UNIVERSITY OF SOUTHAMPTON
2006

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

INSTITUTE OF SOUND AND VIBRATION RESEARCH

THE DEVELOPMENT OF A NEW ENGLISH SENTENCE IN NOISE TEST AND AN ENGLISH NUMBER RECOGNITION TEST

by Stuart John Hall

A dissertation submitted in partial fulfilment of the requirements for the degree of MSc by instructional course.

Abstract

An English synthetic sentence test and a number triplet test have been developed as part of HEARCOM, a wide-ranging European Union project aimed at mitigating the effects of hearing impairment in the information society. There is a lack of such tests in the English language and the present study attempts to create and develop English versions of these speech tests to add to a strategy of unified procedures and methods across Europe. The sentence test was developed following the same principles as the Hagerman sentences (Swedish), the Oldenburg sentences (German) and the DANTALE II test (Danish). The number triplet test was based on a Dutch speech-in-noise screening test for completion by telephone.

The sentences were randomly generated from a base list. This base list consisted of 10 rows of words with the same syntactic structure (name, verb, numeral, adjective and object). The sentences were spoken, recorded and cut into separate words in such a way as to include co-articulation effects, so that when re-combined randomly into new sentences they would be perceived as naturally spoken. The triplets were similarly randomly generated from a base set of nine spoken digits, but without co-articulation.

Word specific speech recognition functions were measured for every word in each test and compared to the overall speech recognition function for the combined material for each test so that level corrections could be calculated. These corrections when applied would achieve homogeneity in the speech material and improved measurement accuracy. Once the corrections have been applied to the words a further evaluation of the material should be carried out to establish if homogeneity has been satisfactorily achieved.

The 50% correct level for the synthetic sentence test corresponded to a signal-to-noise ratio˜9.5 dB. The 50% correct level for the number triplet test corresponded to a signal-to-noise ratio of˜11.8. These findings show good agreement with other similar tests. It is concluded that the new materials form the basis of viable English language tests in the common format being adopted by the HEARCOM project.

ACKNOWLEDGEMENTS

The author would like to thank his supervisor, Professor M E Lutman for his much appreciated help and advice in the preparation of this study. He would also like to thank Kirsten Wagener for her guidance and expertise in the creation of the tests. He also gives thanks his colleagues Ann Stevenson and Elaine Maynard and his parents for all their support during the past year.

CONTENTS

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

List of Tables

List of Figures

Chapter One: Literature review

1.1 Introduction

Speech is the most important sound which we listen to in everyday life. It allows us to communicate with the complex world which surrounds us. To an individual with a hearing loss the degree to which they can perceive and understand such speech is an important feature of their condition. Since its introduction pure tone audiometry has been used to accurately describe the characteristics of a person's impairment and make a significant contribution to the diagnosis of the hearing disorder. However since this time audiologists have had to face the fact that pure tone audiograms do not provide a good measure of hearing loss for speech and as such do not provide any direct measure of the handicap endured in everyday life. A variety of audiometric descriptors have been suggested for best representing performance at speech tests, most commonly averages of hearing threshold levels (HTLs) at frequencies between 0.5-6 kHz, however no single descriptor has been agreed upon. The utilisation of self-assessment has broadened the basis of disability rating and made the descriptors largely redundant. Lutman et al (1987) derived four principal components from the responses of 1470 people who completed a questionnaire. Two of these components deal with speech, they are, 'everyday speech' and 'speech in quiet'. Everyday speech showed a correlation coefficient of 0.63 with the HTLs of 0.5, 1, 2 kHz and 0.62 with 0.5, 1, 2, 4 kHz and speech in quiet showed a correlation coefficient of 0.39 with both of these descriptors.

So while speech tests of hearing have less diagnostic value than pure tone audiometric tests, the minimal level at which a listener can correctly perceive speech in quiet correlates highly with the listener's absolute sensitivity averaged across the frequencies relevant to speech (Festen and Plomp, 1983). So to gain an insight into the problems encountered by those with a hearing impairment speech tests play a vital role as stated by Fry (1961) "[speech reception] depends upon the condition of the peripheral hearing mechanism and the efficiency of the central decoding mechanism, the speech centres of the brain".

1.1.1 The nature of the speech signal

Speech is the communication system for all human beings and these linguistic structures fulfil their intentions if they allow the listener to generate the same thoughts that the talker is trying to express (Gibson, 1998).

The physical description of speech usually commences with the analysis of waveforms and whether or not they are periodic, a repeating waveform, or aperiodic, a non-repeating waveform. At this level speech can be considered to be simply a disturbance of air pressure. We can physically describe speech signals in terms of their amplitude, period, spectrum and duration.

Each of these physical descriptions has an equivalent perceptual description. Perceptually, speech signals can be therefore defined in terms of loudness, pitch, quality and length.

Physical	Perceptual
Amplitude	Loudness
Period	Pitch
Spectrum	Quality
Duration	Length

Table 1.1 Physical and perceptual descriptions (Wright 1997).

1.1.2 Linguistic units of speech

In addition to being defined by physical and perceptual descriptions, speech can also be broken down into linguistic units. The smallest speech sound is the phoneme. Phonemes can be classified into two groups, vowels and consonants. Individually phonemes are abstract units but in relation to one another they distinguish one word from another. For example 'fog' and 'jog' each have three phonemes but it is possible to distinguish each separate word by the phonemes /f/ and /j/. When combined together phonemes form syllables. A syllable usually consists of a vowel surrounded by one or more consonants. The combination of one or more syllables forms units which we would recognise as words. Words when linked together form larger linguistic units termed sentences, the use of which in speech testing will form a large part of this study.

Sentence structure is described in terms of phonology, morphology, syntax and semantics. Phonology describes the phonemes of the language, their formation and how they combine into words. Morphemes are the smallest meaningful linguistic units for example /s/ is a morpheme when acting as a plural marker. Morphology describes how morphemes are combined into words. Syntax refers to the system of word arrangement and how words linked together form acceptable sentences. The study of word meanings is semantics. A sentence must have the correct syntax and semantics for it to seem a natural composition in a particular language.

1.1.3 Speech perception – auditory space

The acoustic properties of speech can be summarised and illustrated using a diagram similar to an audiogram (dB vs. frequency), with areas marked out for the various speech sounds. See diagram below.

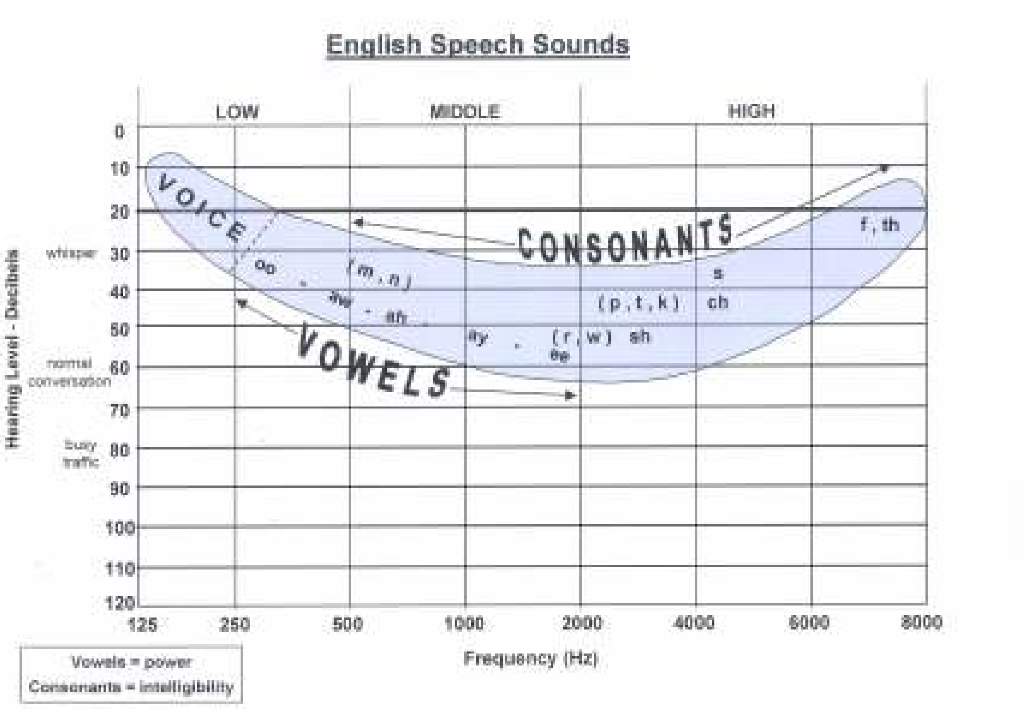

Figure1.1 The frequency component of English speech sounds, adapted from Wright (1997).

Such a diagram is informative but incomplete. It only maps the short term spectral content of speech overlaid on the auditory dimensions of a response to the level of frequency of pure tones. It is mainly a summary of the locations of 'auditory space' involved in vowel contrasts and place/manner decisions.

Detection of pitch is usually neglected and a more informative diagram would show pitch related information as being dependent upon the entire spectrum of the speech signal. The biggest problem with conventional auditory space, for pitch perception and generally for speech processing as a whole is that the time dimension is neglected. The perception of speech is dependent on a three dimensional space (amplitude vs. frequency vs. time), and this type of diagram just shows section through this space. The consideration of time is particularly relevant to pitch perception and sensorineural hearing loss, due to the existence of a temporal mechanism for frequency discrimination.

1.2 Objectives of speech audiometry

Traditional assessments of hearing loss based on pure tone thresholds may not adequately measure the function of the auditory system with wideband signals, nor do these assessments accurately predict speech intelligibility in noisy environments (e.g. American Academy of Otolaryngology Committee on Hearing and Equilibrium, 1979).

Historically speech testing has been used for the differential diagnosis of sensory and neural disorders. Speech tests are also used in assessing a patient's potential candidacy for rehabilitation using hearing aids. The speech test can give information whether a hearing aid would be effective. The amount of initial disability and examination of the individual speech reception scores for each ear may help decide which ear is most suitable for hearing aid fitting.

Many studies have used speech testing in their evaluation of benefit of hearing aid fittings and in comparison of different types of hearing aids by testing the subject in both aided and unaided speech recognition. More recently speech in noise testing has been used to assess the effectiveness of noise cancelling technologies employed by nonlinear hearing aids.

Speech testing is often employed prior to surgical intervention firstly to assess if the integrity of the auditory system is worth preserving and secondly to what extent speech discrimination has been affected post operatively.

The aim of the speech test is to assess how well an individual can correctly identify speech under certain conditions. Speech tests are generally intended primarily to be representative of everyday speech used for communication and the tests aim to give an indication of disability (Lutman, 1997).

1.2.1 Speech audiometry

Carhart (1951) defined speech audiometry as follows: The technique wherein the standardized samples of a language are presented through a calibrated system to measure some aspect of hearing ability. Lyregaard et al (1976) state speech audiometry means any method for assessing the state or ability of the auditory system of an individual, using speech sounds as the response evoking stimuli.

These definitions both place the role of speech audiometry as the assessment of the auditory system. The basic principle of speech audiometry is as follows.
Specified speech materials such as words are presented to the subject, who then indicates what they heard. The tester then compares the reported and the presented material and derives a score for that particular condition. Items from lists of the selected speech material are presented to the subject at a series of intensities and a speech recognition score obtained for each intensity.

These can then be plotted on a speech recognition curve or performance intensity functions so that scores obtained are plotted against the intensity of the speech signal.

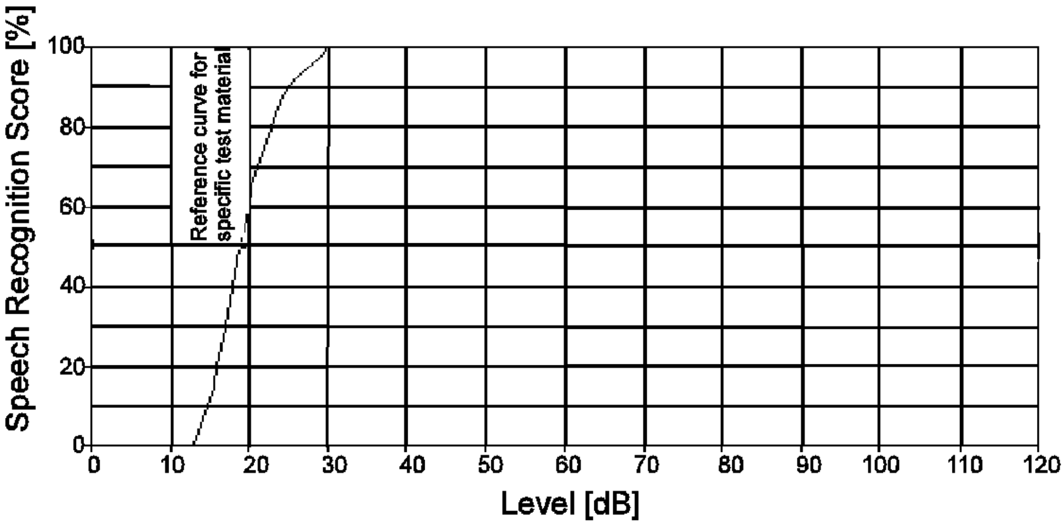

Figure 1.2 Speech audiogram, with reference curve given as an example.

The test validity depends on keeping all the other factors constant other than the test subject.

1.2.2 Accuracy of speech audiometry

If a difference is observed between two scores obtained from one individual under two test conditions, this may be due to variability in the test, variability in the subject or to the difference in the two test conditions. Scores from an ideal speech test would be exactly repeatable. However, in practice, this is not usually the case; if a subject performs a speech test and achieves one score it would be unrealistic to hope that they would obtain exactly the same score every time the test was performed under the same conditions. If this were the case then if they achieved a different score under a different condition we could be confident that this change was the effect of the different condition. In reality the test scores will show some random variation from test to test. This variability will depend on the test itself, and on the conditions under which it took place. There is in fact a trade-off between the sensitivity of the test and its reliability.

There are critical differences that have been noted by researchers Thornton and Raffin (1978) where tables were constructed to display upper and lower limits comparing the use of one or two lists which can be said to be significantly different with 95% confidence.

