1 Introduction

The ability to understand speech in noise has a profound impact on communication in everyday life. Most conversations take place with a background of music, other voices, construction or industrial noise, and/or traffic. Although normal hearing (NH) listeners report that they experience some difficulty in noise, the high redundancy of conversational speech helps NH listeners to understand even a degraded signal (Beattie, Barr, & Roup, 1997; Dirks, Morgan, & Dubno, 1982).

The relationship between puretone thresholds, speech recognition in quiet and speech recognition in noise is highly predictable for those with normal hearing (Dirks et al., 1982). In contrast, it is extremely difficult, if not impossible, to predict to what extent a hearing impaired (HI) listener will have difficulty with speech discrimination in noise solely on the basis of the puretone audiogram and/or performance on speech in quiet (Killion & Niquette, 2000). Listeners that score identically on word recognition tasks in quiet can display significantly different word recognition abilities in noise (Beattie et al., 1997).

A common complaint of HI listeners is that they can *hear* the speaker but are unable to *understand* them. On this basis, Carhart (1951) distinguished two components of hearing loss and termed them 'acuity' and 'clarity'. Acuity was related to hearing sensitivity, which is the ability of an individual to detect sound. Clarity referred to the ability of an individual to recognize complex sound stimuli such as speech. Stephens (1976) and Plomp (1978) built on this framework to characterise two contributing factors to hearing impairment as attenuation (acuity) and distortion (clarity). Attenuation refers to the reduction in ability to hear and process simple sounds in threshold-seeking tasks such as puretone audiometry, whereas the distortion component refers to the increased difficulties experienced with more complex stimuli presented at supra-threshold levels such as speech understanding in quiet and in noise. How these two factors combine to affect speech understanding in noise is still not totally understood.

It has been established that individuals with sensorineural hearing loss (SNHL) demonstrate greater difficulty understanding speech in background noise than do NH individuals under the same conditions (Beattie, 1989; Carhart & Tillman, 1970; Dirks et al., 1982). Speech discrimination ability in noise also tends to deteriorate with age (Barrenas & Wikstrom, 2000; Dubno, Dirks, & Morgan, 1984). This decline has been shown to be attributable to deficits in auditory processing related not only to

audibility, but to supra-threshold deficits such as those related to frequency resolution and temporal processing (Gordon-Salant & Fitzgibbons, 1993).

1.1 Factors contributing to supra-threshold deficits

While audibility, represented by hearing thresholds, has been shown to have the greatest predictive power for performance on speech in noise tests, it cannot account for all variance observed in testing (Houtgast & Festen, 2008). In some cases there may be underlying damage that is not revealed by raised thresholds (Portmann, Harrison, Negrevergne, Dauman, & Aran, 1983). Various studies have shown abnormal otoacoustic emissions in ears that have been exposed to noise, in the presence of what would be classed as normal puretone thresholds (Attias et al., 1995; Lucertini, Moleti, & Sisto, 2002) . In experiments exposing young NH listeners to noise, Feth, Osterle & Kidd (1979) found that frequency selectivity was compromised even at thresholds where there was no temporary threshold shift. As both these processes are assumed to be generated by the outer hair cells, it is possible that there is sufficient activity for a puretone to be transmitted, but damage may already be occurring at a subclinical level. This suggests that puretone thresholds are not always a reliable indicator of performance on tests using more complex stimuli.

In contrast to detection of puretones, speech perception requires the integration of a number of systems. Speech perception is affected not only by the peripheral system ("bottom-up") but also by central processing ability and non-auditory factors ("top-down"), such as working memory capacity and speed of temporal processing (Humes, 2005). This combination of bottom-up or "stimulus-driven" processes and top-down or "knowledge-driven" factors determines speech reception ability (Goldstein, 2002).

Both bottom-up and top-down processes make use of both the intrinsic and extrinsic redundancy of speech. As discussed by Bocca and Calearo (1963), intrinsic redundancy comes from the listener's own knowledge of the language, whereas extrinsic redundancy refers to clues contained within the signal relating to phonology and syntax. For example, the articulation of consonants affects some acoustic properties of the vowels around them (called co-articulation effects), such as the transitions between formants at the beginning and end of vowels (Ohman, 1966). Even if parts of the signal are masked by surrounding noise, listeners use both the acoustic clues in the signal (bottom-up processing) and their knowledge of vocabulary (top-down processing) in order to fill in the gaps left by the masked-out phonemes (Warren, 1970).

1.1.1 Bottom-up factors

a) Frequency resolution

Frequency resolution or frequency selectivity refers to the ability to separate out the different frequency components in a complex sound. In a human ear, the basilar membrane (BM) assists in this separation of frequency components, acting as a kind of frequency analyser (Moore, 1998). Sound energy is converted into a travelling wave running from base to apex along the BM, each point of which is 'tuned' to a particular frequency (its characteristic frequency, or CF). This is partly due to the physical properties of the BM, which is narrow and stiff at the base, where responses are observed to higher frequencies, and wider and more relaxed at the apex, which responds to lower frequencies. This is termed the passive process (Robles & Ruggero, 2001). When a particular frequency is presented to the auditory system, a travelling wave is generated along the BM with maximum displacement caused at the place coinciding with the corresponding CF. Acoustic stimuli of different frequencies will produce maximum displacement at different points along the BM.

The 'sharpness' of the tuning is increased by nonlinearities on the BM known as the active process, also called the cochlear amplifier, first discussed by Gold (1948). The outer hair cells (OHCs) are responsible for giving a mechanical 'boost' to the displacement on the BM caused by low level sounds and enhancing the listener's ability to differentiate acoustic signals that are very close in frequency. Without fully functioning OHCs, the sharp tuning provided by the active process is lost, and frequency selectivity is dramatically reduced.

