DEVELOPMENT OF A NEW SPEECH INTELLIGIBILITY TEST FOR CHILDREN

A C Carey London South Bank University, London UK
B M Shield London South Bank University, London UK
J E Dockrell Institute of Education, London UK

1 INTRODUCTION

Acoustic design of classrooms is currently specified in terms of background noise levels, reverberation times and sound insulation [1,2]. The aim of these standards is to provide optimal listening and speaking conditions for pupils and teachers. In the case of open plan classrooms in England and Wales, requirements are specified in terms of the Speech Transmission Index (STI). However, the STI scale was developed following work on speech intelligibility with adults and it is not known whether or not the values quoted are relevant to the listening and understanding capabilities of young children. In order to determine what STI values are appropriate to young children, and what reverberant/noise conditions are suitable to provide good speech intelligibility for children it is necessary to have some form of subjective test. Such a test could be used to assess spaces to be used by children for activities in which listening is a critical activity, and also to investigate whether currently accepted criteria for speech intelligibility are appropriate for children. The work described here attempts to develop a reliable and valid subjective speech intelligibility test for children that accurately reflects the listening conditions in a room.

2 BACKGROUND

Various commentators have cast doubt on the suitability of the parameters commonly used for specifying and measuring speech intelligibility in classrooms and rooms for general use [3,4]. Validation of these parameters was in general carried out using adult subjects and little is known about their ability to predict the acoustic quality of classrooms as perceived by children [5,6]. A limited amount of work has been carried out previously to investigate the suitability of the parameters to children, but this has mainly been conducted within laboratory environments [7-9]. Furthermore, the work has focused on children aged 8 years and over, leading to a lack of data on younger children [10,11]. In the process of validating parameters and setting guidance levels, some speech intelligibility tests have been used [12,13] but they have, for a variety of reasons often been unsuitable for use by the children doing the test (even those over 8 years of age). It has also been the case that in some situations these tests have been carried out in unrepeatable ways or using very small groups of children resulting in conclusions being potentially unreliable. Some work has been carried out to predict young children's ability to understand speech in real classroom environments, both by extrapolating from empirical data and by using simple models [4,14] However, there is little evidence available to substantiate the predictions.

Most word lists in current usage for the subjective testing of speech intelligibility are designed for use by adult listeners and are therefore unsuitable for children. Some word and picture tests, such as the 'Word Intelligibility by Picture Identification' (WIPI) test, have been used recently in research on classrooms [15], but these tests were developed in the United States some years ago. They therefore contain several words that are unfamiliar to children in the UK. Furthermore, the pictures are dated, sometimes in black and white and in some cases quite difficult to identify, and are unlikely to appeal to today's children. It therefore appears that there is a need for a new test designed specifically to be used by young children in classroom situations in the UK. This paper describes the development of a new picture/word test for this purpose.

3 TEST DESIGN

A test using pictures to identify the words avoids difficulties that may arise if children do not know how to spell a word and therefore do not recognise the word when written. In a word and picture test the children can be presented with pictures representing the chosen words (the 'target' words) plus other pictures representing the 'confusions' or 'foils', and required to mark the picture which they think represents the spoken word.

For the new test, to be used in assessing speech intelligibility in schools, it was decided that there would be two equivalent tests each containing one list of words. The two lists would be of equivalent difficulty and be balanced for spread of consonants in test positions. Each test list would be restricted to around 20 test items to ensure that it did not become too long for the children and both final (bea*ch*-bea*n*-bea*k* etc.) and initial (*c*at-*h*at-*m*at etc.) confusions would be tested. The consonants chosen to be under investigation (in the target word) and those used as confusions (in the foil words) would as far as possible be controlled by their frequency in British English and the frequency of children's consonant confusions.

It was decided that the test would be presented in booklet form, each booklet containing around 25 pages. Each page would show four pictures: the test item (target word) and three acoustically similar foils. Limiting the number of foils to three for each test item increases the chance of the subject guessing the correct answer. However, it was felt that, since the children need to be able to process the pictures at the same time as the test word is being heard, three rather than a larger number of foils should be used, as is typical of such tests. It has also been found that using additional distractors, that is foils which are not acoustically related to the target item, such as the two extra pictures that occur in the WIPI test can be counterproductive [16]. These types of foil are very rarely chosen as errors and make the test more confusing for young children.