Markides (1978b) stated that the test re-test repeatability of the AB(S) word lists is reasonably high, providing correlation coefficients ranging from 0.34 near to threshold to 0.79 at supra threshold for identification scores.

1.3 The Choice of speech materials

1.3.1 Phonemic balance

If the test material is said to have phonemic balance, usually termed phonetic balance (PB), it has a phonemic composition which is equivalent to that of everyday speech. The different phonemes should occur in the test material with the same relative frequencies as is observed in everyday speech. The reasoning behind this is if the subject was unable to perceive a particular phoneme that occurs infrequently in normal everyday speech then the handicap they will experience will be less severe than it would be if the phoneme had been a more common one. Phonemic balance may be described as a relationship between the parent population and the test material

The familiarity of the speech material used in the test is important for both test and interpretation and need to be considered when considering sources of speech material. The term familiarity implies that if a subject has a greater familiarity with one stimulus with regard to another then they will more readily recognize the one which is more familiar. In order to quantify the familiarity of a word the assumption is made that it is equivalent to the frequency with which the subject has been exposed to that word. That is then approximated from the frequency of occurrence of words found in a corpus of word material, sampled to ensure good coverage of written or spoken material. Owens (1961) and Savin (1963) indicate that uncommon words have a lower intelligibility than common words, everything else being equal. The effect that this is likely to have in terms of SRT shift from commonly occurring words to uncommon words has been estimated has been estimated at 15 dB by Howes (1957).

Most speech tests have been constructed to account for this word frequency effect, uncommon words are normally excluded from the word material. Wagener et al (2003) selected the words for a Danish speech test by analysing word frequency in the written language. This analysis was performed on the 5000 most frequently used words in Danish this is to give each test sentence equal difficulty. This gives the test subjects an equal footing as far as familiarity is concerned.

Many variations of material have been developed for speech audiometry. The choice of material depends upon the intended purpose of the test. If the purpose is to measure an individual's speech recognition ability, then the choice of test material should resemble natural conversation as closely as possible.

1.3.2 Nonsense syllables

The advantage of these tests is that they look at the acoustic and phonetic information in speech and contain minimal semantic content. Therefore the subject's performance is not affected by vocabulary or education. The items are usually presented to the listener and various potential responses offered from which the subject can choose, thus allowing for analysis of the phonemic errors made. There are disadvantages to this type of test, as the test material is abstract and the composition of nonsense syllables does not resemble the sequencing found normally in language. It may be difficult to obtain appropriate responses from the subject without considerable training so are not really suitable for general clinical use.

1.3.3 Monosyllabic words

Monosyllabic words are easier to use in clinical testing as they are more familiar to the subject under test and therefore more readily repeated than nonsense syllables and their abundance in the English language provides a large pool of source material.

1.3.4 Sentences

Sentence tests are used to measure the intelligibility of realistic speech material. Sentences assess the phonetic, lexical and semantic information. The advantage of using sentences instead of isolated words or syllables is that they incorporate the whole language system and resembles everyday listening conditions much more realistically than using isolated utterances giving them high face validity.

Most often sentences are used for the determination of the speech reception threshold (SRT) in noise; that is, the signal to noise ratio at which a 50% recognition score is obtained. Depending on how the sentences have been constructed and the contextual content included in the sentence material they can be divided into the following groups.

- Short meaningful sentences
- Syntactically fixed, but semantically unpredictable short sentences
- Carrier phrase type sentences

Short meaningful sentences

Each test list consists of 10-20 short meaningful sentences that are matched with respect to their intelligibility in noise and that approximately represent the phoneme distribution in the respective language. Examples are Plomp sentence test (Plomp and Mimpen 1979a), Göttingen Sentence test (Kollmeier and Wesselkamp 1997),

HINT test (Nilsson et al. 1994) and BKB sentence test (Bench, Kowal and Bamford 1979).

Syntactically fixed, semantically unpredictable sentences

This type of sentence was first proposed by Hagerman (1982). It employs short sentences of the form “name” “verb” “number” “adjective” “object”. For each position of the sentence ten alternatives are available. Within each list all ten alternatives of all five positions are used. They approximately represent the phoneme distribution in the respective language. The advantage is that a large number of different sentence lists can be generated that all use the same word material. Examples are: Hagerman Sentences (Hagerman 1982), Oldenburg Sentences (Wagener et al 1999a).

Carrier-phrase-type sentences

This type of test uses sentences primarily as a carrier phrase to introduce a certain key word. Hence the result of the test is the intelligibility of the key words only, whereas the intelligibility of the remaining elements of the sentence is not considered. The key word can either be predictable from the remainder of the sentence (highly predictable) or not predicable at all (no predictability). Hence these tests are in between word and sentence tests. They can be used to assess the degree to which the listener can make use of the contextual information in understanding the respective key words. SPIN Sentence test (Kalikow et al 1977) and Basel Sentence test (Tschopp and Ingold 1992) are examples.

1.3.5 Running speech

Running speech contains all the semantic, prosodic (rhythm) and contextual features of everyday speech so has high face validity for use as a testing material. However the results are difficult to record and the subject requires a certain amount of training to be able to repeat running speech accurately. There are also significant challenges in quantifying the materials.

1.3.6 Speech in noise

It is common for hearing impaired listeners to complain of increased difficulties hearing in background noise or in reverberant environments. Available evidence suggests that individuals with hearing impairment are more susceptible to the deleterious effects of background competition than are individuals with normal hearing (Dirks, Morgan and Dubno 1982). Speech-in-quiet tests can therefore be insensitive and ineffective measures of the effectiveness of a chosen management strategy, as they have failed to reproduce the listening situation which resembles everyday listening environments.

Speech tests in noise require a suitable noise to be selected to accompany the test material; generally the noise should have sufficient energy at all the frequencies present in the speech signal. Noise can be generated to have a frequency spectrum which approximates the long term spectrum of speech, or the voice of another or several talkers may be used.

There are many variations of noise used for speech-in-noise testing, which combine the speech from several talkers, known as speech babble the speech of several talkers can also be mixed with cafeteria noise, white noise, and speech shaped noise and others. The effect of noise is to interfere with speech understanding and mask some of the speech signal so that the listener has less acoustical information on which to interpret the speech. This requires an increased effort on the part of the listener. Speech babble interferes with speech intelligibility to a greater extent than stationary noise, the degree of masking that occurs is dependent on the number of voices mixed. Optimal spectral masking can be achieved by using stationary noise with the same long-term spectrum as the speech material state (Wagener et al, 2003).

The temporal structure of noise affects its ability to mask the speech sounds. Speech is a spectro-temporal code containing information in both the frequency and time domain. A steady noise is less likely to mask the amplitude modulations of the speech than those with similar amplitude modulations to those in the speech and therefore amplitude modulated noise (babble) has a greater masking effect than steady continuous noise. With amplitude modulated noise there are temporal fluctuations in the masking noise and the listener may be able to detect the speech sounds in the relatively silent gaps. The ability depends of course on the temporal resolution of the ear so is likely to be reduced in hearing impaired listeners.

The main advantage of using speech in noise is its use in testing a heterogeneous population. A suitable range of scores can be determined in advance by manipulating the signal-to-noise ratio. It is therefore possible to avoid or at least reduce the problems of ceiling and floor effects (scores of 100% and 0% respectively) which would lead to the test having a lack of sensitivity and inadequate measures of rehabilitation efficacy by setting the absolute level of materials at a moderately high intensity easily audible to most subjects and adjusting the signal-to-noise ratio. The addition of background noise to a speech test attempts to overcome ceiling effects of testing in quiet.

1.3.7 Redundancy in speech material

The redundancy of speech material relates to the ease with which an individual can determine what is being spoken from the context. Sentences have a much higher redundancy than phonemes. The less choice there is amongst the alternative items the more the items are redundant. The identification of certain speech sound patterns will also be influenced by the duration and silent intervals of speech. Parnell and Amerman (1978) state that the influence of adjacent phonemes on each other is known as 'co-articulation' and represents a bi-directional phenomenon involving overlapping of articulatory movements for two or more phonetic segments which has an obvious effect upon the listener. In the context of speech testing the effect of increased redundancy and reduced choice is to make the slope of the audiogram steeper. Lehmann (1962) suggested that the higher the redundancy, the fewer the acoustic cues needed to recognise the stimulus when related to the shapes of speech intelligibility curves.

Comprehension of a sentence when preceded with information about its context is greater than when presented in a neutral context (Kalikow et al, 1977). Therefore a test using sentences will both be a measure of peripheral hearing impairment and of general cerebral function.

1.3.8 Open vs. Closed sets

The responses to different types of speech tests can be open or closed set. In an open set test (e.g. the AB word list Boothroyd, 1968) the subjects response is unrestricted. The subject is asked to repeat whatever they thought they heard regardless of any context so the extent of possible responses in unlimited .

In a closed set the possible responses have been deliberately limited which makes them easier to perform. The four alternative auditory feature test (FAAF) test (Foster and Haggard, 1987) offers the subject four possible alternatives from which to select. Closed set or forced choice material is usually constructed in lists of similar words presented in a written or touch screen format from which the subject can choose. The sensitivity of such tests can be altered by changing the response set used. Closed response tests can be used repeatedly for the same listener, as redundancy is low. Lyzenga (2005) suggests that closed response tests suffer more from learning effects. However Munro and Lutman (2003) demonstrate that repeated use of the FAAF test does not show significant practice effects.

1.4 Scoring

Compared to scoring keywords or whole sentences, scoring all the words separately requires more effort, time and skill of the tester but gives more fine-grained results. Using all the information in a sentence by scoring all the words or the whole sentence as a block is a better approximation of natural listening than scoring one or a few key words.

1.5 Development of speech tests

Three primary issues are fundamental to the development of a speech test. Firstly the test should be sensitive, demonstrating performance differences for individuals with normal hearing and individuals with hearing impairment. It should also be possible to demonstrate performance differences with different degrees of hearing loss. Second, the test should provide information regarding the underlying auditory processes that govern the perception of speech. Thirdly the test should provide data that can be applied for the purpose intended.

Test	**Language**	**Sentence length**	**Scoring**	**Noise**
Spin, Kalikow et al 1977	American English	5-8 words 6-8 syllables	Keywords % correct	12 talker babble
Dutch speech reception test Plomp & Mimpen 1979a	Dutch	8-9 syllables in total no word with more than 3 syllables	Whole sentences	Speech shaped noise
Alternative Dutch speech reception test Versfeld et al, 2000	Dutch	8-9 syllables in total no word with more than 3 syllables	Whole sentences	Speech shaped noise
Speech in noise sentence test (BKB) Bench et al, 1979	British English	21 lists of 16 sentences with up to 7 syllables	Correct key words	
Test for speech reception thresholds (IHR sentences) MacLeod and Summerfield 1990	British English	Average 5 words	3 keywords, all correct in sentence	Low pass filtered white noise
Closed set Swedish sentences for speech intelligibility Hagerman, 1982	Swedish	5 words	All 5 words	Speech shaped noise
Hearing in noise test HINT Nilsson et al, 1994	American English		Whole sentences all word correct	Speech shaped noise
German sentence in noise test (Göttingen sentences Kollmeier & Wesselkamp, 1997	German	Phonemically balanced lists	Each word is scored and weighted	Fixed level speech shaped noise
Closed set German sentence in noise test (Oldenberg sentences) Wagener, 1999a	German	5 words	All 5 words	Fixed level speech shaped noise

Test	Language	Sentence length	Scoring	Noise
Adaptive version of the Gottingen and Oldenberg speech tests Brand & Kollmeier, 2002	German		Word scores	Speech shaped noise
Dutch speech in noise screening test by telephone Smits, Kapteyn and Houtgast 2004	Dutch	Digit triplets	Complete triplet	Speech shaped noise

Table 1.2 Summary of basic features of common speech tests.

1.6 Development of an automated speech in noise test

The incidence of hearing loss inevitably increases with age; however many people who are aware that they are experiencing some difficulties with hearing do not seek any professional assessment of their hearing. There are also many others who seem unaware that they are experiencing any difficulties possibly because they can only make a subjective assessment of their own hearing ability. Therefore it has been proposed that an objective hearing test for home screening preferably without needing an instructor is needed (Smits and Houtgast, 2004).

The difficulty in understanding speech in noise is considered by many people to be the greatest handicap associated with their hearing impairment (Kramer et al, 1998). So a test which could measure this ability would fulfil these criteria. It has been shown that pure tone audiometry and speech in quiet are not good predictors of this ability (Smoorenburg, 1992). Different tests of speech intelligibility in noise have been developed (some of which have been described above) using sentences as the test material and using fixed signal to noise level or an adaptive procedure (Plomp and Mimpen, 1979a; Kollmeier and Wesselkamp, 1997; Nilsson et al 1994; Hagerman, 1982). The use of sentences instead of words as speech material has the advantage of being more representative of real life listening situations.

The ability to understand speech in noise is generally presented as the speech reception threshold (SRT) which is described as the signal to noise ratio required for a subject to recognize 50% of the speech material.

The stated goal of the project reported by Smits and Houtgast (2005) was *"to develop a SRT$_n$ test that can be done by telephone. The test should be easy quick and suitable for screening purposes (high sensitivity and specificity)"*. It was also stated that a strong correlation between the SRT$_n$ measured from the new test and the SRT$_n$ measured from the original standard Dutch speech in noise test should exist.

It was decided that digit triplets (e.g. 6-2-8) should be used as the speech material. The reasons for this decision were fourfold. Firstly digits are very commonly used words and hence very familiar. Secondly in contrast to sentences they can be repeated as the likelihood of the subject remembering the triplets used is low. Thirdly the use of triplets made possible the full automation of the test using a telephone which was connected to a computer which presents the test and judges the responses, which are given by pressing the telephone key pad. Fourthly it was decided that triplet would give a more accurate response than single digits.