Differences in frequency selectivity are regularly observed in people with SNHL caused by various pathologies such as noise exposure, ototoxicity or presbycusis (Moore, 1982). The common variable in these cases is usually a loss of the OHCs. Florentine, Buus, Scharf and Zwicker (1980) found that people with cochlear damage had reduced selectivity on several different measures of frequency selectivity. Although reduced frequency selectivity appears to be correlated with level of cochlear damage, other investigators have found varying results. Margolis and Goldberg (1980) found that some listeners with SNHL showed almost normal frequency selectivity. Moore (1982) has suggested that this may be due to the pattern of cochlear damage and whether inner hair cells (IHCs) or OHCs are more affected. There may also be some differences in the way frequency selectivity is affected in those with cochlear damage and those with damage to the auditory nerve (that will present in other ways like cochlear hearing loss). Another factor may be the different methodologies used in order to measure frequency selectivity.

b) Temporal processing

Temporal resolution refers to the ability to discriminate rapid changes in sound patterns. The speech signal is rich in temporal cues. Some consist of short-term fluctuations, which encode information about segmental speech properties such as voicing (the difference between 'b' and 'p') and manner of articulation (the difference between 'b' and 'm'). More temporal information is conveyed in the long-term properties of the temporal envelope, which encodes prosodic cues (Reed, Braida, & Zurek, 2009). Temporal processing ability may also affect speech recognition in noise, as it can allow listeners to benefit from brief improvements in the signal-to-noise ratio as they hear more in the gaps. This ability plays a crucial role in the understanding of speech both in quiet and in noise, as without it we would be unable to detect or discriminate the temporal cues that assist listeners in making sense out of the rapidly changing speech signal.

HI individuals have been shown to have temporal processing deficits (Festen & Plomp, 1990). Whether detecting brief pauses in an otherwise continuous sound (Buus & Florentine, 1985), or discriminating soft sounds played rapidly after the continuous sound stops (Nelson & Freyman, 1987), listeners with SNHL do poorly compared to NH listeners. HI individuals typically show smaller improvements in speech reception thresholds (SRTs) in fluctuating noise compared to continuous noise. This may be because their reduced temporal acuity means they cannot take advantage of the improved signal to noise ratio (SNR) during the gaps in the same way that that NH individuals do (Smits & Houtgast, 2007)

c) Auditory streaming

Associated with both frequency selectivity and temporal resolution is the ability to identify certain acoustic patterns as coming from a common source. This is called object formation, also known as auditory streaming (Bregman, 1990; Darwin & Carlyon, 1995). To be able to selectively attend to an acoustic "object", the listener must be aware of the spectral and temporal features of the object, such as perceived location, pitch and timbre (Shinn-Cunningham & Best, 2008). Object formation and object selection is what allows NH listeners to separate out a certain signal, such as a single voice, from background noise (Cameron & Dillon, 2007). As mentioned above, sensorineural hearing loss is often associated with abnormalities in the perception of spectrotemporal features of auditory stimuli. These peripheral degradations may contribute to the difficulties experienced by hearing-impaired people by interfering in and slowing down the processes of object formation and object selection. While these deficits may only have limited impact in quiet, a degraded representation of the

auditory scene will result in a target signal or object that is perceptually similar to the competing signals, and thus much more difficult to attend to using top-down processes. The increased time taken to build up object formation and selection will have an impact on the HI listener's abilities to follow group discussions in noisy environments where the speaker changes constantly.

Exacerbating these issues can be difficulties with localisation of sound within the environment. Localisation of sound depends on processing very subtle acoustic cues, such as the difference in time between the signal reaching each ear (inter-aural time differences) and the difference in volume between the ears (inter-aural level differences). High frequency information is often very important for localising sound, e.g. inter-aural level differences are pertinent mainly for frequencies over 3000 Hz (Middlebrooks & Green, 1991). If these cues are not audible because of raised thresholds, or distorted because of poor frequency selectivity and/or temporal resolution then localisation may be difficult. This may further hinder auditory streaming.

1.1.2 Top-down processes

Cognitive processes have long been recognised as contributing to speech perception in noise. Although it can be challenging to separate out peripheral, bottom-up processes from central, or cognitive, top-down processes, results of a number of studies have suggested that some of the variation in hearing performance in noise can be attributed to cognitive abilities (Sommers, 1997; Wingfield, 1996). This is particularly pertinent in studies looking at speech recognition in older people. Speed of processing is one cognitive process that is likely to decrease with age (Wingfield, Poon, Lombardi, & Lowe, 1985). General cognitive slowing can be observed as part of the aging process, possibly as a result of cellular loss occurring throughout the brain, and this can affect the rapid information processing required for understanding speech (Cerella, 1985). Age-related memory constraints may also have an effect on whether context in sentences can be utilized (Wingfield, 1996). Memory constraints have been shown to have a significant impact on performance on speech perception tasks (Salthouse, 1991).

Other studies have been less conclusive. Results from one study by Humes (1996) showed that the degree of SNHL was the sole or at least the primary predictor for performance of elderly people on speech-in-noise tests. Pichora-Fuller, Schneider and Daneman (1995) found that neither hearing sensitivity nor age-related changes in cognition could fully account for the difficulty that elderly listeners experience in

noise. They speculated that hearing impairment not only makes speech in noise harder to perceive, but that the extra effort and concentration required to listen in noise means that there are fewer resources available to the cognitive processes necessary for speech understanding.

Studies involving tests of other modalities have been used to tease out the various effects of peripheral and cognitive processes. George, Zekveld, Kramer, Goverts, Festen, and Houtgast (2007) measured the speech reception abilities in noise for both NH and HI individuals. In addition, they assessed non-auditory factors by measuring the ability of each individual to accurately identify written text sentences that were partially obscured by black lines of varying thickness. This visual analogue of the listening task enabled them to conclude that, for NH listeners, non-auditory (i.e. cognitive and processing) factors were the most important source of variance in speech reception ability. For HI listeners, in situations such as in fluctuating noise both auditory and nonauditory factors accounted for the variance in speech reception ability.

1.2 Speech audiometry in audiological testing

Speech audiometry is an integral part of the audiological test battery. It is a key measure of overall auditory perception skills, providing an indication of the individual's ability to identify and discriminate phonetic segments, words, sentences and connected discourse (Mendel, 2008). Scores on speech tests are often used as a crosscheck of the validity of puretone thresholds (Brandy, 2002). Performance on these tests has consequences for the diagnosis and management of hearing impairment and recommendations for intervention, such as amplification, personal FM systems, and cochlear implants. It is also important in audiological (re)habilitation in areas such as speech reading, auditory training, and perceptual training, and for monitoring progress in these areas. Speech perception skills should be routinely assessed using valid and reliable clinical assessment methods which are suitable for the target population, whether adults or children. Materials used for speech perception testing should meet several criteria: (a) familiarity, (b) phonetic dissimilarity, (c) representative sample of English speech sounds, and (d) homogeneity with respect to audibility (Hudgins, Hawkins, Karlin, & Stevens, 1947).