To aid the children in scanning all four illustrations in the given time period, it was decided that pictures would be presented in a two by two structure as this is a familiar format to them, and therefore easier to process than other arrangements. The final booklet was laid out so that the first four sheets are used for practice and the next 20 sheets for test items and their foils. Training times for small closed set intelligibility tests are minimal and were estimated to be around five minutes for the proposed test [17].

The words would be recorded and played back to the children in their classrooms. Each child would have a test booklet and be instructed to mark with a sticker the word they thought they heard.

3.1 Choice of potential words

Potential words for use in the speech intelligibility test were chosen using the following criteria.

- The words must be meaningful to children of the appropriate age
- The words must be age appropriate by objective measures
- The words must be monosyllabic
- The words must contain consonants in initial/final positions with frequency approximately equal to that with which they occur in spoken British English
- The words must be able to be easily illustrated
- The words must test the most common children's consonant confusions
- The words must be of approximately equal "familiarity" to the children

As many potential test items as possible had to be found. It was decided, due to the limited vocabulary available for the tests (owing to the age of the children) that minimal difference testing, that is one phoneme change, for example "tie-pie" (/taɪ/-/paɪ/), could not be adhered to. Therefore, a decision was made that double phoneme changes (eg "tie-fly") and removing the test phoneme (eg "tie-eye") would be allowed but their numbers kept to a minimum, and the proportion of double, single and no phoneme changes would be balanced as far as possible between the two lists.

Potential test items were collected from a range of sources such as lists used by speech and language therapists [18,19], words provided for teachers by the Department for Education and Skills as vocabulary to be taught at each key stage, psycholinguistic databases and children's vocabulary lists [20-22]. The age of acquisition familiarity and verbal frequency of words were also considered. Altogether 212 potential words were identified as suitable for use.

Possible items were phonetically transcribed and sorted into acoustically similar groups, for example. "pig-pill-pin-pip" and "wall-tall-fall-shawl-ball-stall". It should be noted that in this instance "acoustically similar" refers to rhyming words in the case of initial confusion sets (pig-pill-pip), or words which start with the same consonant vowel pair in the case of final confusion sets (wall-shawl-ball), and not acoustic descriptors such as voicing, aspiration or nasality.

3.2 Choice of words and foils for two word lists

A literature search of all the data on frequency of consonant confusions made by children, plus the most well known adult data was carried out [eg 23,24]. Analysis of the data is made difficult because different studies test different stimuli types (words, phonemes etc.), consonant pairs and also use different marking and scoring systems. In addition much of the work has been in languages other than British English. However, sufficient information was obtained to identify common confusions which have been used to inform the selection of words and foils.

As a result of the data search it was decided that 14 of the 20 items in the children's test would be initial confusions and 6 would be final confusions. Some compromise was necessary in the choice of consonants for testing in the final position because of limitations caused by the restricted vocabulary of young children. Foils were chosen as far as possible to present the confusions more frequently made by children (as far as the restricted vocabulary would allow). To obtain a sufficient number of potential foils a search of dictionary databases, rhyming dictionaries and crossword dictionaries was carried out.

Once all the target items had been chosen the two lists were balanced so they had the same consonants (under test) in the initial/final position in both. Minimal, double and no phoneme differences were also distributed evenly between the two lists. The number of times where a word appears as a target and a foil in the same list and across lists was minimised, as was the number of times foils are repeated within a list. Some examples of test words and related foils are shown in Table 1.

Table 1: Examples of test words and related foils

Item under test		Target	Foils		
Initial /h/	List 1	Hat	Rat	Fat	Bat
	List 2	Ham	Jam	Lamb	Pram
Final /p/	List 1	Cap	Cat	Can	Cash
	List 2	Map	Man	Mat	Match

3.3 Illustration of words

Pictorial representations of as many of the potential words as possible were found from sources such as the web and commercial clip art packages [25]. An informal study to assess the suitability of the pictures was carried out with a small group of children aged between four and seven. The pictures were tested in three different ways: some children were asked to state what the picture represented, while others were asked to point to the picture representing a particular word. The children were also asked for their general opinions of the pictures. Many of the illustrations were not correctly identified when the children were asked what the picture represented. There had been some concern that the children would have problems with images representing verbs or adjectives, but this did not appear to be the case. The children had strong ideas and opinions about the use of colour and styles of the illustrations. They disliked the use of black and white, preferring full colour

cartoon style pictures to both line drawings and photographs. The use of the same image to represent two different words did not seem to confuse the children at all.