An adaptive test procedure described by Plomp and Mimpen was used with ten extra presentations resulting in 23 presentations per SRT$_n$ measurement. The noise level is fixed and the level of speech varies. The triplet is judged correct only when all digits are entered correctly. The first triplet is presented in 4 dB steps until the triplet is entered correctly. The speech level is decreased by 2 dB and the second triplet presented. Based on the subject's response, the subsequent triplets are presented 2 dB higher (incorrect response) or 2 dB lower (correct response)

The SRT$_n$ is calculated as the average signal to noise ratio of triplets 5-24. The last triplet is not presented but is imputed from response to triplet 23.
It was decided that only those numbers which were monosyllables should be used, so 7 and 9 were removed as they have two syllables when spoken in Dutch. The remaining digits were 0,1,2,3,4,5,6 and 8.

Digit triplets were used to reduce the chance of a subject correctly guessing the correct response but not being influenced by cognitive ability. Increasing the number of independent items increases the measurement efficiency (Versfeld et al, 2000). Five lists were compiled made up of 23 different triplets (115 triplets in all). Triplets were chosen so that digits were distributed as equally as possible in the different positions.

All triplets were spoken by a female speaker with each digit pronounced individually and with pauses between digits. It was found that the last digit was pronounced more softly than the first so to equalize the intelligibility across the separate digits; amplitude was increased by 0 dB to 6 dB for every triplet. The noise was shaped to match the long term average speech spectrum.

The selection and equalization of the speech was carried out using eighty normally hearing subjects who completed between one and five of the test lists using their home telephone. The order of the triplets in each list was randomised for each subject and noise level had been fixed at 62 dB (A).

For every triplet presented the signal-to-noise ratio was corrected for inter-individual differences by adding the difference between the $\mathrm{SRT_n}$ for that individual and the average $\mathrm{SRT_n}$ for all individuals. As each triplet was presented at different signal-to-noise ratios during the adaptive procedure and it was known if the response at that level was scored correct or incorrect it was possible to fit a psychometric function to the data.

The function used was a logistic function given as

$$P(\mathrm{SNR}) = \frac{1}{1+\exp[-(\mathrm{SNR}-\mathrm{SRT_n})4s]}$$

Where SNR = signal-to-noise ratio, $\mathrm{SRT_n}$ = speech-reception thresholds (i.e. signal to noise ratio corresponding to 50% intelligibility), and s = slope of the psychometric function at 50% intelligibility.

Only triplets with steep slopes ($s \geq 9\%/\mathrm{dB}$) and $\mathrm{SRT_n}$s between −2 dB and −12 dB were selected for the final set of triplets. So there were 80 triplets in all with an average SRT_n of −7 dB. Equal intelligibility was achieved for all triplets by applying a level correction to them.

1.7 Development of a Europe wide sentence in noise test.

The availability of tests, testing procedures and usage of tests differ from country to country across Europe. This has meant that results cannot be compared across national borders or across languages. By developing speech test with a common structure that can be used in various languages it is hoped that comparisons across different populations and languages can be made as well as providing an agreed European standard for speech testing.

Three tests have already been developed: in Swedish (Hagerman sentences) German (Oldenburg sentence test) and in Danish (Dantale II test Wagener et al, 2003) using the same syntactic structure.

The Oldenburg sentence test determines the speech reception threshold in noise and in combination with an adaptive procedure. It has its basis in the Swedish Hagerman sentences (Hagerman, 1982). The sentences are of low predictability and follow the format name/verb/numeral/adjective/object. The sentences are drawn from a base list of ten sentences of five words each.

This base list approximates the mean phoneme distribution of the respective language. The sentences are then generated by randomly choosing one of the ten alternatives for each part of the sentence, so each list contains the same word material.

Index	Name	Verb	Numeral	Adjective	Object
0	Anders	owns	ten	old	jackets
1	Birgit	had	five	red	boxes
2	Ingrid	sees	seven	nice	rings
3	Ulla	bought	three	new	flowers
4	Niels	won	six	fine	cupboards
5	Kirsten	gets	twelve	lovely	masks
6	Henning	sold	eight	beautiful	cars
7	Per	borrows	fourteen	big	houses
8	Linda	chose	nine	white	presents
9	Michael	finds	twenty	funny	plants

Table 1.3 English translation of the basic test list of the Dantale II test.

One hundred sentences are spoken and recorded in a way that all words in a given column are recorded in combination with all words in the following column. Speech simulating continuous noise is derived from a random superposition of the words. Sentences are selected to have a high homogeneity; that is sentence specific SRT are restricted to −7.1 dB ± 0.16 dB.

The lists are optimised with respect to same number of phonemes, same number of words, approximate phoneme distribution and intelligibility over a range of SNR.

The sentences are spoken by a male speaker at a speech rate of 233 syllables/min taking co-articulation into account for a more natural sound.

Each test list consists of at least 20 short sentences. The subject listens to the sentence and then repeats what has been recognised. The experimenter then strikes out all words that have been incorrectly repeated.

The test uses an automated computer controlled adaptive procedure. The level of the subsequent sentence is based on the response to the previous sentence. This is calculated by

*Delta L = −f (i) *(prev- tar)/slope*

tar = target recognition value at which the procedure should converge.

prev = the recognition value obtained in the previous sentence.

slope = set to 0.151/dB which is median value of Göttingen and Oldenburg sentence test in normal hearing and hearing impaired subjects. *f (i)* controls the velocity of convergence and its value depends on the value of *i* of reversals of level.

The computer controlled procedure can either measure just the SRT (converges on 0.5) or both the SRT and the slope at SRT (has to converge on two concurrent recognition values 0.2 and 0.8). It is possible to use a manual adaptive procedure but it is less efficient. The manual measurement is divided into two parts. The first for a rough adjustment and the second for a fine adjustment of the SRT.

PART 1: Sentence 1 to 5		PART 2: Sentence 6 to 10 (20, 30)	
Number of correct responses	Delta L	Number of correct responses	Delta L
5	−3	5	−2
4	−2	4	−1
3	−1	3	0
2	1	2	0
1	2	1	1
0	3	0	2

Table 1.4 The two part adjustments of the SRT.

During the two parts the previous sentence changes the level of presentation level as above.

The result of the test is the individual speech recognition score at the chosen signal-to-noise ratio. The adaptive procedures converge on a speech reception threshold (SRT), a value in dB S/N. The whole speech recognition curve can be obtained if several measurements at different signal-to-noise ratios are made.

1.8 Rationale of study

There has been recent interest in the idea of producing a unified strategy of procedures and methods across Europe. The NATASHA consortium was set up to look at this idea and they have described various tests which the consortium believes are important tests currently, or will become important tests used in clinical Audiology for diagnostic and rehabilitative purposes across Europe (www. Phon.ucl.ac.uk/home/andyf/natasha.htm).

Up to now the availability of tests and testing procedures has varied greatly across European countries. In most cases it means that comparisons of results cannot be made over national borders, certainly not over language borders as many of the tests are language dependent. The consortium put forward various ideas as to which tests should be included in this strategy. Included in these tests were speech tests and in particular sentence tests.

The NATASHA consortium's preferred method of speech material to be used in the future was that of syntactically fixed but semantically unpredictable short sentences, as first described by Hagerman (1982).

The HEARCOM project followed on from the ideas put forward by NATASHA. The HEARCOM project has EU funding to develop a common set of speech recognition tests for use in a wide-ranging project aimed at mitigating the effects of hearing impairment in the information society. These common sets of speech tests include a sentence test as described above and a number recognition test. The reasoning behind the number test is that the test can be designed for self-completion over the telephone or the Internet (www.Hearcom.com).

There is a lack of such tests in the English Language and the present study is an attempt to create and develop English versions of these speech tests to add to this concept unified strategy of procedures and methods across Europe.

1.9 Aims of the study

Therefore the main aims of the study are firstly to create the sentence material to be used in an English language version of these speech tests; then secondly to evaluate each word in the material with regard to equal intelligibility. More specifically the aim is to assess whether each word is equally intelligible and therefore are all the sentences across the tests lists homogeneous.

Chapter Two: Experimental Design

2.1 Research objectives

The major objectives of the research were, firstly to create the materials needed for the synthetic sentence in noise test and the number triplet test; then secondly to obtain normative data using adult subjects, using these test materials. Specifically the study aimed to compare scores obtained at different signal-to-noise ratios for each word, with the overall scores of the combined word material. Finally the study would obtain for each word an amount in dB by which it needs to be corrected to make all words equally intelligible.

2.2 Creation of the test material

2.2.1 Synthetic sentences

Base Material

The synthetic sentences are generated from a base list of fifty words which have been combined into ten sentences of five words each.

Index	**Name**	**Verb**	**Numeral**	**Adjective**	**Object**
0	Peter	got	three	large	desks
1	Kathy	sees	nine	small	chairs
2	Lucy	bought	*five*	old	*shoes*
3	Alan	gives	eight	dark	toys
4	Rachel	sold	four	*thin*	spoons
5	*Barry*	*likes*	*six*	green	*mugs*
6	Steven	has	two	cheap	*ships*
7	Thomas	kept	*ten*	*pink*	rings
8	*Hannah*	*wins*	twelve	red	*tins*
9	Nina	wants	*some*	*big*	*beds*

Table 2.1 The word matrix of the fifty words from which the test material has been derived.

Many of the words in this base list had originally been used in an American English version of the test; the words which are in italics have been revised from the American test as it was felt they were not suitable for a British English version of the test. The revised corpus has equal numbers of phonemes within each column and is also phonetically balanced based on the phoneme frequencies of Fry (1961)

The syntactic structure of all the sentences is identical *Name verb numeral adjective object.* This structure has been used already in other tests which have been discussed in Chapter One: the German Oldenburg Sentence test and the Danish Dantale II test (Wagener et al, 2003), all these having been based on the Swedish Hagerman sentences (Hagerman, 1982). As already stated it is an aim of this project to create an English version based on these principles.

The base list of fifty words is then used to randomly generate the test sentences by selecting one of the ten possible alternatives for each part of the sentence, so each sentence is made up of the same word material.

2.2.2 Number triplets

Base material

The base list for the generation of the number triplets consisted of the digits zero (pronounced 'oh' in the test material) to nine. These numbers were chosen as they are found on a standard telephone keypad. One aim of the number triplet test is to create an English language test that can be completed over the telephone, like that already in use in Dutch language . To make the group homogeneous it was decided to use only words which were monosyllabic; this then excluded the number seven having two syllables. Having two syllables would make its identification during testing too easy.

Using digit triplets reduces the chance of a subject guessing the correct response and makes the measurements more accurate (Smits, Kapteyn and Houtgast, 2004). It is well known that increasing the number of independent items increases the measurement efficiency (Versveld et al, 2000).

These numbers are then randomly grouped together as three sets of nine number triplets. In the sets of number triplets each number appears in the first, second and third positions once.

Set 1	Set 2	Set 3
024	000	962
135	111	815
246	222	534
359	333	628
468	444	093
591	555	401
680	666	340
802	888	289
913	999	156

Table 2.2 Base sets of the number triplets, as originally recorded.

2.2.3 Recording the speech material

For the initial recording of the speech material one hundred sentences were generated in a manner which meant that each word in a given column would be recorded in combination with all the words from the following column. This was so that in the cutting of the speech material the correct co-articulation between words could be used, giving the sentences used in the final test a more naturally spoken pattern.

The one hundred sentences were randomly compiled into ten lists for recording, comprising ten sentences each. These lists were then recorded in three separate takes. These lists were recorded in a different order for each take. The same female speaker was used in each take. The speaker was instructed to maintain the same speed and pronunciation throughout each take. Any sentences in which it was felt that a difference in pronunciation could be detected were recorded again at the end of each take.

Three takes were used to provide enough recorded material so that the best material could be used and any words which were identified during the cutting as not ideal could be replaced by one from another take. The recordings were recorded initially onto Digital Audio Tape (DAT) in a recording studio with low reverberation, at a sampling rate of 44.1 kHz. These recordings were transferred digitally to CD as 16-bit waveform (.wav) files for editing.

2.2.4 Recording the number triplets

The number triplets were recorded at the same time as the synthetic sentences and under the same conditions. The numbers were recorded as three sets of nine triplets as shown in Table 2.2. The first and third sets were randomly generated with each individual number appearing in each position once and the second set as triplets of identical numbers (one, one, one, two, two, two etc).

These sets were recorded in a number of takes with the triplet sets being spoken at slightly different speeds. The sets of triplets were also recorded starting with the first triplet in the set and then starting with the last triplet in the set. As with the synthetic sentences the same female speaker was used and repeats of any triplet which it was felt did not fit the pattern were repeated after each set. Additionally the sets of triplets were recorded with and without a carrier phrase in this case ‘the digits’ followed by the triplet. The carrier phrase was included as in the completed test this would be included repeating the instruction to the subject for example ‘please repeat the digits’ followed by a number triplet.

2.2.5 Cutting the speech material

The different takes were examined by listening to the recordings and the one which was considered to be the best in terms of speed of delivery and pronunciation was selected as the starting point from which to begin cutting the speech material. As the final test materials are to be produced by combining the 10 alternatives for each word randomly the original one hundred recorded sentences need to be cut into single words.

The cutting was performed using the Adobe Audition program. This allows the sentences to be visualised as both waveforms over time and as spectrograms. The recordings of the speech materials were imported from the CD and stored in the program, each take as a separate file. Before cutting individual words the required sentence was selected by listening to the required take, highlighting the whole sentence and then copying it into a new file. The silence at the beginning and end of the sentences was then almost entirely deleted. A small duration of silence (15 ms or as close as possible) was left. After the silence had been removed the entire sentence was averaged with regard to RMS level.

An RMS level was selected after examining a number of sentences and noting their RMS values, as these were all found to be close to a value of −30 dB relative to maximum limits this figure was taken as the value to which all of the sentences should averaged.

Once the silence has been removed and the sentences averaged with regard to RMS level then it becomes possible to begin cutting the individual words. As the words will be chosen at random (with the constraint of correct co-articulation) it is important that the same procedure is used throughout the cutting process. The cutting point needs to be identified that separates the file into two parts so nothing is duplicated. To do this requires careful listening to find the point at which the second word of a selected pair begins. The sentences were cut into individual words and the saved as single files (.wav) except for the last two words of each sentence (the adjective and object can be left as one file).

The words were cut concisely at the beginning of each word, as if representing the first word of a new sentence, but included the co-articulation to the following word so that the following word would be perceived as naturally spoken.

