exceed unity in spite of the subjective nature of the test.

It can be seen that both noise and reverberation have a dramatic effect on the degree of difficulty in understanding the sentence content.

It was also interesting to note that when the number sentences were independently scored on the ease of understanding basis, they scored badly even though the numbers were invariably scored correctly. Once again highlighting the insensitive nature of the number tests.

Finally, since the sentence tests were conducted under the same conditions as the Word Score tests, it is possible to correlate the sentence test with the Word Scores and these results are shown in fig. 7.

We are of the opinion that the scatter is quite small given the nature of the test.

CONCLUSIONS

We believe the London Underground intelligibility investigation has demonstrated the validity of the binaural/jury approach. We have used a similar approach on other projects again to good effect.

Word Score testing methods do have their drawbacks and are particularly sensitive to influences which do not form part of the test. We believe that fatigue and environment are good examples. The jury method overcomes this. Considerable care however still needs to be taken.

The jury method also allows for retesting in the event of an anomalous result.

We were concerned regarding the apparent improvement over reverberant conditions achieved by stereo recordings although the exact mechanism is not clear. We are presently conducting further work into this aspect.

We are also of the opinion that the Ease of Understanding (EOU) method is worthy of further examination.

Post Processing Equipment Set Up

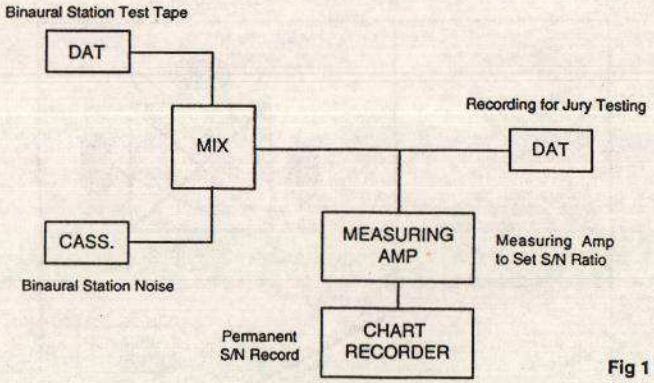

Fig 1

Sample Noise & Speech Charts

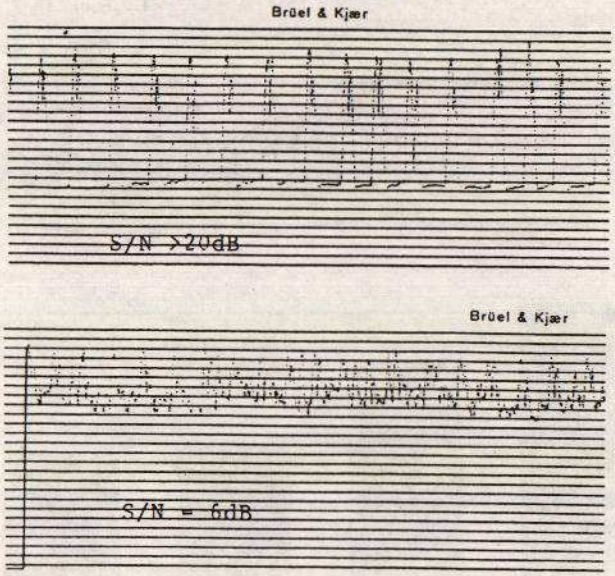

Fig 2

JURY ROOM SET UP

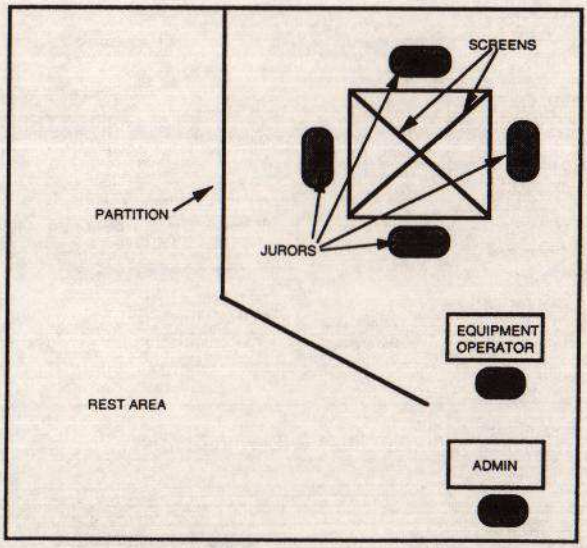

Fig 3

COMPARISION OF PB & MRT STATION TESTS

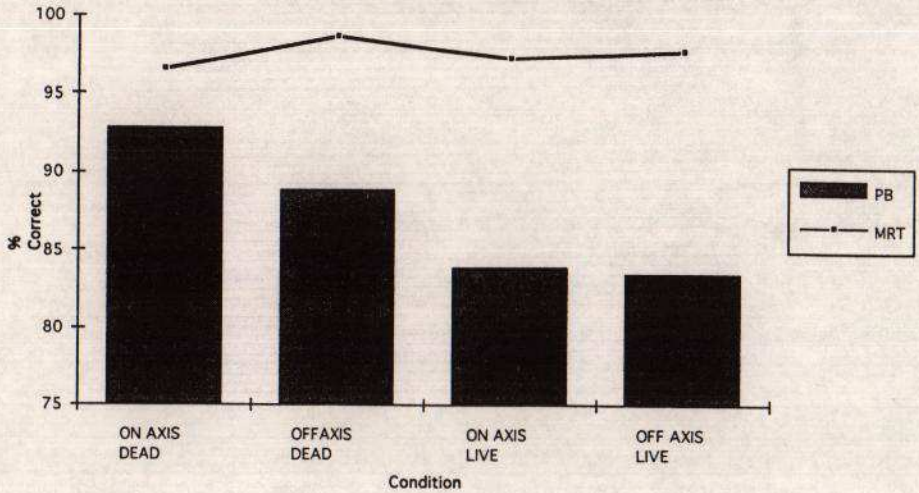

Fig 4

STATION TESTS vis. BINAURAL & STEREO RECORDINGS

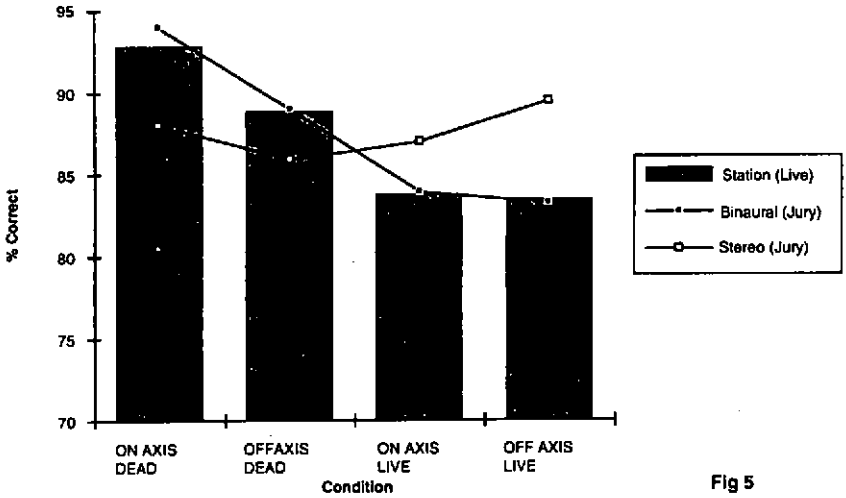

Fig 5

EFFECT OF NOISE ON PB WORD SCORES

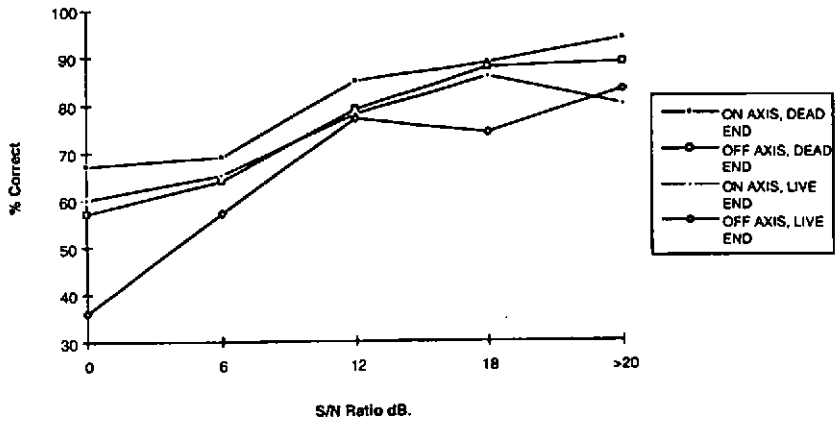

Fig 6

GRAPH OF % INT AS SCORED USING THE PB WORD METHOD
vis
EASE OF UNDERSTANDING (EOU) USING THE SENTENCE METHOD

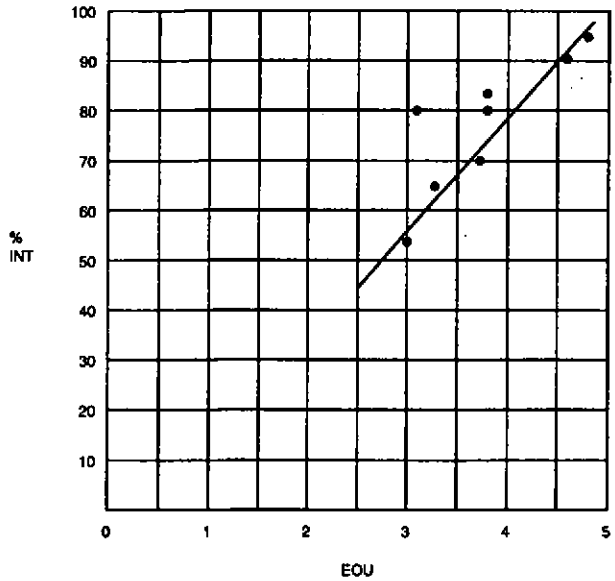

Fig 7