Speech audiometry is performed regularly as part of an initial audiological assessment. One important aim of speech audiometry is to find the patient's speech reception threshold (SRT), defined as the point at which the patient can correctly identify 50% of the speech stimuli (Plomp, 1978). It gives an indication of the

patient's sensitivity to speech sounds and provides a measure of the individual's speech processing abilities along the entire auditory pathway. This figure can then serve as a crosscheck for the validity of the puretone audiometry results, by comparing it to the puretone average (PTA). This is typically calculated from the puretone thresholds at 500, 1000 and 2000 Hz. Agreement between the SRT and PTA is considered good if within ±6 dB, fair if between ±7 dB and ±12 dB, and poor if greater than ±13 dB (Brandy, 2002). Inconsistencies between the SRT and PTA may result from the influence of extrinsic factors, such as equipment malfunction, errors of calibration, or misunderstanding of the instructions by the patient. However, if the SRT is assessed to be significantly lower than the PTA, and this is not explained by the slope of the audiogram, pseudohypacusis may be indicated (Martin, 2009). If the SRT is significantly higher than the PTA then the possibility of a retrocochlear lesion, some other central auditory disorder, or language impairment should be investigated. In clinical practice, the SRT is used to roughly gauge the level of supra-threshold distortion present in the auditory system. If the maximum score reached is low, and presenting at higher levels results in no increase (or even a decrease) in understanding, this may suggest that 'turning up the volume' with amplification will not be as successful as predicted from the puretone thresholds.

1.3 Monosyllabic words in quiet

The materials used in speech audiometry clinically are generally monosyllabic word lists presented in quiet, such as the Meaningful CVC (Revised AB) Words (Boothroyd & Nittrouer, 1988), North-Western University Auditory Test number 6 (NU-6; Tillman & Carhart, 1966), and the Central Institute for the Deaf Auditory Test W-22 (CID W-22) (CID W-22; Hirsch et al., 1952). Items are presented in lists, often after a carrier phrase, such as "say (the word) _____". Performance is scored by word or by phoneme repeated correctly, to arrive at a percentage correct score for the level at which the list is presented. Speech audiometry in current audiological practice consists of presenting a single list of monosyllabic words to each ear, at a level where the listener would be expected to get close to 100%. This is generally estimated using formulas based on the puretone average. For greater validity, a number of word lists are presented at two or more different intensity levels in order to describe a performance-intensity (PI) function, from which the SRT, or 50% correct point, can be estimated. This is standard clinical practice in New Zealand, but due to time pressures and other factors, it is not always carried out to its fullest extent.

The conditions under which speech audiometry is performed in the clinic are optimal compared to those habitually encountered in the real world. Speech materials are presented through headphones, in a soundproof room, with maximum concentration from the patient and minimum external distraction. Monosyllabic words are presented in isolation, without context, so that patients must repeat what they hear without relying on contextual clues. The aim of testing patients with monosyllabic words in quiet is to attempt to isolate the problem of audibility from other confounding factors such as working memory and use of context (Wilson, McArdle, & Smith, 2007). Although testing patients with these word lists is efficient and targeted, single word recognition tests are not representative of spoken language and the validity of these lists of words for predicting the social adequacy of one's hearing has been extensively questioned (Beattie, 1989; Orchik, Krygier, & Cutts, 1979). The most important factor is that speech-recognition testing in quiet does not address the main problem experienced by the majority of HI patients, which is difficulty understanding speech in noise.

The assessment of receptive communication abilities ideally should involve speech materials and listening conditions that are likely to be encountered in the real world. There is less redundant information in single monosyllabic words than there is in sentences, which yield multiple contextual clues involving syntax and semantics. Sentences are far more representative of everyday communication than isolated monosyllabic words since they include natural intensity fluctuations, intonation, contextual cues, and temporal elements that are associated with conversational speech (Nilsson, Soli, & Sullivan, 1994). Conversational speech is highly redundant, as knowledge of the subject in question, and visual cues from lip-reading and body language can assist the listener in deciphering the signal. Speech materials consisting of sentences and phrases yield a much more valid measurement of speech reception abilities as practised in the real world.

Diagnostic tests must be sufficiently sensitive to discriminate between listeners with NH and patients with various hearing impairments. Speech testing in quiet does not appear to exhibit this characteristic, as listeners with identical word recognition abilities in quiet can have significantly different word recognition abilities in background noise (Beattie et al., 1997). It also has little predictive value, as it is impossible to predict the potential improvement in noise when amplification is provided from scores on monosyllabic words in quiet. Using monosyllabic words in quiet as the sole measure of speech recognition is insufficient for making recommendations for amplification and instilling in the patient realistic expectations

as to the benefit they may receive from hearing aids. An efficient, reliable test of speech perception in noise would be more valuable than testing in quiet if the aim was to document a person's reported degree of handicap before and after providing amplification.

1.4 Speech-in-noise tests

Speech-in-noise tests have long been recognised as an important addition to the audiological test battery, although they are only just starting to be introduced clinically (Beattie, 1989; Carhart & Tillman, 1970; Dirks et al., 1982; Killion & Niquette, 2000). Speech-in-noise testing enables the clinician to test HI patients in the kind of realistic, 'real-world' situation in which they report having the greatest difficulty. Results from testing in noise may help determine which ear is suitable for amplification (Beattie et al., 1997). It can have benefits for hearing aid selection and counselling, giving a more realistic assessment of the likely benefit the patient will receive from hearing aids, and whether a certain type of hearing aid or feature will ensure maximum benefit for the patient. This may be useful in order to reconcile a patient's desire for a certain type of hearing aid with their goals for listening in certain situations. If a patient scores poorly on a speech-in-noise test, but hopes that their new hearing aid will help them to hear and take part in the conversation while dining with friends in a restaurant, then careful counselling will need to be given so that the patient's expectations of their performance in noise is not unrealistically high.