Following this preliminary study the following set of criteria were created for the illustrations:

- They should be modern in appearance and culturally appropriate
- They should be the same style, for example. all photographs or all diagrams
- They should be in full colour and have a consistent colour scheme
- They should all have a border
- A method of isolating parts of pictures (other than arrows) should be found that young children understand.
- They should be recognisable to the modern child of the appropriate age

A professional illustrator was commissioned to produce a sample set of 12 illustrations. These were shown both to a selection of adults and to some of the children used in the informal study. Comments about the illustrations were then reported back to the illustrator who rectified any problems, for example animals were re-drawn so that their whole body could be seen in the illustration rather than just a portion.

The illustrator was then asked to produce illustrations for all 212 potential words. Some words that could not be easily illustrated (eg float) were withdrawn from the lists.

3.4 Picture recognition testing

Before a final decision was made on composition of the word lists the illustrations were shown to children to assess if each picture was recognisable to children of the appropriate age. Each illustration was tested for recognition a minimum of 10 times and a requirement of 75% correct recognition was set which must be achieved if the illustration was to be included in the final word lists.

Ninety-eight children, from year 1 and year 2 classes in seven schools in and around London, participated in the picture recognition testing. As the final speech intelligibility test is aimed at year 2 and above, most of the illustration testing was done with year 1 to ensure that year 2 children and children with special needs would understand the illustrations.

The illustrations were laid out in sets of 12 on landscape orientated A3 sheets, as shown in Figure 1. Sets of 12 were used in an attempt to prevent children from using elimination to guess the correct picture. Illustrations which were to appear together in sets in the final test were not presented together in this study, in order to avoid acoustic similarity confusion at this stage.

The children were tested individually in a quiet area of the classroom, during an afternoon session of the school day. Each child was asked to point to the picture representing the word they heard, words being repeated up to five times if a child did not hear it at first.

Each child completed between two and four test sheets (24 to 48 words). Each illustration was tested by between 10 and 15 children, and in at least two different schools in different areas. A rejection criterion of less than 75% correct recognition was set. A study was made of all failed illustrations. If the majority of responses from the test subjects for a failed illustration were due to children not knowing the word then the word was removed from the test, regardless of whether it was a target item or foil. The remaining failed illustrations were separated into two categories. If the illustration was a target item or a critical foil (i.e. where there was no other viable foil to replace it) then it was re-drawn. If the illustration was a foil which could be easily replaced by an alternative then the illustration was removed from the test. On occasion a word failed because there were two similarly meaning words on the same sheet (eg coat and shawl) so the word was tested again on a new sheet against other words.

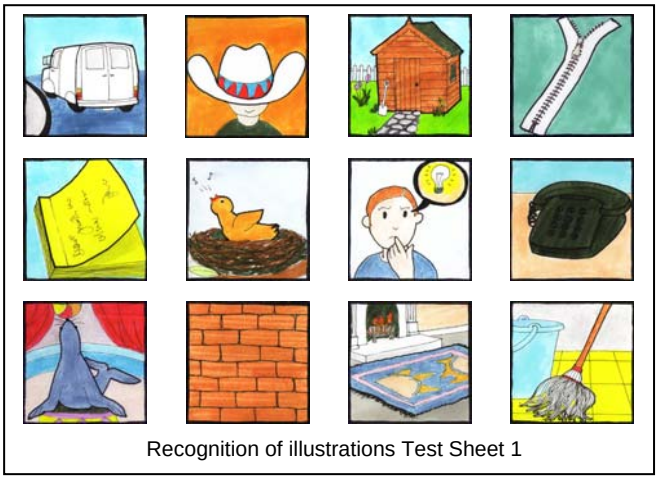

Figure 1: Example of picture recognition test sheet

In total 30 words were removed as the word was not recognised, 16 as the picture was not recognised and 26 as, once the other words had been removed, there were no longer sufficient words left to provide three suitable foils for a target word.

3.5 Production of final word lists

Once all the words and pictures had been tested as described above the final word lists were drawn up. Because of the reduced number of available words it was necessary to use some words more than once. However, no items appear as the target within a list on more than one occasion although some items have been used as both a target item and a foil within a list. A small number of words appear in both lists as target and/or foil.

