Proceedings of the Institute of Acoustics

SUBJECTIVE SPEECH INTELLIGIBILITY IN PRACTICE

H.M. Goddard AMS Acoustics, Rayleigh House, Bush Hill Park, EN1 2QB.

1. INTRODUCTION

Subjective intelligibility testing has, in my opinion, gained an unfair reputation for inconsistency with an emphasis on estimation rather than measurement.

In fairness it is not difficult to understand why this has come about since clearly we are using individuals rather than a piece of inanimate electronics or some electro-mechanical device.

The vagaries of humans apart, it would seem that a subjective method is preferable since the parameter under measurement is subjective.

Clearly any objective method is at least one step removed from the measurement parameter and indeed the present preferred objective method STI (RASTI) has been shown to relate to speech intelligibility via subjective testing methods.

if we briefly list those issues that affect communication between a talker and listener we can appreciate the difficulty with objective methods.

1. Acoustics of space
2. Acoustic anomalies in space
3. Frequency response
4. Distortion
5. Noise
6. Talker proficiency
7. Listener proficiency

Clearly to satisfy items 1-7 is a tall order for objective methods.

In fairness subjective testing is not without its problems and this Paper is concerned with reducing extraneous effects that produce inconsistency, variability and uncertainty in the results.

Clearly it is important to strive towards a talker-listener proficiency of unity, thereby the results are directly attributable to the parameter under test.

In order to truly 'measure' speech intelligibility using a subjective test, all possible variables have to be accounted for within the control of the method. If any variable is allowed to influence this situation under test, the data becomes a measure of an entirely different subject matter. Whilst this is the common problem with all test methods, subjective intelligibility is a veritable minefield.

The question is then, what parameters are we trying to quantify through our subjective testing? Subjective speech intelligibility testing through Phonetically Balanced Word Scores (closed sets) are used to assess the performance of a system If the variables are controlled PB words can be used to 'measure' the speech intelligibility of a system. The system under test is the PA system chain and the space in which it is located.

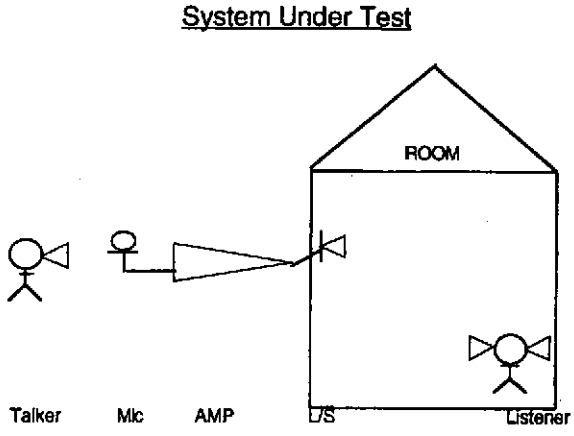

Fig.1

The PB Word Score test then audits all of the seven components previously listed as affecting communication between a talker and listener.

1. Acoustics of the space - the input speech signal interacts with the space via the loudspeaker.

2. Acoustic anomalies also will interact with the speech signal in a 'real' way such that degradation may occur due to masking in the case of a flutter echo.

3. The frequency response of the system is also tested since the speech has passed through the signal chain.

4. The speech received by the listener will have the systems distortion characteristics inherent in it. The importance of this being that other objective methods take no account of distortion which in reality affects speech.

5. The noise that may be typical in the space can be present at the time of the tests or added at a later post-processing stage to replicate a set of noise circumstances.

6. Talker proficiency is also measured - that is the quality of the original speech signal. Talker proficiency adheres to the old 'garbage in - garbage out' law.

7. Listener proficiency also superimposes it's effects onto the results. This proficiency is a function of the subjects training, willingness, state of mind and hearing.

In the subjective testing carried out by AMS Acoustics items 5, 6 & 7 are carefully controlled through developed processes. The Word Scores are recorded binaurally in the space under quiet conditions i.e. S/N ratio >+25dB. Calibrated noise is added during post-processing at a known level. A talker proficiency of unit is maintained using an excellent talker in an anechoic environment. Listener proficiency is controlled through rigorous training and a controlled listening environment.

It can be seen that if these items are controlled, held constant, then their effect is negated and the system under test becomes purely the system and its interaction with the room.