The function and benefit of different hearing aid technology can be evaluated using speech-in-noise tests in the soundfield. Valente, Fabry and Potts (1995) found that using a directional microphone significantly increased scores on the Hearing in Noise Test (HINT). Lin, Bowditch, Anderson, May, Cox and Niparko (2006) used the HINT to quantify the increase in benefit from a Bone Anchored Hearing Aid (BAHA) compared to Contralateral Routing of Signals (CROS) amplification. Ricketts and Dhar (1999) used modified HINT stimuli to compare the performance of three different hearing aids. The increased difficulty of these tests helped to minimize ceiling effects, thereby separating out the performance of different hearing aids.

To achieve results in noise that are comparable to NH listeners, HI listeners require a better signal-to-noise ratio, or SNR (Beattie et al., 1997). That is, HI listeners need the signal to be louder in relation to the noise in order to discriminate speech as well as NH listeners. Irrespective of the location of the lesion or deficit(s) that cause poor speech in noise discrimination, the resulting disability can be understood by using

the concept of “SNR loss (signal-to-noise ratio loss)”. This term refers to the “increase in signal to noise ratio required by a listener to obtain 50% correct words, sentences, or words in sentences, compared to normal performance” (Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004, p. 2395). One person, with a PTA of 50 dB, may have an SNR loss of 2 dB once speech is made audible by hearing aids. This indicates that they will have no more problems in noise than a NH person would. Another person with an identical PTA may have such a severe SNR loss that they require at least directional microphones and ideally an FM system in order to understand speech in noise at all. Determining the SRT in noise gives a quantitative measure of the degree of this disability. This information can be used to give an indication of whether amplification will be successful, or whether other technologies such as FM systems will be required to improve speech understanding in noise.

Speech-in-noise testing must be carried out to get a true picture of the listener’s communication abilities. As there is no correlation between self-report measures of difficulty understanding speech in noise and objective measurements of this ability (Rowland, Dirks, Dubno, & Bell, 1985), efficient, reliable objective tests should be part of the audiological test battery.

1.5 Types of noise/masker

Different kinds of maskers are commonly used to provide the ‘noise’ component of speech-in-noise tests. A broadband noise masker provides equal energy at all frequencies. Speech is a spectrally complex signal, with one peak around 500 Hz, above which the spectrum slopes downward. If using a spectrally flat masker (e.g. white noise), the overall level must be very high to mask out the 500-Hz peak region of speech, and thus louder than other types of noise in order to have an equivalent masking effect on speech. However, stationary broad band noise can be synthesised to match the long-term average spectrum of the target speech. The advantage of a speech-shaped masker is that it ensures that on average the S/N ratio will be approximately equal at all frequencies, as well as being not as loud as a white noise masker. A spectrally matched masker can also increase the sensitivity of the test to changes in speech discrimination (Nilson, 1994).

Non-stationary or fluctuating noise has also been used in speech discrimination studies. This type of noise is also matched to the long term average spectrum of the signal, and can be varied by the frequency, depth, and/or regularity of modulation (e.g. interrupted noise with a duty cycle of 50%, amplitude modulated noise, speech modulated noise). These noises are often used in clinical or research settings but are

not representative of everyday masking sounds (Rhebergen, Versfeld, & Dreschler, 2008).

Another type of non-stationary masker is multitalker babble, which consists of four or more individual voices speaking at once, so that no one voice is intelligible. This has been the masker of choice for many recently developed speech-in-noise tests (see 1.9 Available speech tests below). It gives good face validity, as it mimics the 'cocktail party effect' (first discussed by Cherry, 1953) where masking is provided by multiple co-occurring speech signals coming from other people in the room. It may be even more sensitive to hearing impairment, as the natural fluctuations in amplitude may separate out HI listeners from NH listeners more easily (Smits & Houtgast, 2007). As there will be amplitude modulated fluctuations in the masker, it is important when establishing equivalent intelligibility of speech materials to fix the same passage of babble to each portion of speech, so that the peaks and troughs in both the signal and the noise always fall in the same place (Killion et al., 2004).

1.6 Adaptive speech tests

Speech tests in quiet of the kind that are currently used in audiological practice in New Zealand fall into the category of non-adaptive tests. This non-adaptive method, where the distribution of trials is pre-determined at different fixed intensities, is called the method of constant stimuli. Percentage correct scores are used to form a performance intensity function, like that used with the AB word lists.

Another method is the descending presentation level paradigm, which is a pseudo-adaptive procedure involving the presentation of a set of target stimuli at a fixed SNR followed by further sets of target stimuli at decreasing levels. The number of target stimuli and decibel step sizes can be varied, but all are administered in a systematic fashion. The Spearman-Karber equation (Finney, 1952) is used to calculate the SRT of 50%.

In contrast, a true adaptive procedure is one in which the stimulus level on any one trial is determined by the response to the preceding stimulus (Levitt, 1971). This differs from the method of constants where sets of stimuli are presented at fixed intensity levels. Instead, the participant's previous responses help determine the presentation level of the next stimulus item until their threshold can be determined and testing is suspended. Threshold is defined as the stimulus intensity at which the listener can identify the stimulus correctly for a certain percentage of trials. By honing in more quickly and efficiently to the "region of interest" in which each individual patient's threshold is likely to fall, they can be more effective than tests

that use the method of constants (Levitt, 1978, p. 287). This has important ramifications for clinicians and time management while making the task less onerous for the patient.

There are a number of experimental variables that may influence the results obtained with a given adaptive procedure. These include the step size, defined as the difference in level between sequential stimuli; the initial starting level or intensity of the stimulus; the tracking algorithm, which controls how the presentation level of the stimuli is varied; and the stopping rule, which determines when to finish testing for each round (Leek, 2001). Despite differences between procedures, the aim is to accurately measure characteristics of the subject's performance over the shortest amount of time and thus maximise the potential benefits of an adaptive approach.

The two main categories of adaptive procedures are maximum-likelihood and staircase procedures. Maximum-likelihood procedures evolved from modifications to the parameter estimation by sequential testing (PEST) methodology which resulted in improved reliability and efficiency in determining the threshold for a psychometric function (Pentland, 1980). After each trial, a new psychometric function is generated based on the proportion of correct responses associated with each stimulus level. The level of the next stimulus presentation is decided upon using the current best estimate of the underlying psychometric function. As the number of trials increases, more data points are generated and the psychometric function becomes more accurate and better defined. Various studies have shown that reliable, accurate threshold estimates can be obtained in 20-25 trials (Leek, 2001).