The frequency of occurrence of consonants in test positions is evenly balanced across the two final word lists and is also representative of the frequency of consonant occurrence in spoken English. The relative acoustic intensity of consonants was also considered. The aim was to have more occurrences of weaker consonants in the word lists as it was thought that these would provide a more accurate assessment of the speech intelligibility of a space. However, in practice, because of the relatively small number of words available, this was not possible while also fulfilling the other aims for balancing the lists.

3.6 Recording of word lists

The words lists were designed to be presented to children in their classrooms using recordings played through a loudspeaker with directivity characteristics similar to that of the human speaker. The target test items were recorded by a professional actor in an anechoic chamber using the carrier sentence, “Mark the picture for ---- please”. Recorded test items were equalised during post processing and inserted into the reference carrier sentence for playback.

4 RELIABILITY TESTING

Before the tests can be used to assess speech intelligibility it is necessary to ensure that the tests themselves are valid and reliable. Initially, it is necessary to check that each test is repeatable, that the two word lists are equivalent and that there are no item order effects.

Validation of the tests is therefore currently being carried out in order to investigate the issues above. To date, the tests have been used in nine classes in three different schools, with 228 children aged between 7 and 9 participating in the validation.

In the validation of the test four versions of the test were used: the two word lists plus each list in reverse order. Each class of children was given a test on two occasions, between two and six weeks apart. Children completed both tests in their usual classroom in the presence of their normal class teacher.

For each test the words were presented through a loudspeaker at the front or in the middle of the classroom, with the children sitting in their usual positions.

4.1 Results of reliability testing

In total there are 28 valid sets of results for use in investigating repeatability, order effects and equivalence of the two word lists. These provide seven pairs of results for investigating repeatability; three pairs of results for investigating order effects and four pairs of results for examining equivalence of the two word lists.

For each test pair, the scores for each word on the two test occasions have been compared and ranked pairs correlation used to investigate statistical differences between the two sets of results. (It was necessary to use ranked pairs correlation, rather than Pearson correlation, as the data are not normally distributed.)

4.1.1 Repeatability

Seven classes repeated the same word test, with a period of at least four weeks between the two times of testing. For five of these seven pairs of test/retests there are no statistically significant differences between results. For two classes of 7/8 year olds in the same school there was a significant difference between the results on the two times of testing.

4.1.2 Order Effects

Three classes did the same word test in both forward and reverse order, with a period of at least two weeks between the two times of testing. There was no statistically significant difference between the results for any of the three classes.

4.1.3 Equivalence

Four of the classes did both word tests, with a period of at least four weeks between the two times of testing. Of these four classes, two showed no statistically significant difference between the results while for the other two classes the results were significantly different.

4.2 Discussion of reliability test results

The statistical comparison of the test results on two occasions of testing using the same children in the same classroom shows that order does not affect the results of the word test. The classes for which there was a difference for both test retest reliability and list equivalence were the same. These were classes of 7/8 year olds, in the same school. It is possible that the data from these classes may be unreliable. A similar difficulty was reported recently by Picard [26] who, when doing word tests with similar age children, found that children of 7/8 years old gave inconsistent and unreliable results. The fact that for five of the seven test/retest pairs there was no significant difference in results suggests that the test may be regarded as repeatable. Further investigation is being carried out to examine the issue of the equivalence of the two word lists.

4.3 Practical difficulties

During the validation several practical problems became apparent which will need to be resolved before the test is fully rolled out for use in testing speech intelligibility in classrooms and other spaces.

Although the children were told that it was not a test of them, that there were no right and wrong answers and that they had to make up their own minds and not look to see what other children had put there was a lot of copying among the children. Also, although the children were also told they must not tell other children what they had put, and that they should do the test in silence, in many classes one child (usually a boy!) would shout out some of what he thought were the correct 'answers'. A possible solution to this in future would be to present the test on individual laptops or computer, or for the test to be carried out with small groups of children under adult supervision.

The children marked the picture they thought was 'correct' using coloured stickers. However, this proved to be quite difficult for some of the children, particularly the younger ones. Children became quite distressed if they missed a word or were unable to answer because they had dropped their stickers. This then affected their performance for the remainder of the test. The use of laptops or computers in future would overcome this problem.

5 CONCLUSIONS

A new subjective speech intelligibility test for children is being developed. It is a word and picture test designed specifically for young users of British English, using words which are known to be understood and familiar to children aged 7 to 8 years of age. Preliminary trials of the test suggest that it is repeatable and that the order of words does not affect results. Future work involves using the test in rooms of different acoustic characteristics to see whether the test is sensitive enough to indicate differences in speech intelligibility in rooms.