The selection of these cutting points was undertaken by listening carefully to the recorded material using headphones. This gave the general position of the cutting point. It was then often necessary to examine the selected point with the sentence shown as a spectrogram; this made it easier to locate the point as which to cut and include the co-articulation.

The cutting points also needed to be made at the zero crossings of the low frequency components of the waveform. Each individual file need to begin with 0° phase and end with 180° phase so that correct phase can be maintained when the individual files are randomly combined together to generate the test sentences. This could only be done accurately by zooming the view of the file to the selected cutting point and adjusting its location to the nearest zero crossing point.

2.2.6 Labelling the files

The files of the cut material were uniquely labelled so that they could be identified not only by the word but also the co-articulation which is included. To do this the base list of the sentences was given an index. The sentences were indexed by the numbers 0-9 and the letters a, b, c, d, e for the word types. Each file was then labelled by the word and the co-articulation.

For example if we take the sentence, “Lucy has five cheap shoes”, the word ‘has’ would be labelled firstly as the sixth verb of the base list (see table 2.1) b6 and secondly by the co-articulation of the following word in this case ‘five’ which is the second number of the base list hence c2. The file label would therefore be ‘b6c2.wav’. The adjective-object combinations were labelled in the same way, for example ‘cheap shoes’ would be ‘d6e2.wav’.

2.2.7 Cutting the number triplets

The number triplets were examined by listening and it was decided to reject the triplet sets made up of one number repeated three times as the pronunciation sounded different with the number repeated. This left two sets randomly generated triplets. The sets which it was felt were spoken at the correct pace were then selected from the different takes for cutting.

The same software and cutting procedure as described for the cutting of the synthetic sentences was used, except it was not necessary to consider co-articulation as the digits were pronounced discretely.

The labelling of the number files was undertaken in a different manner. As there were only two sets of nine triplets one was called set one and the other set two. The individual waveform files were then labelled by number, position and set.

For example using the triplet 0-2-4, the two would be labelled 2b and the set from which it came in this case set 1 so the file would be labelled 2b1.wav.

2.3 Generating new test material

2.3.1 Generating new test sentences

The cutting of the sentences produced a total of 400 waveform files which could be used to generate new test sentences.

The generation of new test sentences was performed using a specially written program which randomly generated a new list of ten sentences by combining the cut waveform files. In constructing the sentences a word in a given column is selected to produce the correct co-articulation for the following word, regardless of the previous word. The sentence generating program produced a single waveform file for each new sentence of each of the twenty lists of ten sentences (two hundred in totals) each made up of the four original waveform files. These were labelled 0101.wav for the first sentence of list one, 0102.wav for the second sentence of list 1 and so on.

The program also produced a text file for each of the lists of ten sentences describing the contents of each sentence. The text files needed to be cut and pasted into a new text file combining all the text files from all of the lists.

Once all the new test sentences were created it was then necessary to check them by careful listening to improve or remove any sentences which did not seem naturally spoken or contained clicks or other audible unwanted inclusions. It was found by listening to the sentence material that the female speaker had found the word pair ‘cheap chairs’ difficult to pronounce and this was carried through to the final sentences, so any sentences with this word combination were removed. New sentences were generated in place of these. No other major problems were found with the generated material.

2.3.2 Generating the test triplets

To generate the new test triplets a similar program was used to that which generated the synthetic sentences. The program randomly selected three numbers from one of the two original sets. The numbers from each set were kept separate and each new test list of nine triples was sourced from only one set. Each set was used alternately to generate a list. Twenty lists of nine triplets were generated (ten from each original set) one hundred and eighty triplets in to total.

The individual triplet files also contained the carrier phrase ‘the digits’ and three 10 ms gaps of silence all originally in separate files. So the new triplet file 4-9-8 would be heard by the subject as ‘the digits’ silence ‘Four’ silence ‘nine’ silence ‘eight’. These seven separate files would then be saved as a single complete file. This file would then be labelled by the test list and its position in the list; for example, triplet 2 in test list 1 would be 0102.wav. As with the synthetic sentences the program also produced a text file describing the triplets these also needed to be compiled into one file.

2.4 Generating the noise

It is the sound of one or more competing speakers that provides the major source of interference in common everyday listening situations so it is preferable when designing tests such as these to use noise with the long term spectrum of speech. In addition, the noise should have the same spectral content as the speech signal. So by using the original speech material for each test as the source for the competing noise this can be achieved. An unwanted effect of having different spectral contents in the speech and in the noise would be for some frequencies to get more weight than others (Plomp and Mimpen, 1979a).

The interfering noise can be generated by superimposing the different types of speech material, in this case the synthetic sentences and the number triplets. So a unique set of noise is created for each test. In fact, an alternative approach was used here. For the sentence test, the long-term average spectrum of the corpus of material was calculated and then white noise was filtered to have the same spectrum. The same approach was used for the triplet test using the corpus of recorded digits.

The long-term spectrum of the resulting noise is comparable to the mean long term average speech spectrum of various languages (Byrne et al, 1994).

Once the noise files had been generated all the materials required to run the synthetic sentence and number triplet test in noise had been created. All these files were made available to the Automated Sentence Test (AST) software, which was used to run the actual sentence and triplet tests, as described below.

2.5 The Automated Sentence Test

The Automated Sentence Test is designed to run sentence testing primarily with the BKB sentence materials, and has been successfully used in a number of studies undertaken at Institute of Sound and Vibration Research. In this instance the synthetic sentences and the number triplets were both added to a new version of the test. Additionally the noises which had been generated for each of the tests (using the test material) were also made available in the software.

The design of the software allows the levels of speech and noise to be altered by the tester. The signal-to-noise ratio was computer controlled following the general principles outlined by Lutman and Clark (1986). The speech level was fixed and the signal-to-noise ratio was altered by adjusting the background noise level. For the purpose of this study the speech level was fixed at a free-field equivalent sound pressure level of 55 dB. This was achieved via headphones, which had been calibrated to produce 65 dB in the IEC artificial ear.

The software allows the tests to be run with the noise being altered adaptively, or run non-adaptively.

2.5.1 Adaptive testing

The adaptive method used a 2-down-1-up rule as described by Levitt (1971). The subject must respond correctly twice in succession for the task to become more difficult by one step, but only once for the task to become easier by one step. A series of steps in a single direction is defined as a run, whilst a change in direction is termed a reversal. The 2-down-1-up rule tracks the subject's performance around a level of 70.7% correct. In this study the adaptive staircase continued until nine reversals had occurred. A larger step size (6 dB) was used initially to alter the noise level but after three reversals the step size was reduced to 3 dB. The final score was calculated by averaging the noise level at the final six reversal points.

This adaptive technique was used for both the synthetic sentences and the number triplets to establish the subjects initial thresholds for the test material in both the left and right ears. This threshold figure would then act as a reference point to calculate the levels of the fixed signal-to-noise ratios required in the non-adaptive testing.

2.5.2 Non-adaptive testing

The main bulk of the testing was performed using non-adaptive methods according to the method of constant stimuli. In order to plot the intelligibility function of each individual word and to be able to determine both the 50% correct level as a dB value and the slope of the function, scores above and below 50% would also be required. It was initially decided to have three fixed signal-to-noise ratios approximating to the scores of 30%, 50% and 70% correct word scores. These fixed levels were to be determined using a small pilot study (for details of pilot see section). Following the pilot study, it was decided to have four fixed signal-to-noise ratios, principally as the number of sentences and triplets was not easily divided into three. The four fixed signal-to-noise levels were –2, –4, –6 and –8 dB from the initial level determined using the adaptive method described above.

The synthetic sentences test session was subsequently split into four parts. Each part would be made up of five test lists of ten sentences presented at one of the four fixed signal-to-noise ratios. Similar procedures were used for the triplet test.

Randomisation was used to reduce any order effects. Each subject started with a different test list (subject 1 with list 1, subject 2 with list 2, etc) and the four signal-to-noise ratios were randomly attributed to one of the parts.

The right and left ears were also alternated after every two parts of the session (10 lists) and between subjects.

Whether the subject started with the synthetic sentences of the number triplets was also alternated between subjects.

2.6 Subject sample

The instructions for recording, cutting and validating Oldenburg sentence test types (Wagener, 2005) have guided much of the work of creating the test materials described here. It is suggested that in these instructions that 12 or more normal-hearing subjects should be used during the optimisation measurements.

It was possible for 14 subjects to be recruited 11 female and 3 male. The subjects consisted of otologically normal adults. Each subject was only included in the study if they were able to meet certain criteria outlined below.

Subjects aged 18-30 years.
Hearing threshold levels of ≤20 dB HL across all frequencies and in both ears.
No occluding wax present in ear canal.
Had English as a first language. The subjects used for the optimisation of the word material need to be English native speakers so that they are familiar with all of the words of the base material.
No history of ear disease.
No previous operations on their ears.

No history of exposure to loud sounds. Subjects not to have been exposed in the past to significant durations of noises in the work place, gunfire of explosions.

No exposure to loud sound in past 48 hours. Subjects who have been exposed to recent loud sound may have some temporary threshold shift.

No significant tinnitus. This may affect a subject's ability to distinguish the speech in noise.

Not suffering from any colds or congestion.

No medical or other reasons that may prevent them from taking part.

All prospective subjects were asked to complete a screening form (see Appendix 1) and were assessed using otoscopy, pure tone audiometry and tympanometry.

2.7 Experimental variables

The dependent variable of this study is individual word score, expressed as how many times each word has been scored correctly by the subjects. The main independent variable is that of noise level. Other factors which are capable of influencing the results were deemed undesirable and the effects of these were minimised wherever possible as outlined below.

Familiarity of word material. Different subjects may be predisposed to pick certain words over others when they are finding it difficult to hear in the presence of competing noise. It is possible determine the familiarity of words within the base material by examining data for the occurrence of particular words in everyday usage. However by giving the subject a matrix of all the words used in the test and a list of all the digits this limits the possible responses of the subject in effect making the responses a closed set, this should make all the word equal with respect to familiarity of the words.

Variance within subjects. Levels of performance may be influenced by varying levels of motivation, concentration and effort (resulting from the subject being distracted or uncomfortable, or because of the subject's attitude towards the task or tester). These are the most difficult variables to control. Verbal encouragement was given to the subject by means of a simple acknowledgement of a response. All the subjects were given identical instructions at the beginning of the task to prevent inconsistencies arising due to misunderstanding the instructions. Breaks from testing were offered after the completion of each quarter of the test protocol.

Tester/equipment. For the duration of the data collection period the tester and all the equipment remained the same.

Familiarity with the test material. None of the subjects were acquainted with the test material prior to taking part in the study.

2.8 Equipment and testing

2.8.1 Arrangement of equipment

A Heine otoscope with disposable speculae and Grason Stadler GSI 33 tympanometer were used for screening. A Grason Stadler GSI 16 clinical audiometer in conjunction with Telephonics TDH 50P earphones was used to present the speech materials.

A Hi-Grade computer with 16-bit soundcard conforming to the Windows sound system requirements and running the Automated Sentence Test v8.01 provided the speech materials. The output from the soundcard was fed to the Tape/CD input of the audiometer.

2.8.2 Test room

The test room was a quiet room situated in the Hearing and Balance Centre, Institute of Sound and Vibration Research, University of Southampton. Figure 2.1 illustrates the layout of the room during testing.

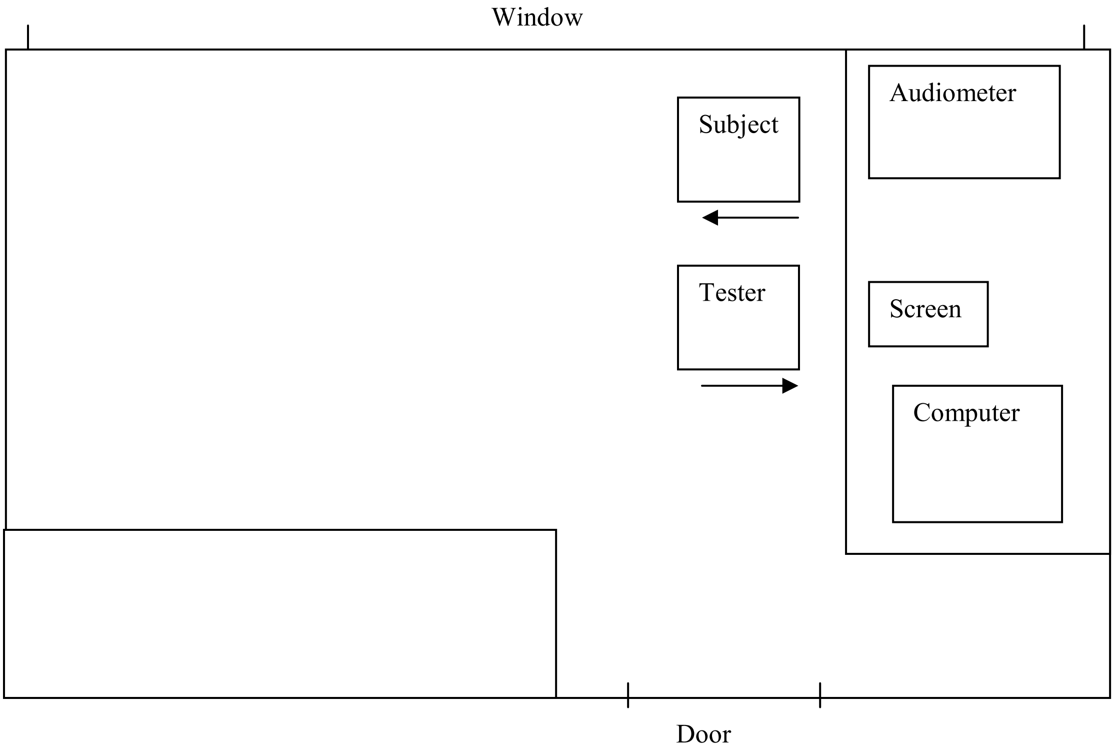

Figure 2.1 The layout of the room during testing.

For speech testing, the subject was seated comfortably facing away from the computer screen. They heard the test materials through earphones and repeated what they heard to the tester, who scored the responses on the computer.

Otoscopy, PTA and tympanometry were performed in a separate sound-proofed room also situated in the Hearing and Balance Centre.