Staircase procedures use a simpler, more flexible adaptive methodology which have been used in many studies (e.g. Dirks et al., 1982; Kamm, Morgan, & Dirks, 1983). In these procedures, an incorrect response causes an increase in the presentation level of the following sentence, while a correct response causes a decrease in presentation level. Each 'step' can either be of the same or different magnitude, depending on the strategy used. This tracking method targets the 50% performance level on the psychometric function, which stands halfway between 0% correct performance (at chance) and 100% correct performance. Eventually, successive presentation levels will oscillate above and below the 50% correct point. The SRT is then estimated by averaging the presentation levels recorded at the points at which the responses change direction (reversals). The number of reversals required to estimate a threshold can vary between four and forty (Leek, 2001). Averaging more reversals can result in a more accurate and reliable threshold, but this comes at the expense of efficiency (Kaernbach, 2001).

While simple staircase procedures target the 50% performance level, transformed staircase procedures can be used to determine other points along the psychometric function (Levitt, 1971). This can give useful information on the slope of the function. Whereas the simple staircase procedure requires that the level of the stimulus changes after every response, transformed procedures demand a sequence of correct responses before decreasing stimulus intensity. However, a single incorrect response will cause the stimulus intensity to increase. For example, a 'three-down, one-up' procedure will target the 79.4% step level. Weighted step procedures, where different fixed steps are used for increases and decreases in intensity, have been shown to be even more flexible and efficient than simple or transformed staircase procedures (Kaernbach, 2001).

Adaptive measures have a long history in the field of audiology, especially in speech testing. Kamm, Morgan and Dirks (1983) used an adaptive speech test protocol based on Levitt (1978) to estimate maximum speech recognition (PB-max) ability. This study found that an adaptive staircase procedure gave a more accurate estimation of PB-max than presentation of a single list at a fixed intensity corresponding to a specific sensation level (e.g. 30-40 SL). Dirks et al. (1982) used an transformed up-down adaptive procedure (Levitt & Rabiner, 1967) to determine the effects of competing babble on the speech discrimination of HI adults by obtaining SRTs at different target levels on the performance-intensity curve. In this study the signal to babble ratio was adaptively varied by manipulating the noise in 2 dB steps while keeping the speech stimuli at a fixed intensity.

1.7 Advantages of adaptive tests

Adaptive tests have been shown to be efficient, accurate and reliable (Buss, Hall, Grose, & Dev, 2001; Leek, 2001). Administration time can be considerably shortened, with no loss of accuracy or reliability. The use of a maximum likelihood adaptive procedure in a test of visual field sensitivity enabled thresholds to be obtained in half the time that it took using a method of constant stimuli (Turpin, McKendrick, Johnson, & Vingrys, 2002). Zera (2004) implemented an adaptive maximum-likelihood procedure in the context of a modified rhyme test to measure speech intelligibility. In this study, it was found that the adaptive method was an accurate and highly efficient procedure for estimating the SNR required to obtain different percentage correct scores against a constant level of background noise.

The potential advantages of adaptive speech intelligibility testing have been investigated in studies by Levitt and Rabiner (1967), (1971), and Bode and Carhart

(1974). Subsequent studies have found that speech reception performance in quiet and at fixed signal-to-noise conditions are less likely to reveal differences than the use of speech stimuli in noise presented using an adaptive procedure (Frisina & Frisina, 1997). For the speech-in-noise SRTs, Dubno et al. (1984) presented spondee, low predictability and high predictability materials at fixed intensity levels in a background of babble (see section 1.9.1 SPIN test below). The first presentations were at a highly favourable signal-to-babble ratio, after which the noise level was varied adaptively in 10 dB steps, following a simple up-down procedure. Once the starting level was determined the step size was reduced to 2 dB. Dubno et al. (1984) found a constant age effect on scores obtained on the SPIN materials by individuals with NH and others with a mild hearing loss. According to Gordon-Salant (1987), this was due to the adaptive test design rather than the sensitivity of the SPIN items. This suggests that the benefits of using the adaptive method can overcome shortcomings in certain test materials.

Adaptive test procedures display further advantages over methods of constant stimuli. One benefit of measuring the SRT directly, rather than eliciting percentage-correct scores, is that it avoids floor and ceiling effects (where a number of participants obtain scores of, or close to, 0% or 100%). These effects can distort results and make it difficult to reveal significant differences in speech recognition ability (Gifford, Shallop, & Peterson, 2008). Once scores close to 100% are attained then no further improvement can be recognised, as the testing materials are not of sufficient difficulty to challenge the patient's abilities. Certain populations require tracking of word and sentence recognition performance over time after a clinical intervention, such as providing amplification or cochlear implantation. Many cochlear implant wearers reach close to maximum performance (100%) within three months of implantation when tested on sentences in quiet at a fixed level (e.g. 65 dBSPL) (Litovsky, Parkinson, Arcaroli, & Sammeth, 2006). Although tests using lists of single CVC words are naturally harder (due to lack of context) and therefore tend to be free of ceiling effects, sentence tests in quiet using methods of constant stimuli are unable to track improvements due to MAP adjustments because ceiling effects appear rapidly. Gifford et al. (2008) trialled four different speech recognition tests on 156 newly implanted patients and 50 hearing aid users: monosyllabic words (CNC) in quiet; the HINT and AzBio sentences in quiet (Spahr & Dorman, 2005); and the Bamford-Kowall-Bench Speech-In-Noise (BKB-SIN) sentences in noise. They found that 28% of the population tested achieved a perfect score of 100% and 71% of the population achieved scores greater than 85% correct on the HINT sentences in quiet. The BKB-SIN, a pseudo-adaptive test, showed no such ceiling effects. However,

Luxford et al. (2001) has suggested that the HINT be employed in cochlear implant assessment as originally intended, i.e. as an adaptive speech-in-noise test, as adaptive tests have the advantage of being able to be made harder if required by simply reducing the SNR.