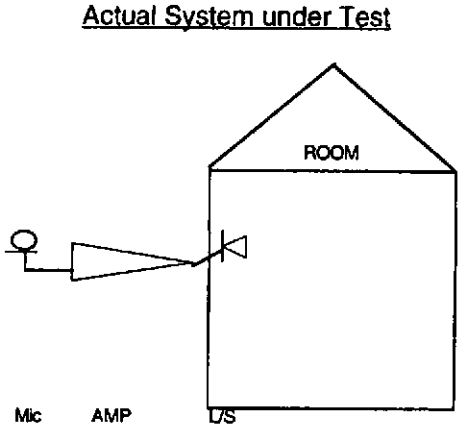

Fig.2

The process of preparation, data collection and analysis is long and involved. it requires rigour to be applied to every process.

2. THE PROCESS

As stated earlier, in most cases the primary function is measurement rather than to gauge or observe a reaction. As such it is important to take account of potential errors and as far as is practically possible, reduce them.

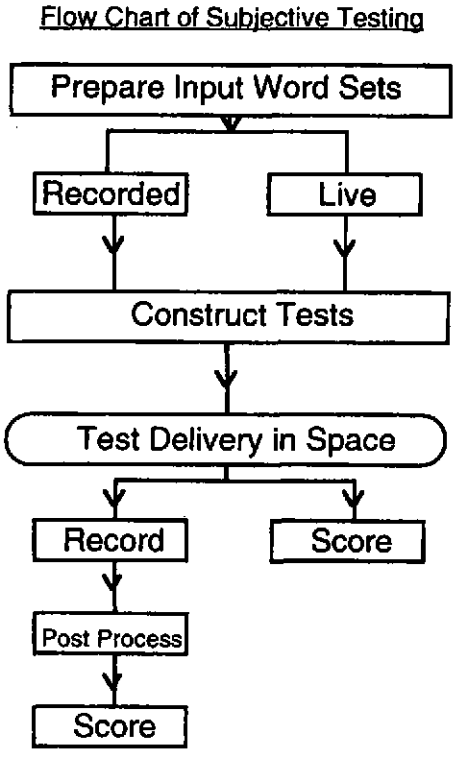

Fig.3

The flow chart given in fig. 3 shows the process of subjective testing (measurement) broken down into its component parts. Clearly some of the steps and interim tasks are in themselves involved and complex and worthy of individual treatment, with the limited time available the treatment can of necessity only be cursory.

It is worth noting that fig. 3 shows only the tasks and not the controls.

Finally it is worth noting that the bane of subjective testing is monitoring a talker/listener proficiency of around unity.

3. TEST CONSTRUCTION

Since this is a measurement exercise we are not in favour of 'live' tests since the obvious controls are different.

The recorded material is then input to the space either via the PA/VA system or via a mouth simulator with appropriate equalisation.

The recording or measurement positions will be dependent upon the task at hand.

4. PREPARATION OF SOURCE MATERIAL

The published standard Phonetically Balanced Word Scores are 20 x lists of 50 x words published in ANSI S 3.2 (1989). This list includes words which are not in normal usage in the UK. These Americanisms were removed and replaced with a phonetic equivalent.

The talker was chosen for her excellent speech skills and trained in the rate of delivery required. During her period of familiarisation with the works the talker was able to watch her level on a chart recorder and find a level that she was both comfortable with and able to maintain.

The recordings were then made of each list in the format of the word in a carrier sentence. The carrier sentence used was 'Please write WORD now' which serves several purposes:

a) To allow the talker to set the pace.
b) To allow the talker to set the level.
c) To prepare the listener for the word.
d) To lull the listener into a rhythm of listening and writing.
e) To excite the reverberant field and begin masking effects.

At the time of the anechoic recordings the acoustical effect of carrier sentence is obviously nil but it becomes the vehicle fro the space to exhibit its own acoustical characteristics. The recordings are made using a high quality microphone and put onto the DAT medium.

Sanitisation of the anechoic lists then takes place. The lists are put onto a computer and the gaps between the sentences regulated. The lists are then put through chart level recorder and the levels assessed for uniformity. This is done by looking at the average level of the carrier words. The level of entire sentences may be changed using the computer. Once this is complete, the word levels are never changed. The sanitised list is then recorded onto chart paper and the list committed to CD format. An L_{eq} of the entire list is measured for future reference.