2.8.3 Daily checks

The Automated Sentence Test software was checked by listening to synthetic sentences and the number triplet to confirm that they were being correctly presented. The output was calibrated using the 500 Hz test tone provided by the software to set the input sensitivity of the audiometer to a predetermined level, in this case 0 dB VU meter reading.

2.8.4 Test procedure

The order of the test procedure was the same for all of the fourteen subjects, with the exception of randomisation and balancing of conditions described above.

Explanation

The purpose and nature of the experiment was explained to each subject as well as the expected duration of the session.

Consent form

All the subjects were asked to complete and sign a form consenting to be a volunteer participant in the experiment in accordance with the Human Experimentation Safety and Ethics Committee guidelines.

Screening form

Prior to beginning any testing each subject was asked to complete a screening form (see appendix). The form had been designed to give information regarding the age of the subject, history of any ear diseases, the presence of tinnitus, any history of or recent noise exposure any recent illnesses. As well as confirming that English was a first language.

Otoscopy

Otoscopy was carried out on each subject prior to testing. This was to identify any wax or foreign bodies as well as the condition of the tympanic membrane and external auditory meatus.

Audiometry

Pure Tone Audiometry was performed on all subjects in accordance with British Society of Audiology recommended practice. The subjects were required to have thresholds ≤20 dB HL across all frequencies and in both ears.

Tympanometry

Tympanometry was carried out on all the subjects before testing to confirm that they had middle ear pressure and compliance within normal limits. Normal middle ear pressure was taken to be between −50 and + 50 daPa and a middle ear compliance of 0.3 to 1.5 ml was considered to be normal. The shape of the compliance peak was also examined.

Once the screening had been satisfactorily completed the testing was able to commence. The order in which the synthetic sentences or triplets were presented, and to which ear would have already been determined prior to the subject's arrival. The subject would be seated facing away from the computer screen. The subject was then asked to read the following instructions (if the session was to begin with the synthetic sentences)

You will be played 20 lists of 10 sentences (200 sentences in all) through headphones. Some sentences will be easier to understand than others. You will hear sentences in both your left ear and your right ear (but not at the same time).

All the words which make up the sentences are found on the word matrix provided. Each sentence is made up of five words one from each of the columns.

During the test you should listen to each of the sentences as they are played in turn then repeat back what you have heard to the tester. If you are unsure of a word (or words) then select a word from the matrix which you think it might have been.

A new sentence will not start until you have repeated your answer to the previous sentence.

If you have any questions ask the tester before beginning the testing.

After confirming that they had fully understood the instructions, they were given a copy of the word matrix (see Table 2.1) and a few moments to familiarise themselves with the information. The earphones were then carefully placed over the ears.

The adaptive method was selected and the noise level set to 55 dB (i.e. a signal-to-noise ratio of 0 dB) making the task relatively easy at first. This allowed the subject to become familiar with the material and the test procedure as well as the task of listening in background noise.

The sentence material would then be presented to the subject in one ear only, depending on how the session had been divided. The subject then repeated the sentence back to the tester. The tester then scored the sentence accordingly. In this adaptive procedure it was only possible to score the entire sentence correct or incorrect. In this context, correct was defined as repeating all words correctly. After nine reversals a noise level in dB was given and this was noted. This level was found for both ears separately and used to calculate the noise levels needed to obtain the four fixed signal-to-noise ratios already discussed.

Once these noise levels had been calculated the non-adaptive method was selected and the noise level altered. The subject was then presented with five test lists at one of the selected signal-to-noise ratios. The subject again repeated the sentences back to the tester but this time each word was scored (correct or incorrect) individually. The data generated by the Automated Sentence Test was saved to disk after each test list was completed. After five lists at one of the fixed signal-to-noise ratios the subject was asked if they would like a break or wished to move on to the next set of five test lists. Following any break the next five lists would be presented at a different signal-to-noise ratio. This would then continue until all twenty lists had been completed at the four signal-to-noise ratios.
A compulsory break of between five and ten minutes was then taken before commencing with the number triplet test.

Before beginning the number triplet test the subject was asked to read the following instructions.

You will be played 20 lists of 9 number triplets (180 triplets in all) through headphones. Some number triplets will be easier to understand than others. You will hear number triplets in both your left ear and your right ear (but not at the same time).

The numbers which make up the triplets are found on the sheet provided.

During the test you should listen to each of the triplets as they are played in turn then repeat back what you have heard to the tester. If you are unsure of a number *(or numbers) then select a number from the sheet which you think it might have been.*

A new triplet will not start until you have repeated your answer to the previous triplet.

If you have any questions ask the tester before beginning the testing.

After confirming that they had fully understood the instructions, they were given a copy of the numbers and a few moments to familiarise themselves with the information. The earphones were then carefully placed over the ears.

The same procedure was then followed as described above for the synthetic sentences. The adaptive method was used first to determine the subject's threshold then the non-adaptive method using the same four signal-to-noise ratios. Breaks were offered at the same intervals. The data was also saved to disk. That concluded the test session.

2.9 Pilot study

The current study was given approval by the Human Experimentation Safety and Ethics committee in July 2005. Following receipt of this a pilot study was undertaken.

The main reason for undertaking a pilot study was to evaluate which signal-to-noise ratios gave word scores which approximated the 30, 50 and 70 percent correct levels.

It also gave the tester experience in administering the test, including practice in scoring the test and saving the data. It also helped ensure the smooth running of the test protocol highlighting any modification needed to the procedure and the instructions, and gave an idea of the time needed to complete a full session of testing. In this way tester practice effects were reduced.

2.9.1 Subject sample

Two adults agreed to participate in the pilot study. Both had hearing and middle ear function within the limits already outlined. After the completion of the test they were asked for any comments relating to the running of the tests.

2.9.2 Results of the pilot study

The subjects were given the same instructions as above. The adaptive technique was then used to find the threshold at which to start. They were then played a single list of ten sentences and the number of words correctly scored noted down. Depending on the % correct the signal-to-noise ratio was either increased or decreased until scores approximating 30, 50 and 70 percent correct had been achieved. The noise levels were then noted down and a further list at the same noise level presented to confirm that the score was repeatable.

Synthetic sentences

	Adaptive threshold (dB)	Noise level for 30% correct (dB)	Noise level for 50% correct (dB)	Noise level for 70% correct (dB)
Subject: 1	59	62	64	66
Signal-to-noise ratio		−3	−5	−7
Subject: 2	59	62	65	66
Signal-to-noise ratio		−3	−6	−7

Table 2.3 Pilot study results synthetic sentences.

Number triplets

	Adaptive threshold (dB)	Noise level for 30% correct (dB)	Noise level for 50% correct (dB)	Noise level for 70% correct (dB)
Subject: 1	61	63	67	68
Signal-to-noise ratio		−2	−6	−7
Subject: 2	62	65	68	69
Signal-to-noise ratio		−3	−6	−7

Table 2.4 Pilot study results number triplets.

Having examined the results it was noted that suitable signal-to-noise ratios for 30, 50, and 70% correct would be −7, −5 and −3 for the sentence test and −7, −6 and −2 for the triplet test, all relative to the threshold obtained on the adaptive test.

It was at this point that it was noted that due to the number of lists, three fixed noise level would make the division of the sessions uneven. It was therefore decided that four fixed noise levels would be used. The levels of −2, −4, −6 and −8 were then chosen as they fitted the spread of the previous levels and would still allow the individual word functions to be accurately plotted, and equal division of the sessions.

Both the subjects experienced few difficulties with the test procedure and commented that the instructions described the protocol adequately. Both subjects did make enquiries about how well they were doing. To avoid affecting the subjects motivation it was decided that instead giving a direct answer the subject was encouraged to carry on and assured that they were ‘doing well’.

The pilot further aided the tester with familiarisation with both the administration of the test and scoring. It also gave an insight into potential subject reactions and responses. The pilot also gave an insight to the length of the entire test session. This was now gauged to be approximately from one hour thirty minutes to two hours depending on the time taken by the subject to respond.

Chapter Three: Results

3.1 Subjects

Of the fourteen subjects who comprised the original sample all passed the screening process and were therefore eligible participate in the study.

3.2 Raw data

The data obtained from the study required a small alteration to be made to the Automated Sentence Test (AST) operating procedure to allow the data to be exported. The data had to be first exported from the AST software then cut and pasted into a new file so that analysis could begin. The analysis was undertaken using Microsoft Excel 2003 and SPSS for Windows version 13.0. Once the data had been input it was checked for transcription errors.

3.3 Exploration of data

3.3.1 Synthetic sentences

Before exploring the data obtained for each individual word, an exploration of the mean correct scores against noise level for each of the twenty lists making up the synthetic sentence test was undertaken. Pooling the data for each list and the four fixed signal-to-noise ratios it was possible to derive mean scores for each list at the various noise levels used and to plot this information graphically.

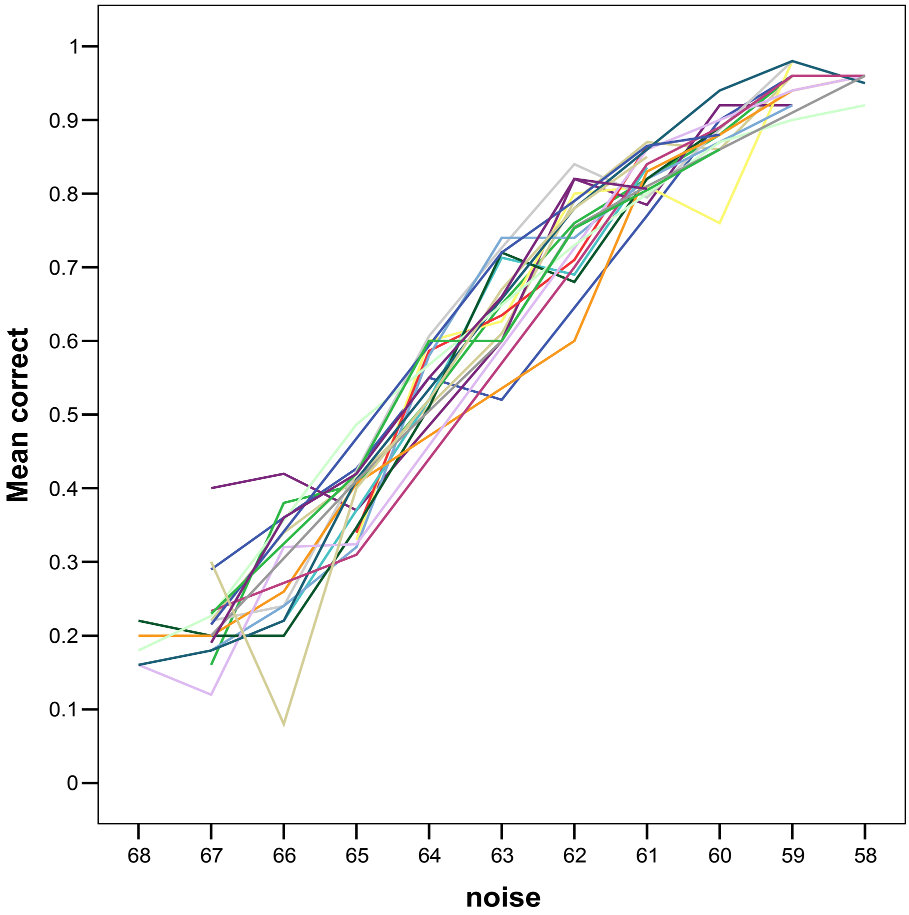

Figure 3.1 Proportion correct word scores against noise for individual synthetic sentence lists. Each curve is for a different list (1-20).

Inspection of the Fig 3.1 reveals that for each of the twenty lists as the relative noise level is reduced (the speech presentation level being fixed at 55 dB) the mean correct score increases. This follows the expected pattern of distribution for speech tests as the signal-to-noise ratio is improved in favour of the subject.

It can also been seen that as the noise level is increased (particularly beyond 65 dB) that the individual lists become more widely dispersed.

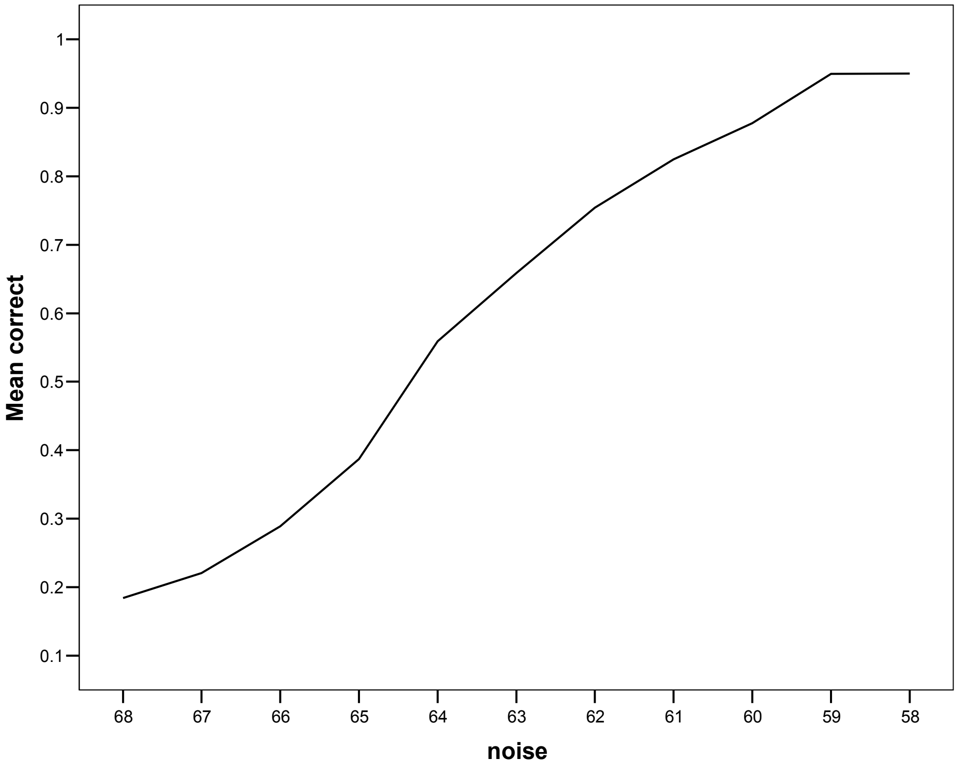

Figure 3.2 Proportion correct word scores plotted against noise for entire sample population of synthetic sentences.