1.8 Available speech-in-noise tests

A number of tests have been developed to assess speech communication in noise over the last few decades. Some examples include: the Connected Speech Test (CST; Cox, Alexander, & Gilmore, 1987); the City University of New York topic-related sentences (CUNY sentences; Boothroyd, Hanin, & Hnath, 1985); and the Speech Perception in Noise Test (SPIN; Kalikow, Stevens, & Elliott, 1977). These tests have been designed to measure percent intelligibility at fixed speech and/or noise levels, and are said to produce reliable estimates of performance. However, percent intelligibility measures, whether for speech in quiet or speech in noise, are inherently limited by floor and ceiling effects (Nilsson et al., 1994). In order to obtain scores at different SNRs, multiple test lists must be presented to the listener, resulting in a long and unwieldy testing process. Although the descending method has been used to develop more clinically useful speech-in-noise tests such as the Words in Noise test (WIN; Wilson, Abrams, & Pillion, 2003) and the Quick Speech in Noise test (QuickSIN; Killion et al., 2004), only the Hearing in Noise Test (HINT; Nilsson et al., 1994) is a truly adaptive speech test.

A description follows of some of the more common speech-in-noise tests used for clinical and research purposes in the English-speaking world. The advantages and disadvantages of each test will be discussed, along with information as to whether the test is used in New Zealand.

1.8.1 Speech Perception in Noise (SPIN)

The SPIN test (Kalikow et al., 1977) consists of eight sets of 50 sentences recorded by a male speaker and presented in speech babble. The listener is required to repeat back the last word of the sentence, a monosyllabic noun. Half of the sentences in each list are high predictability (PH), i.e. the last word is highly predictable from the semantic context, and the other half are low predictability (PL), or contextually neutral. The same word is used once in a PH sentence and once in a PL sentence in an attempt to separate out the various contributions of acoustic and linguistic information. In the selection of the final words, familiarity, predictability and phonetic content was carefully considered. Kalikow et al. (1977) reported

performance of young and old participants on high predictability (PH) and low predictability (PL) sentence items at four SNRs (-5, 0, 5 and 10 dB). Bilger, Nuetzel, Rabinowitz and Rzeczkowski (1984) queried the equivalency of the test forms and also the scoring and interpretation of the test, which was based on raw scores, allthough Gelfand, Ross and Miller (1988) found the PH materials to be reliable. This test was developed for other purposes than measuring SRTs and SRT measurements are thus time consuming and inefficient. This is due to the limited number of sentences, and the fact that scoring is based on individual words (one per sentence). This test has been used in a number of clinical studies overseas (Bentler, Niebuhr, Getta, & Anderson, 1993; Dirks, Kamm, Dubno, & Velde, 1981; Hutcherson, Dirks, & Morgan, 1979) but is not commonly used clinically. It is not used in New Zealand.

1.8.2 Connected Speech Test (CST)

This test of intelligibility of everyday speech has been developed primarily for use as a criterion measure in investigations of hearing aid benefit (Cox et al., 1987). The test consists of 48 passages of conversationally produced connected speech, presented in competing speech babble. Each passage is made up of 10 sentences containing 25 key words for scoring purposes, which are all related to the same topic. The listener is told the topic of each passage before testing begins. At least four passages are administered, and the results averaged, to yield a single intelligibility score. Performance is measured for at least 2 fixed signal-to-babble ratios during the test. This process can be performed in less than 10 minutes. Audiovisual versions of this test have also been developed. However, significant floor and ceiling effects have been noted. This test is available for purchase from the Hearing Aid research Laboratory at the University of Memphis and is used in a number of research facilities, but is not used in New Zealand.

1.8.3 City University of New York Topic Related Sentences (CUNY)

The CUNY test (Boothroyd et al., 1985) was designed so that the 12 sentences in each of the 72 lists were topic related. A percentage correct score is obtained from the number of words correct out of a total of 102 words per list. Within each list are four statements, four questions, and four commands. The sentence length varies from 3 to 14 words and is counterbalanced between the lists. Lists are presented in quiet at 65 dB A and in the presence of multi-talker speech babble at +10 dB signal-to-noise ratio. There are sufficient lists for repeated use of this test. However, the 95% confidence intervals for a score of 50% based on one list are approximately +/- 20 percentage points. Four sets reduce the confidence limits to +/- 10 percentage points,

and the use of 16 sets reduces this to +/- 5. As each set takes 2-3 minutes to administer, the time required to ensure test-retest reliability is beyond the scope of clinical use. Although the confidence intervals get smaller as scores approach 0% or 100%, this is still a limiting factor as floor and ceiling effects are also evident in this test. This test is not used clinically in New Zealand.

1.8.4 Speech in Noise/Quick Speech in Noise (SIN/QuickSIN)

The SIN (Etymotic Research, 1993) and later the QuickSIN (Killion et al., 2004) take a slightly different approach. Rather than measuring percent intelligibility, these tests attempt to find the SNR-50, or the SNR at which an individual can correctly repeat the speech material 50% of the time. This is similar to the concept of SRT. Despite reports of good face validity, it was felt that the SIN test was too time-consuming and difficult to score. Bentler (2000) noted that not all of the test blocks were equivalent, and that ceiling and floor effects had been recorded. The QuickSIN is a shortened version of the SIN that is administered as a block of six sentences, each containing five key words, presented in four-talker babble. The sentences are pre-recorded at different SNR decreasing in 5 dB steps from 25 dB SNR to 0 dB SNR. SNR-50 can be derived from the total number of key words repeated correctly using the Spearman-Karber equation, which then is used to find the SNR loss. This is a quick and useful test, with high face validity. A single test also takes about one minute to administer, with the recommendation that two tests are presented and the results averaged. In New Zealand, this test is currently being used in at least one clinic in Christchurch (P. Peryman, p.c).

1.8.5 Words in Noise (WIN)

The WIN test (Wilson et al., 2003) consists of blocks of 70 words taken from the NU-6 and presented in multi-talker babble at a fixed level (60 dB HL). Ten words are presented at each of seven signal-to-babble ratios starting at 24 dB and decreasing to 0 dB in 4 dB steps. As with the QuickSIN, the WIN uses the Spearman-Karber equation to determine the SNR at which speech recognition performance is 50%. Testing is stopped as soon as all ten words at one level are incorrectly identified. Subsequent studies have suggested dividing the 70 word blocks into two lists of 35 words that are administered sequentially and then combining the results (Wilson & Burks, 2005). This halved the time taken to administer the test, meaning that both ears could be tested in approximately five minutes. The WIN has been shown to be very sensitive to hearing impairment, with only 1% of HI listeners performing in the

normal range (Wilson et al., 2007). This test is not currently used clinically in New Zealand.