At this stage the list is still in alphabetical order which is not valid for test presentation. Further preparation ensues during which the sentences are shuffled creating random variations of the original set. These shuffles are then put onto CD format.

Each list has 20 x shuffles giving a total number of test lists of 400 in total.

In taking all of the above precautions with care we believe that Pt = 1.

5. COLLECTION OF DATA

It is often not practical to expect a representative sample of the population to stand in a space and score Word Scores for any length of time and the presence of a healthy sized listening crew would change the acoustics of the space itself. For these reasons the PB word lists are recorded in the space using a binaural recording torso.

The sanitised shuffled list is played into the system at a normal operational level and the sanitised lists and the system and the space is recorded onto DAT via the torso. The system level must not be changed during the tests.

In order to ensure the data collected is good, a S/N ratio of at least +25dB is maintained throughout. Outside influences are kept to a minimum by undertaking the recordings during engineering hours. In order to audit the space in a representative way recordings are made at a selection of positions. The issue of the significant of the results obtained is dealt with by recording a minimum of two lists per position.

The use of the binaural torso is fraught with difficulties for this type of test. Consideration has to be given to the recording position, ear height and head orientation. Since the recording is binaural the orientation of the head must not be overlooked as an important issue as the listening crew will have difficulty scoring lists which arrive at the back of their heads. The level is also an important issue since the torso will move between positions and the system level must not change. Many tests have been aborted when the test furthest from a source has not achieved enough S/N ratio. If this is the position, left until last, all data is invalid. It has become common site practise to set the level of the tests in an area according to the most OFF-axis position which ensures that no gain changes are necessary.

If however the data collection is "live" then the situation is far more complex and these fundamental rules shall apply:

1. Travel to site by the listening crew should be monitored i.e. no lengthy car or tube journeys.

2. An adequate period of rest on arrival should be provided.

3. If a listener has had journey difficulties then the results should be monitored.

4. The listeners shall have undergone at least 2 days training.

5. The first set of tests should be control tests to ensure that a listener has maintained his or her rating.

6. PREPARATION OF RECORDED MATERIAL

Recordings are mostly made during periods of non-occupation and prior to presentation to a jury the word lists will need to be post-processed with the addition of noise to replicate the noise of occupancy. We generally use pink noise with suitable spectral shaping, actual noise often is too variable and also contains items of distortion.

A sample of pink noise is recorded in the space as a matter of course and this sample is used for the noise since it contains the artefacts and characteristics of the space and this would not sound out of context.

All mixing and data manipulation is carried out in the digital domain.

7. THE LISTENERS

The listeners are without doubt the source of the greatest potential errors. The objective of all subjective assessors or researchers is to ensure that listener proficiency approaches unity i.e. $P_l \Rightarrow 1$.

The various standards offer guidance on this issue but we believe fall short of the basic requirements and we have therefore introduced our own controls and procedures.

The simplest controls involve:

1. Hearing tests.
2. Training and motivation exercises.
3. Maintaining good comfort conditions.

We also maintain a permanent register of listeners mostly in the age group 18-45 but with some for specific tasks >65 years.

Each listener undergoes detailed training which includes around 10 hours scoring control tests (some 60 x control tests) to which we have the known answers.

Listeners are graded in the categories A-E where:

A : scores the average +3%
B: scores the average +1-2%
C: scores the average +/-1%
D: scores the average -1.2%
E: scores the average -3%.

A+ and E- jurors who consistently score either +4% or greater or -4% or less are discarded. We endeavour that a jury or listening crew does not contain a significant bias of attributes.

The scores of the listeners are continuously monitored and if a listener jumps more than one class, then subject to review, the listener will promptly be discarded.

Frequent tests are provided and we have found that 15 min. scoring, 15 min. resting is preferred by the listeners. In this way we accomplish around 6 x word lists an hour per listener (each list takes 4-5 mins.).

Listeners undergo constant and continuous monitoring and in this way, and are looked after and closeted throughout their brief term.

8. CONCLUDING REMARKS

With the reservation that the task is that of measurement, maintaining a talker/listener proficiency approaching unity is vital. The maintenance of the listeners is of paramount importance and is the crux of the matter.