Figure 3.2 shows the distribution of mean correct scores against noise for the entire sample population of fifty words in the synthetic sentence test combined. This also follows the expected trend that as the noise level is decreased relative to the fixed speech level of 55 dB the number of words scored correctly increases.

3.3.2 Number triplets

The data for the number triplets will be examined in a similar manner to that described above for the synthetic sentences.

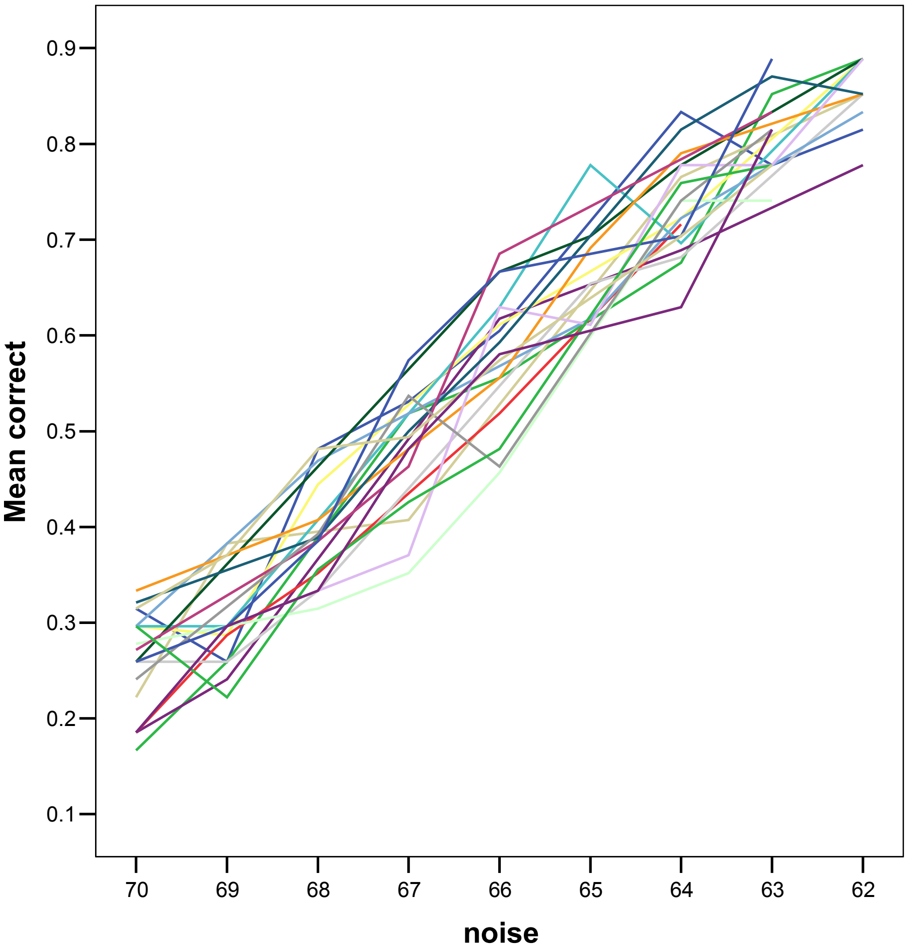

Figure 3.3 Proportion correct word scores against noise for individual number triplet lists. Each curve is for a different list (1-20).

Fig. 3.3 reveals a similar picture to that of the sentences. For each of the twenty lists as the relative noise level is reduced with presentation level being fixed at 55 dB the mean correct score increases.

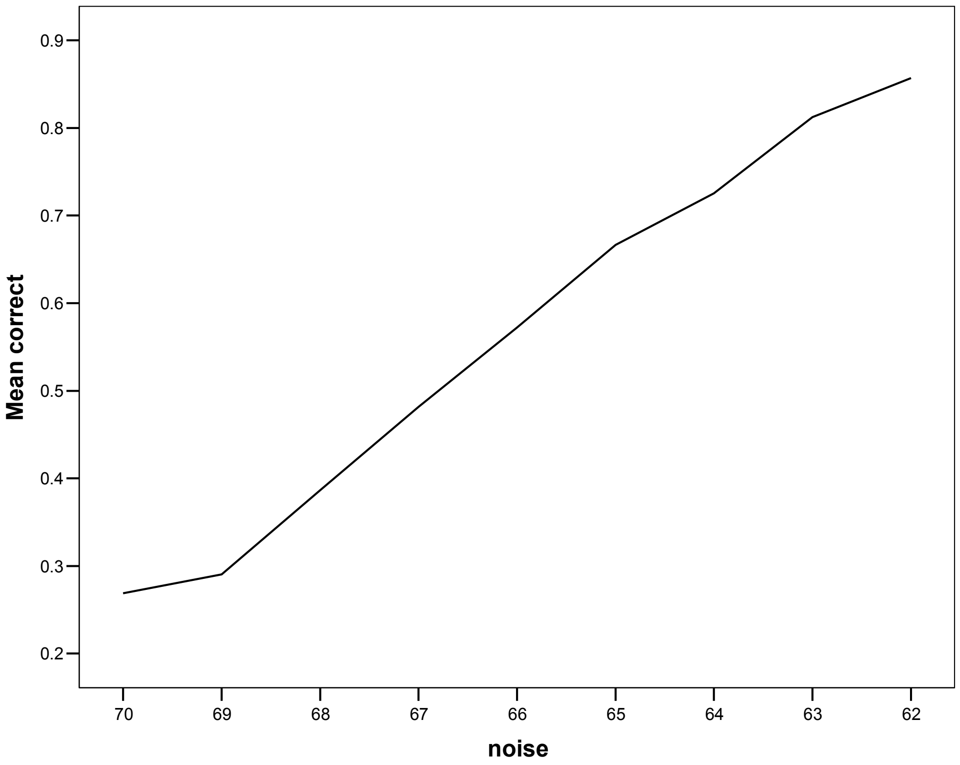

Figure 3.4 Proportion correct word scores against noise level for the entire sample population of number triplets.

Figure 3.4 shows the distribution of mean correct scores against noise for the entire sample population of all the words comprising the number triplet test. This can be seen to be following a similar trend to that described above.

3.4 Individual word scores

It is the main aim of this study to be able to derive the mean noise level in dB at which each word is correctly scored 50% of the time. This can then be compared to the overall noise level in dB at which all the words combined are correctly scored 50% of the time.

The comparison of these two dB noise levels will then reveal how much the mean of each individual word 50% correct score differs from the overall mean 50% correct word score for all the words combined. This difference then becomes the amount that each individual word needs to be adjusted so that the words all become equally intelligible. This adjustment can be done relatively easily using the same software that was used in the cutting of the test materials.

The derivation of these 50% noise level has been achieved using similar plots to those described above. The data relating to each individual word are first selected and the mean word scores plotted against the noise levels relating to the fixed signal-to-noise ratios. A trend line (2nd order polynomial) was added to all of the graphs. The intersection of this trend line and the 50% correct score (0.5 on the Y axis) was then used to establish the relative noise level at which 50% correct was achieved.

3.4.1 Synthetic sentences word scores

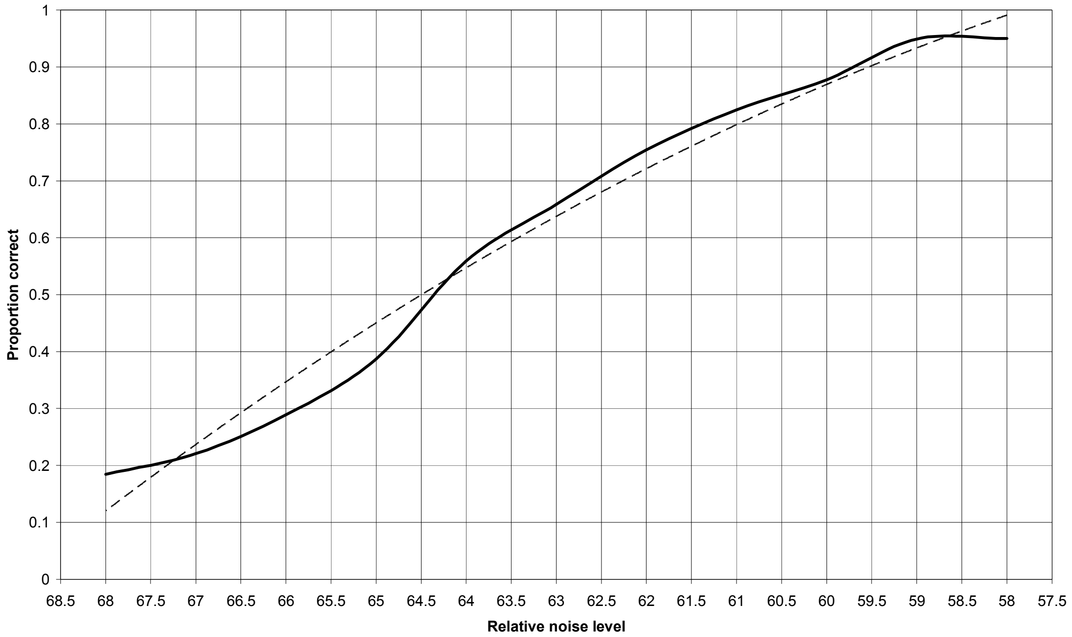

Figure 3.5 Proportion correct across all words for synthetic sentences combined as a function of noise level.

Figure 3.5 shows the plot (the unbroken line) for all the words of the synthetic sentences combined and plotted against the noise levels representing the fixed signal-to-noise ratios. The dashed line is the trend line for the data. The intersection of the 50% correct point on the Y axis and the trend line gives the relative noise level at which all the words are scored correctly 50% of the time (a value of 64.5 dB in this case).

The same analysis was then applied to each of the fifty words. Firstly deriving the mean correct word scores for the different noise levels. Then secondly representing the data graphically and reading off the relative noise level at which the 50% correct intersects with the trend line.

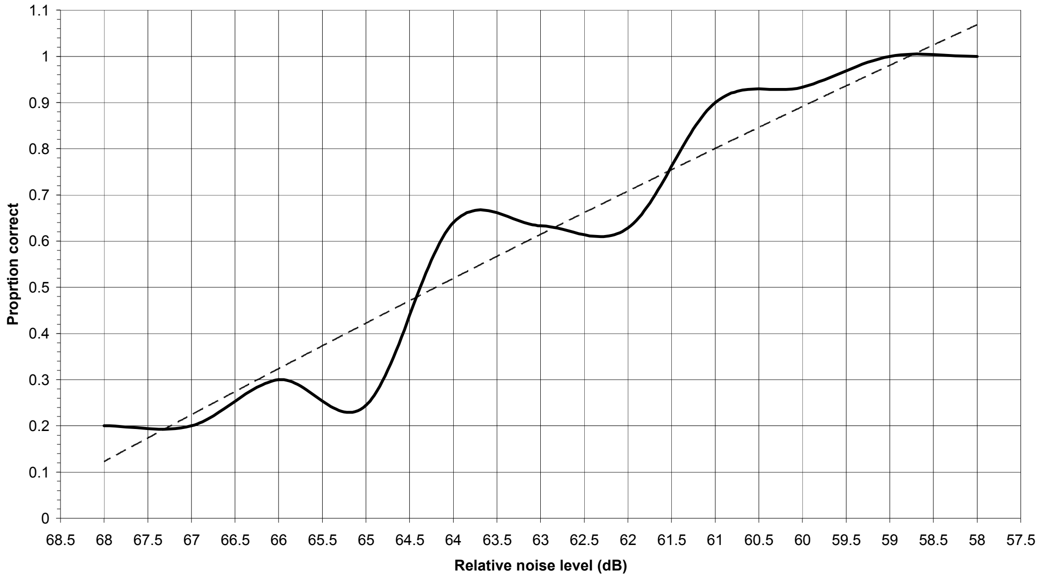

Figure 3.6 Proportion correct for word “Barry” as a function of noise level.

Figure 3.6 illustrates individual words correct against relative noise level for the word ‘Barry’. The dashed line is the trend line and its intersection with the 50% correct score line gives a relative noise level of 64.2 dB.

Comparing this noise of 64.2 dB with the overall level of 64.5 dB gives a relative value of –0.3 dB. This is the amount that the word ‘Barry’ needs to be adjusted to make both values equal. This procedure needs to be carried out for all the fifty words in the synthetic sentence test and once all the adjustments have been made all the words should be equally intelligible. This data is however an average across the ten different version of the word “Barry” which were originally recorded (which may not all be equally intelligible). Likewise the triplet data each number was cut from three different positions in two versions.

3.4.2 Number triplet word scores

The number triplets were analysed using the same methods described above.

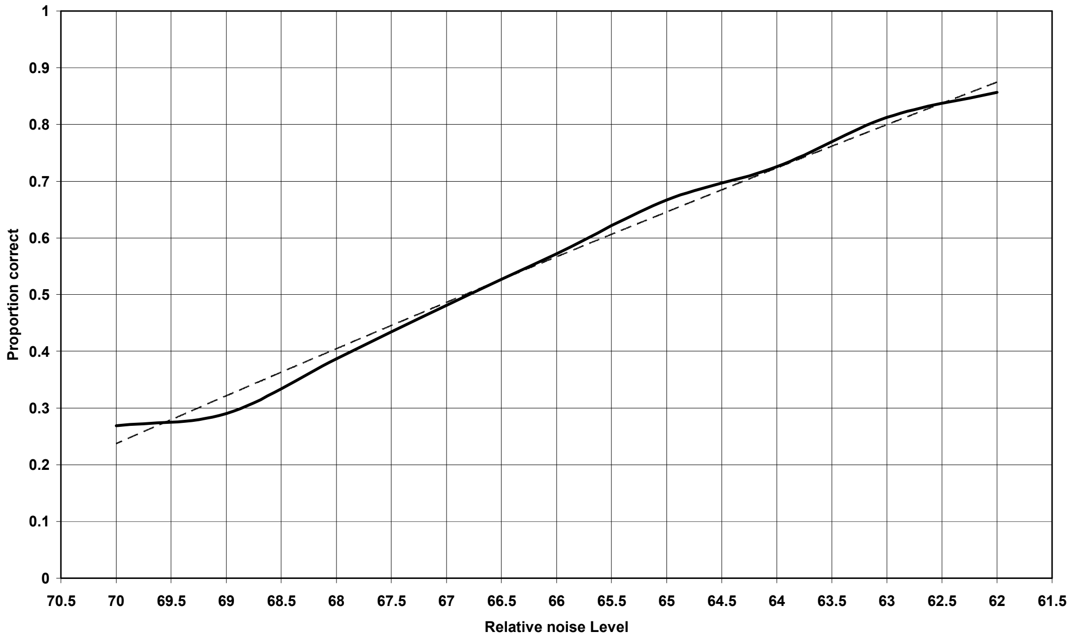

Figure 3.7 Proportion correct across all words for number triplets combined as a function of noise level.