1.8.6 The Hearing in Noise Test (HINT)

A speech-in-noise test that uses an adaptive method to find the speech reception threshold (SRT) in noise is the HINT (Nilsson et al., 1994). The HINT was developed at the House Ear Institute in order to provide a reliable measure of SRT for sentences in quiet and in background noise (Nilsson et al., 1994). The test can be administered under headphones or in the soundfield, to allow for the assessment of different amplification options. The speech materials are based on the Bamford-Kowal-Bench sentences, transcribed from the utterances of British HI children (Bench, Kowal, & Bamford, 1979). Obvious Britishisms were replaced with more appropriate American phrases and the sentences were recorded with an American English speaker. This was the basis of the Hearing In Noise Test (HINT) which was normed for an American population. The HINT is also used as a screening measure to assess functional hearing for workers in hearing critical jobs, such as the coast guard, the police force and other law enforcement jobs. This test has recently been adapted for a large number of populations and languages: Canadian French (Vaillancourt et al., 2005); Latin American Spanish (Baron de Otero, Brik, Flores, Ortiz, & Abdala, 2008); Brazilian Portuguese (Bevilacqua, Banhara, Da Costa, Vignoly, & Alvarenga, 2008); Turkish (Cekic & Sennaroglu, 2008); Castilian Spanish (Huarte, 2008); Bulgarian (Lolov, Raynov, Boteva, & Edrev, 2008); Korean (Moon et al., 2008); Norwegian (Myhrum & Moen, 2008); Malay (Quar et al., 2008); Japanese (Shiroma, Iwaki, Kubo, & Soli, 2008); Cantonese (Wong, 2008); Taiwanese Mandarin (Wong & Huang, 2008); and Mainland Mandarin (Wong, Liu, & Han, 2008). For each language, a large number of sentences are collected and then prepared for testing following a common methodology. This system has produced a number of tests that are consistent across languages and can even be directly compared using a standardised HINT-score. Also, the HINT enables assessment of speech in quiet and in noise using the same materials and speaker with the same method. The HINT sentences have been rerecorded with a New Zealand speaker and are used clinically as part of cochlear implant assessment, as a fixed level test of speech in quiet. However, the HINT as administered adaptively in noise is not currently in use in New Zealand.

1.9 New Zealand English

New Zealand English (NZE) is a relatively young, homogeneous dialect of English spoken in New Zealand in the South Pacific. During the nineteenth century, the process of European settlement brought settlers from all over the British Isles and Australia and by the end of that century the emergence of a distinctly New Zealand accent had already been noted. The largest number of immigrants came from the south of England, meaning the influence of Southern English was substantial (Gordon et al., 2004). This, along with the relatively small size of the population and geographic area, prevented the formation of marked regional dialects. Southland and some parts of Otago were settled mainly by Scottish immigrants, and there is linguistic evidence of a semi-rhotic regional dialect still prevalent in these areas. This is the only generally accepted regional dialect within New Zealand.

1.9.1 Phonology

New Zealand English differs from American English (AmE) in a number of ways, most noticeably in the vowel system. NZE is essentially a non-rhotic dialect, in contrast to the General American unmarked dialect spoken in the New England, Midland and West regions of the United States. This dialect is considered more intelligible than the marked Northern and Southern dialects spoken in other parts, (Clopper & Bradlow, 2008; Clopper, Levi, & Pisoni, 2006). It is also the dialect spoken in the original AmE HINT recordings.

NZE vowels have a very different formant structure and place in the vowel space (Maclagan & Hay, 2007). Vowels can be discussed within the framework of lexical sets as introduced by Wells (1982). Standard lexical sets of keywords can be useful in specifying vowel phonemes, as they are unambiguous in whatever dialect is being used. The front vowels in the words '*heed*', '*head*' and '*had*', referred to by the lexical set labels FLEECE, DRESS and TRAP respectively, have raised and fronted in NZE causing them to be pronounced much higher in the mouth, similar to Australian and South African English. DRESS is sometimes pronounced so high in the vowel space by some speakers that it can overlap with FLEECE (Maclagan & Hay, 2007). There is further neutralisation of front vowels before /l/, such as in the word pairs *celery /salary,* and *doll, dole* and *dull*. The vowel in *'hit'* (KIT) has centralised and lowered even further than when Wells described it (1982), leading speakers of other dialects of English (mainly Australians) to poke fun at the perceived pronunciation of *'fish and chips*' as *'fush un chups'*. NEAR and SQUARE are completely merged for many speakers, so that '*cheer*' and '*chair*', '*beer*' and '*bare'* are pronounced identically. The

GOOSE vowel is also very central, even fronted in some cases, except before /l/. Native NZE speakers do not usually consider that the words *goose* and *school* have the same vowel. These differences can be observed from the acoustic vowel plots of each dialect (figures 1 and 2).