6 ACKNOWLEDGEMENTS

This research has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC). The authors would like to thank the schools, teachers and children that have taken part in the project.

7 REFERENCES

1. Department for Education and Skills, Building Bulletin 93, "Acoustic Design of Schools", The Stationery Office, 2003, www.teachernet.gov/acoustics, (2003)
2. American National Standards Institute. ANSI S12.60-2002: Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools, (2002)
3. P. Mapp and C. Boyden, "Measuring speech intelligibility in classrooms, with and without hearing assistance", in Proceedings of the Institute of Acoustics, 25(7), (2003)
4. M. Hodgson and E. Nosal, "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms", Journal of the Acoustical Society of America 111, 931-939, (2002)
5. T. Houtgast, H.J.M. Steenken and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics", Acustica, 46, 60-72, (1980)
6. J.S. Bradley, "Predictors of speech intelligibility in rooms", Journal of the Acoustical Society of America, 80(3), 837-845, (1986)
7. T. Finitzo-Heiber and T.W. Tillman, "Room Acoustic Effects on Mono-syllabic Word Discrimination Ability for Normal and Hearing-Impaired Children", Journal of Speech and Hearing Research 21, 440-458, (1978)

8. W.S. Yacullo and D.B. Hawkins, “Speech recognition in noise and reverberation by school aged children”, Audiology 26, 235-246, (1987)
9. P.G. Stelmachowicz, “The relation between stimulus context, speech audibility, and perception for normal-hearing and hearing-impaired children”, Journal of Speech, Language and Hearing Research, 43, 902-914, (2000)
10. T. Houtgast, “The effect of ambient noise on speech intelligibility in classrooms”, Applied Acoustics 14, 15-25, (1981)
11. J.S. Bradley, “Speech intelligibility studies in classrooms”, Journal of the Acoustical Society of America, 80,.846-854, (1986)
12. D. Markham and V. Hazan, “Speech intelligibility of adults and children”, Proceedings of International Conference for Spoken Language Processing, (2002)
13. D.N. Kalikow, K.N. Stevens and L.L. Elliott, “Development of a test of speech intelligibility in noise sentence materials with controlled word predictability”, Journal of the Acoustical Society of America, 61, 1337-1351, (1977)
14. M. Picard and J.S. Bradley, “Revisiting Speech Interference in Classrooms”, Audiology 40, 221-244, (2001)
15. M. Picard. Speech recognition in classroom noise by school-age children. Proc. 19th International Congress on Acoustics, Madrid (2007)
16. J.R. Foster and M.P. Haggard, “The four alternative auditory feature test (FAAF) – linguistic and psychometric properties of the material with normative data in noise”, British Journal of Audiology, 21, 165-174, (1987)
17. International Standards Organisation, ISO/TR 4870:1991(E), “Acoustics – The construction and calibration of speech intelligibility tests” (1991)
18. J.M. Wepman and W. Haas, “A spoken word count: children”, Western Psychological Services, Los Angeles, (1969)
19. J. Morris, S. Franklin, A.W. Ellis, J.E. Turner, “Remediating a speech perception deficit in an aphasic patient”, Aphasiology 10, 137-158, (1996)
20. Medical Research Council, Psycholinguistic Database*, www.psych.rl.ac.uk*
21. L. Fenson, P. Dale, J. Reilly, S. Reznick, D. Thal, and E. Bates, “The MacArthur Communicative Development Inventories: Manual and Norms”, San Diego, CA: Singular Publishing. (MacArthur Communicative Development Inventories, (1993)
22. J. Masterson and J. Druks, “Description of a set of 164 nouns and 102 verbs matched for printed word frequency, familiarity and age-of-acquisition”, Journal of Neurolinguistics, 11(4), 331-345, (1998)
23. M.A. Redford and R.L. Diehl, “The Relative Perceptual Distinctiveness of Initial and Final Consonants in CVC Syllables”, Journal of the Acoustical Society of America, 106, 1555-1565, (1999)
24. H. Steeneken and T. Houtgast, “Phoneme-group specific octave-band weights in predicting speech intelligibility”, Speech Communication, 38, 399-411 (2002)
25. Nova Development Corporation, “Gold Edition Art Explosion 125000 images”, *www.novadevcorp.com* (1996)
26. M. Picard. Personal communication, September 2007