Figure 3.7 shows the plot for all the words in the number triplet test. The noise level corresponding to 50% correct in this case is 66.8 dB.

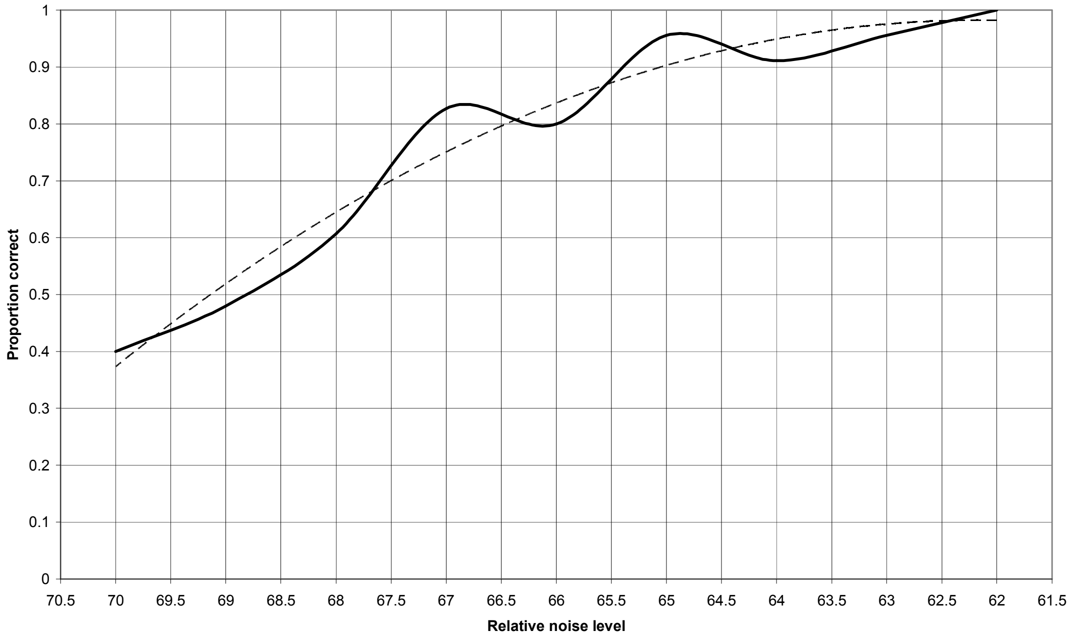

Figure 3.8 Proportion correct for the triplet digit “eight” as a function of noise level.

Figure 3.8 shows the plot for the triplet digit “eight”, where it can be seen that the noise level corresponding to 50% correct is 69.1 dB.

Each of the nine words of the triplet number test was represented in this graphical manner. The relative noise levels representing scores of 50% correct could then be measured for each word.

3.5 Summary of data

Word	Noise level at 50% correct	Correction in dB
Peter	64.0	–0.5
Kathy	62.3	–2.2
Lucy	66.7	2.2
Alan	68.5	4.0
Rachel	68.7	4.2
Barry	64.2	–0.3
Steven	65.0	0.5
Thomas	65.4	0.9
Hannah	57.3	–7.2
Nina	63.2	–1.3
got	63.5	–1.0
sees	65.4	0.9
bought	63.7	–0.8
gives	63.4	–1.1
sold	64.3	–0.2
likes	66.0	1.5
has	67.3	2.8
kept	63.6	–0.9
wins	60.5	–4.0
wants	62.8	–1.7
three	63.8	–0.7
nine	65.4	0.9
five	64.9	0.4
eight	66.5	2.0
four	65.6	1.1
six	66.7	2.2
two	66.1	1.6
ten	65.5	1.0
twelve	64.6	0.1
some	65.7	1.2

Word	Noise level at 50% correct	Correction in dB
large	64.9	0.4
small	63.9	–0.6
old	63.3	–1.2
dark	65.1	0.6
thin	57.6	–6.9
green	63.6	–0.9
cheap	65.4	0.9
pink	60.5	–4.0
red	63.4	–1.1
big	60.8	–3.7
desks	65.4	0.9
chairs	65.6	1.1
shoes	66.1	1.6
toys	63.7	–0.8
spoons	64.1	–0.4
mugs	64.7	0.2
ships	66.0	1.5
rings	62.8	–1.7
tins	64.4	–0.1
beds	62.1	–2.4

Table 3.1 summaries of the noise levels representing 50% correct scores and the level adjustment required for each word in the synthetic sentence test.

Number	**50% correct**	**correction in dB**
oh	65.5	–1.3
one	64.3	–2.5
two	68.1	1.3
three	62	–4.8
four	65.7	–1.1
five	66.4	–0.4
six	71.4	4.6
eight	69.1	2.3
nine	65.5	–1.3

Table 3.2 summaries of the noise levels representing 50% correct scores and the level adjustment required for each word in the synthetic sentence test.

Chapter Four: Discussion

4.1 Limitations of study

The experimental goals of this study were to collect normative data from a sample of adults for two newly created speech-in-noise tests; then to use the data relating to the intelligibility of each word to calculate word specific values in dB which when applied would make all the words equally intelligible. Prior to drawing any conclusions the limitation of the experimental design should be discussed.

Individual words/ individual word files.

The main objectives as discussed above have been to calculate a word specific values in dB which when applied would make all the words equally intelligible. This has been achieved; however each of the fifty words in the synthetic sentence test is actually made up of ten different waveform files (four hundred in total as the adjective/object files are in the same file) which have been cut from different source sentences in the original recorded material. This means that either further work is required to derive these correction values for each individual waveform files or a method determined to use the values already calculated for all the different files.

A similar limitation is found in the number triplets the study has only calculated nine correction values but the test has been created from 54 individual waveform files.

Adjective and Object files.

The adjective and object word at the end of each sentence have been cut together into the same waveform file, however adjustment values have been calculated for each separate word. This then becomes a problem when applying the adjustments without splitting the two words. For example, the adjective object combination “thin spoons” requires and adjustment of −6.9 dB for the word “thin” and only a −0.4 dB adjustment for the word “spoons”.

Sample size.

Whilst the size of the sample fits in with the guidance set out in Wagener’s 2005 notes for validation of these types of speech test, the larger the sample size the more representative of the normative data will be of a wider population. If time had not been a factor then a larger sample of subjects could have been used.

Subject recruitment

The subject sample was taken from the University of Southampton student population and therefore was not balanced in terms of socio-economic group or racial background. Whilst no subjects were excluded as a result of the screen process, it was not a particularly easy task recruiting subjects in light of similar experiments taking place which offered monetary inducements for participation.

Sex distribution

The subjects recruited were 11 females and only 3 males, so there is a significant bias towards females in the sample population. This may have had no effect on the results, but this has not been statistically tested. It is possible that this has had an effect on the results as Lutman (1991) reveals that based on a sample of over 1000 subjects, females performed 2.5% better on performance tests (including sentences in noise) than males. In the further evaluation of the test material it may be wise to recruit equal number of both sex and investigate any statistical difference in the scores.

Attention span/ motivation

This could potentially vary within and between subjects as a result of discomfort or distractions, although efforts were made to reduce that likelihood. It was however perceived by the tester that there were noticeable differences in the amount of effort that different subjects gave to the task especially at the lower signal-to-noise ratios.

4.2 Correction of individual word levels

The study has succeeded in producing a speech intelligibility function for all the words in the synthetic sentence test and the triplet number test. It has also been possible to use these functions to determine the relative noise level of the position at which each word is correctly scored 50% of the time. This has then been compared to the function for the entire sample populations of the respective tests and the relative noise level that represents correct word scores of 50% has been determined. Simple subtraction of these two values then gives the amount in dB which the individual word should be altered to achieve equal intelligibility. This information is found in tables 3.1 and 3.2.

The plots of the individual word intelligibility functions have shown some interesting and unexpected results. On examining some of the intelligibility functions it was noted that certain words, after following the expected pattern of correct scores becoming lower as the noise level is increased some words actually showed an improvement in correct word scores after a particular noise level had been reached. Further increases in the noise level then continued to improve the word scores.

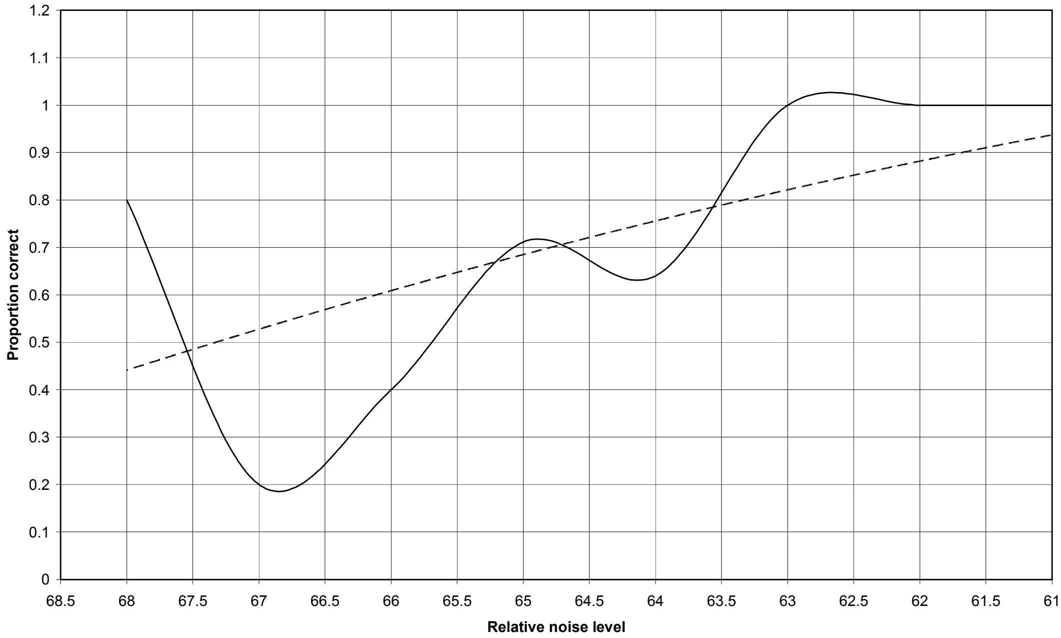

Figure 4.1 Proportion correct for word "has" as a function of noise level.

Figure 4.1 shows this phenomenon for the word 'has'. As the relative noise level is increased the correct scores fall to 18% correct at 66.8 dB then as the noise level is increased further the correct score rises dramatically to 80% correct at a relative noise level of 68 dB. This apparently indicates that some words become easier to identify after a certain noise level has been reached. This trend was noted not only in the word 'has' but also in plots for 'Alan', 'beds', 'Peter', 'red', 'thin' and 'two' although these have been much less dramatic. This seems not to have been noted in any other written work. However other researchers have privately acknowledged that this phenomenon exists. Given that different subjects contributed different data points, it is possible to question the reliability of this information. It might be possible that by chance having "good" subjects for some points could cause this.

The word 'Hannah' was the only word where the trend line failed to intersect with the 50% correct level and an estimate of the 50% correct score was made by extrapolation from lower scores (Figure 4.2). This indicates that the recording of 'Hannah' was too difficult to recognise relative to other words.

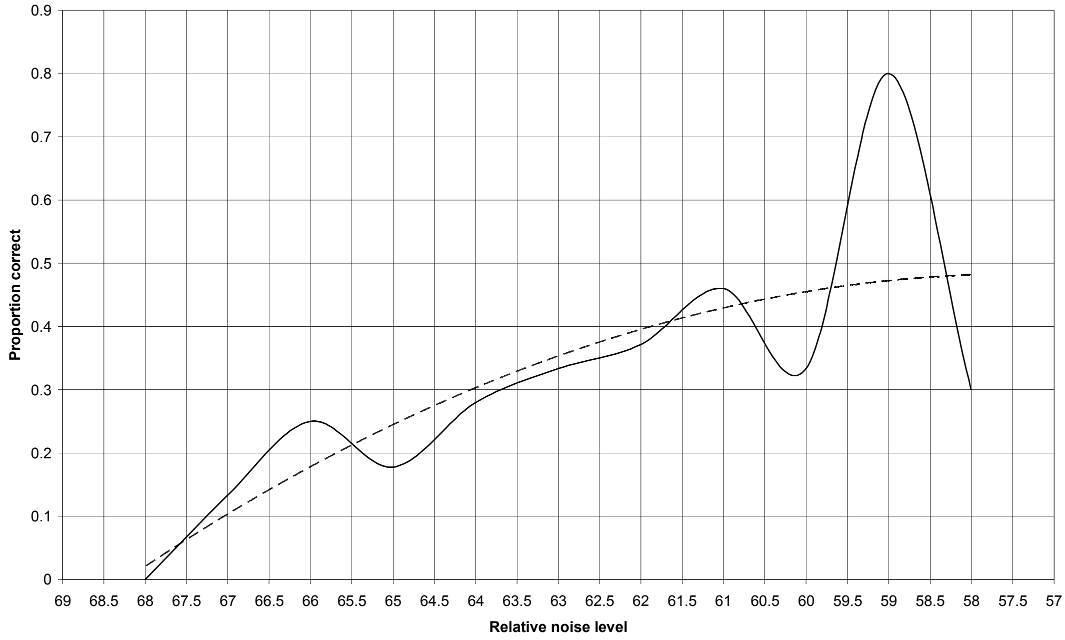

Figure 4.2 Proportion correct for word "Hannah" as a function of noise level.

The tables 3.1 and 3.2 show all the words of the synthetic sentence test and the number triplet test and the correction levels in dB. It has been suggested by Wagener et al (2003) that if a natural intonation is to be preserved then level adjustments to individual words should be restricted to ±4 dB. The words 'Hannah' (−7.2 dB adjustment required), 'Rachel' (4.2 dB adjustment required) and 'thin' (−6.9 dB adjustment required) from the synthetic sentence material and the words *'six'* (4.6 dB adjustment required) and *'three'* (−4.8 dB adjustment required) from the triplet number test would all fall outside this range. Further work would need to be undertaken to decide if these words should be adjusted by the required amounts or by a smaller amount or removed and replaced by other examples from the word material of the two tests. Alternatively the whole sentences which included those words could be removed and replaced with new sentences without those particular words. Of course this would alter the *a priori* probabilities of the words in the tests.

The problem of the adjective object files also needs to be addressed. The options include averaging the adjustment and applying this data to the combined file or splitting the file and applying the level adjustment to each word individually. As the levels have been calculated for each individual word and the some of these values are quite different the latter would seem the better option.

4.3 Comparisons with other studies

It has already been mentioned that other similar work has been undertaken. It is therefore possible to compare the signal-to-noise ratios at which words are correctly scored 50% of the time of the various tests.

The mean 50% correct noise level for the synthetic sentence test was 64.5 dB the speech signal was fixed at 55 dB which corresponds to a signal-to-noise ratio of −9.5 dB. The DANTALE II Danish sentence test (Wagener et al 2003) has a mean SRT of −8.43 dB SNR.

The mean 50% correct noise level for the number triplet test was 66.8 dB. As the same software was used for both tests the speech signal was also fixed at 55 dB so the corresponding signal-to-noise ratio is −11.8 dB. This can be compared to the Dutch speech-in-noise test developed for telephone use (Smits et al 2004) which measured a mean SRT for normal hearing subjects of −11.2 dB when presented via earphones.

So these mean SRT scores show good agreement especially with the number triplets. Differences may be explained by the differences between the three languages used in the different tests and the vagaries of individual speakers and recordings. However, the differences are small enough, especially for the triplets, to be due to uncertainties of estimation in the respective validation studies.

4.4 Further research

Having first resolved the problems already discussed with the correction of the correct waveform files. It would also need to be decided if the words outside the ±4 dB ranges need to be replaced.

The second stage of the study would be to apply this adjustment then new material should be evaluated to determine if equal intelligibility has been achieved. The further evaluation should involve a larger number of normal subjects, preferably large enough to be statistically representative of the population and to give larger numbers of data points for each intelligibility function especially when subdivided into each recording of each word. Wagener et al (2003) used sixty normally hearing subjects for the DANTALE II sentences test and Smits et al (2004) evaluated the Dutch number triplets with eighty normally hearing subjects. On completion of the second stage evaluation normative data can then be generated.

As the completed test is to be used with groups other than those with normal hearing, such as the hearing impaired and possibly older children. It would therefore be wise to collect further standardised data for these groups. The normative data may be unreliable for use with these groups and conclusions drawn of limited diagnostic value. So further research could involve testing adults and children with varying degrees and configurations of hearing loss. Such studies could also show how the test scores vary with hearing threshold level, which will be important if they are to be used as screening tests where a cut-off value must be decided upon.

The number triplets are to form part of a test that will be self completed over the subject's home telephone. So having made the adjustments and second stage evaluation further work involving the use of telephone systems will be required. In particular the in was noted by Smits et al (2004) that SRT values obtained using two different telephone systems were −7.1 dB and −6.9 dB respectively whilst using headphones −11.2 was achieved. These lower scores may arise from bandwidth of the telephone system, telephone instruments, system noise, compression within the system or other distortions. This implies that new normative data for these conditions should be collected.

4.5 Conclusion

This research has created the materials needed for the synthetic sentence in noise test and the number triplet test. It has followed same principles as the Hagerman sentences (Swedish), the Oldenburg sentences (German), DANTALE II (Danish) and the Dutch speech-in noise screening test for completion by telephone.
Data regarding the intelligibility of the material has been collected from normally hearing adult subjects, using these test materials.

Scores have obtained at different signal-to-noise ratios for each word, and compared with the overall scores of the combined word material.
The mean SRT scores for the two tests show a good agreement with scores obtained from similar studies. However it was noted that the individual word scores were in fact an average of all the combinations of that particular word from the base material.

Correction values have been obtained for each word in dB so that when applied all words will equally intelligible. The correction levels now need to be applied to each of the words so that equally intelligibility can be achieved and then a further evaluation carried out to establish if this is in fact the case.

References:

American Academy of Otolaryngology Committee on Hearing and Equilibrium, and the American Council of Otolaryngology Committee on the Medical Aspects of Noise (1979). "Guide for the evaluation of hearing handicap," Journal of the American Medical Association; 241 (19): 2055-2059.

Bench, J., Kowal, A, and Bamford, J. (1979) The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. British Journal of Audiology; 13: 108-112.

Boothroyd, A. (1968) Developments in speech audiometry. Sound; 2: 3-10.

Brand, T. and Kollmeier, B. (2002) Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. Journal of the Acoustical Society of America; 111: 2801-2810.

Byrne, B., Dillon, H., Tran, K., et al (1994) An international comparison of long term average speech spectra. Journal of the Acoustical Society of America; 96: 2108-2120.

Carhart, R. (1951) Basic principles of speech audiometry. Acta Otolaryngologica 40: 62-71.

Dirks, D.D., Morgan, D., Dubno, J. (1982) A procedure for quantifying the effects of noise on speech recognition. Journal of Speech and Hearing Disorders; 47: 114-123.

Duquesnoy, A.J. (1983) The intelligibility of sentences in quiet and in noise in aged listeners. Journal of the Acoustical Society of America; 74: 1136-1144.

Festen, J.M. and Plomp, R. (1983) Relations between auditory functions in impaired hearing. Journal of the Acoustical Society of America; 73: 652-662.

Festen, J.M. and Plomp, R. (1990) Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing. Journal of the Acoustical Society of America; 88: 1725-1736.

Foster, J.R. and Haggard, M.P. (1987) 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.

Fry, D.B. (1961) Word and sentence tests for use in speech audiometry. Lancet, 2, 197-199.

Gibson, L.J. (1998) Comparison of equal-level and equal-intelligibility construction of materials for sentence recognition in noise. University of Southampton; MSc Dissertation.

Hagerman, B. (1982) Sentences for testing speech intelligibility in noise. Scandinavian Audiology; 11: 79-87.

Hagerman , B. (1993) Efficiency of speech audiometry and other tests. British Journal of Audiology; 27: 423-425.

Hagerman, B. (1997) Attempts to develop an efficient speech test in fully modulated noise. Scandinavian Audiology; 26: 93-98.

Hagerman, B. (2002) Speech recognition threshold in slightly and fully modulated noise for hearing-impaired subjects. International Journal of Audiology; 41: 321-329.

Hagerman, B., and Kinnefors, C. (1995) Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. Scandinavian Audiology; 24: 71-77.

Howes, D. (1957) On the relation between the intelligibility and frequency of occurrence of English words. Journal of the Acoustical Society of America; 29: 296-305.

Kalikow, D. N., Stevens, K.N. and Elliot, L.L. (1977) Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. Journal of the Acoustical Society of America; 61: 1337-1351.

King, P.F., Coles, R.R.A., Lutman, M.E. and Robinson, D.W. (1992) Assessment of Hearing Disability. London: Whurr Publishers Ltd.

Kollmeier, B. and Wesselkamp, M. (1997) Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. Journal of the Acoustical Society of America; 102: 2412-2421.

Kramer, S.E., Kapteyn, T.S., and Festen, J.M. (1998) The self-reported handicapping effect of hearing disabilities. Audiology; 37: 302-312.

Lehmann, R. (1962) Etude psychophysique de l'intelligibilite du langage. Theses de l'Universite de Paris, Editions de la Revue d'Optique Theorique et Instrumentale.

Levitt, H. (1971) Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America; 49, 467-477.

Lutman, M.E. (1991) Hearing disability in the elderly. Acta Otolaryngologica; Suppl. 476: 239-248.

Lutman, M.E. (1997) Speech tests in quiet and noise as a measure of auditory processing. In Martin, M (1997) Speech Audiometry. (2nd Ed) London: Whurr Publishers Ltd, 63-73.

Lutman, M.E., Brown, E.J., and Coles, R.R.A. (1987) Self reported disability and handicap in the population in relation to pure-tone threshold, age, sex and type of hearing loss. British Journal of Audiology; 21: 45-58.

Lutman, M.E., Clarke, J. (1986) Speech identification under simulated hearing aid frequency response characteristics in relation to sensitive frequency resolution and temporal resolution. Journal of the Acoustical Society of America: 80: 1030-1040.

Lyregaard, P.E., Robinson, D.W. and Hinchcliffe, R. (1976) feasibility study of Diagnostic speech audiometry. Teddington: National Physical Laboratory, Acoustic Report AC73.

Lyzenga, J. (2005) Literature overview on sentence test for assessing speech intelligibility in noise and quiet. Personal communication.

MacLeod, A. and Summerfield, A.Q. (1990) A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise. British Journal Audiology; 24: 29-43.

Markides, A. (1978a) Whole word scoring versus phoneme scoring in speech audiometry. British Journal of Audiology; 12: 40-46.

Markides, A. (1978b) Speech discrimination functions for normal hearing subjects with AB isophonemic word lists. Scandinavian Audiology; 7: 239-245.

Martin, M (1997) Speech Audiometry. (2nd Ed) London: Whurr Publishers Ltd.

Munro, K. J. Lutman, M.E. (2003) The effect of speech presentation level on measurement of auditory acclimatization to amplified speech. Journal of the Acoustical Society of America; 114: 484-495.

Nilsson, M., Soli, S.D. and Sullivan, J.A. (1994) Development of a hearing in noise test for the measurement of speech reception thresholds in quiet and noise. Journal of the Acoustical Society of America; 95: 1085-1099.

Owens, E. (1961) Indelibility of words varying in familiarity. Journal of Speech and Hearing Research; 4: 113-129.

Parnell, M.M. and Amerman, J.D. (1978) Maturational influences on perception of coarticulatory effects. Journal of Speech and Hearing Research; 21: 682-701.

Plomp, R. and Mimpen, A.M. (1979a) Speech reception threshold for sentences as a function of age and noise level. Journal of the Acoustical Society of America; 66: 1333-1342.

Plomp, R. and Mimpen, A.M. (1979b) Improving the reliability of testing the speech reception threshold for sentences. Audiology; 18: 43-52.

Sahakian, A. (1998) A normative study of an automated adaptive BKB sentence-in-noise test. University of Southampton. MSc Dissertation.

Savin, H.B. (1963) Word frequency effects and errors in the perception of speech. Journal of the Acoustical Society of America; 35: 200-206.

Smits, C, Kapteyn, T.S. and Houtgast, T (2004) Development and validation of an automatic speech-in-noise screening test by telephone. International Journal of Audiology; 43: 15-28.

Smits, C. and Houtgast, T. (2005) Results from the Dutch speech-in-noise screening test by telephone. Ear and Hearing; 26: 89-95.

Smoorenburg, G.F. (1992) Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. Journal of the Acoustical Society of America; 91: 421-437.

Thornton, A.R. and Raffin, M.J. (1978) Speech-discrimination scores modelled as a binomial variable. Journal of Speech and Hearing Research; 21: 507-18.

Tschopp, K., and Ingold, L. (1994) Entwicklung einer deutschen version des SPIN-Tests (Speech Perception in Noise). In: Moderne Verfahren der Sprachaudiometrie, edited by B. Kollmeier. Heidelberg, Germany: Median Verlag; 311-329.

Versfeld, N.J., Daalder,L., Festen, J.M. and Houtgast, T. (2000) Method for the selection of sentence material for efficient measurement of the speech reception threshold. Journal of the Acoustical Society of America; 107: 1671-1684.

Wagener, K. (2005) Description of recording, cutting and validating Oldenburg sentence test types. Personal communication.

Wagener, K., Brand, T., and Kollmeier, B. (1999a) Entwicklung und evaluation eines satztests fur die deutsche sprache I. Design des Oldenburger Satztests. Z. Audiology; 38: 4-15.

Wagener, K., Brand, T., and Kollmeier, B. (1999b) Entwicklung und evaluation eines satztests fur die deutsche sprache II. Optimierung des Oldenburger Satztests. Z. Audiology; 38: 44-56.

Wagener, K., Brand, T., and Kollmeier, B. (1999c) Entwicklung und evaluation eines satztests fur die deutsche sprache III. Evaluation des Oldenberger Satztests. Z. Audiology; 38: 86-95.

Wagener, K., Josvassen, J.L. and Ardenkjaer, R. (2003) Design optimization and evaluation of a Danish sentence test in noise. International Journal of Audiology; 42: 10-17.

Wright, R. (1997) Basic properties of speech. In Martin, M (1997) Speech Audiometry. (2nd Ed) London: Whurr Publishers Ltd, 1-33.

www. Phon.ucl.ac.uk/home/andyf/natasha.htm

www.Hearcom.com

Appendix 1

Screening questionnaire

Screening questionnaire to be completed by all volunteer subjects.

Do you consider yourself to have English as a first language? Yes / No
(please circle)

If no please state your first language.

Do you have any history of ear disease? Yes / No
(please circle)

If yes please give details.

Have you ever had any operations on your ears? Yes / No
(please circle)

If yes please give details.

Do you have a history of exposure to loud sounds? Yes / No
(please circle)

If yes please give details.

Have you been exposed to loud sounds in the past 48 hours? Yes / No
(please circle)

If yes please give details.

Do you suffer from noises in the head or ears which last longer than 5 minutes? Yes / No
(please circle)

If yes please give details.

Are you suffering with any colds or congestion today? Yes / No
(please circle)

If yes please give details.

Do you have any medical or other reasons which you feel may prevent you from taking part in this experiment? Yes / No
(please circle)

If yes please give details.