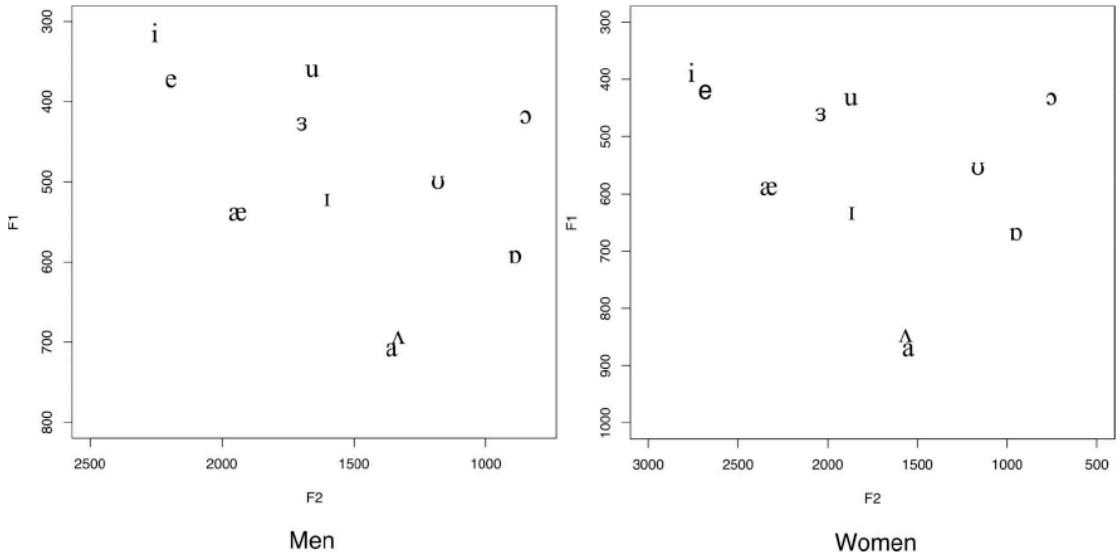

Figure 1. Acoustic plot of first and second formants of the vowels of 20 men (left) and 20 women (right) born c.1970. Data from Maclagan and Hay (2007).

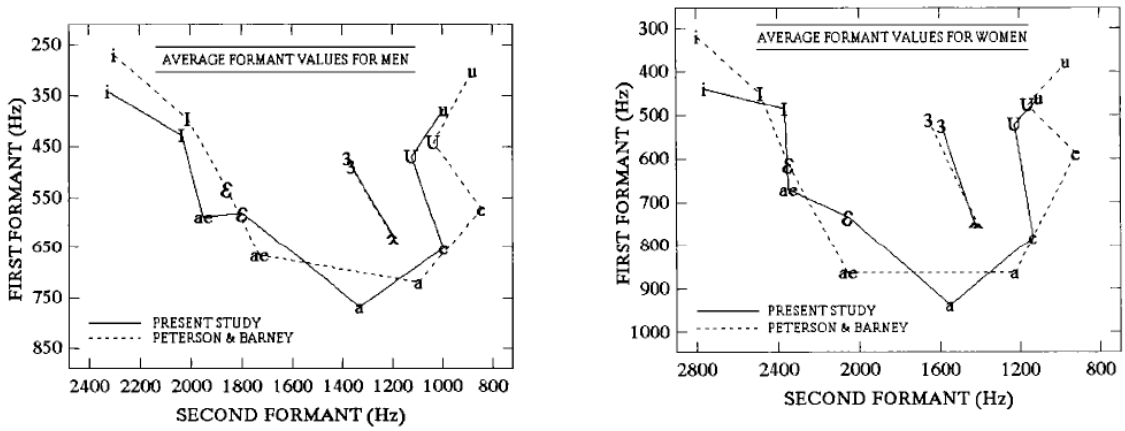

Figure 2. Acoustic vowel plot showing average formant frequencies for men from both Hillenbrand, Getty, Clark and Wheeler (1995) and Peterson and Barney (1952). Figure taken from Hillenbrand et al. (1995).

There are a number of differences between NZE and AmE consonants. AmE has a clear-/l/ which is used in syllable initial positions in words such as *like* and *leap*. The dark-/l/ or postvocalic /l/ appears in syllable final position. NZE has maintained the

clear-/l/ in syllable initial positions, but dark-/l/ in syllable-final positions is often not pronounced with the tongue on the alveolar ridge. This is called /l/-vocalisation, where the /l/ is essentially being replaced by the FOOT vowel. In contrast to AmE, there are a number of words in NZE where /j/ is inserted between certain consonants and the vowel /u/, e.g. *student* and *new*. AmE and NZE also differ in rhoticity, as NZE is a non-rhotic dialect. A relatively recent change in NZE is the affrication of the consonant clusters /tr/, /dr/ and /str/. The production of /tr/ now sounds more like [tʃɹ]. One way in which NZE is becoming more like AmE is in the increasing use of t-flapping, which is the flapping or voicing of inter-vocalic /t/ in words like *butter* and *city*.

These differences may cause NZE speakers to perform more poorly than expected on the HINT. As an example, in the sentence 'The letter fell on the floor' the first vowel in the word *letter* as spoken by an American speaker has a much higher first formant (F1) and a slightly lower second formant (F2) than NZE /e/. This formant structure is closer to NZE / æ /. In addition, inter-vocalic /t/ is often "flapped" in AmE, causing the NZE listener to perceive the word as *ladder*.

1.9.2 Vocabulary

There are significant differences in vocabulary between AmE and NZE. A number of words included in the original HINT would not be considered familiar in a New Zealand context, such as 'rancher' (farmer), 'faucet' (tap), 'vacation' (holiday), 'pitcher' (jug) and 'jelly jar' (jam jar). (See Hay, Maclagan and Gordon (2008) for a full description of NZE.)

Given that speech perception materials will be presented to HI individuals under challenging listening conditions, these differences in phonology and vocabulary could have an impact on the performance of NZE speakers on the original HINT. List equivalency could potentially be affected, as some lists may be more difficult than others, especially when listening at a low SNR. Thus, the validity and reliability of the AmE HINT test may be compromised when used for NZE speakers.

1.10 Necessity for a New Zealand speech-in-noise test

There is no speech-in-noise test currently in existence that is specifically tailored for a New Zealand population. The QuickSIN and HINT tests are both used in clinical and research situations in New Zealand. However, no normative data has been collected that describes the acceptable range of performance on these tests for NH New Zealand listeners. This raises serious issues of validity. The tests are administered,

scored and interpreted based on assumptions that all sentences are equally intelligible and that all lists will give equivalent results (the same SRT) when presented to the same listener.

Performance on speech-in-noise tests has been shown to be affected by certain characteristics of speech materials and speaker, such as the phonetic similarity of words, speaking rate, the gender and dialect of the speaker, and even the naturalness of the speaker's voice (Picheny, Durlach, & Braida, 1985). These factors all contribute to variations in the intensity and/or duration of specific consonant and vowel segments, the stress of the syllables and modifications to the prosody of the speech materials (Gordon-Salant, 2005). Elderly participants appear to be particularly affected by speaker characteristics, especially accent. A study comparing the speech recognition performance of young and elderly native speakers of English for speech of non-native English speakers found that elderly participants scored lower than younger participants, and these scores were associated with the strength of the speaker's accent (Burda, Scherz, Hageman, & Edwards, 2003). As the great majority of the HI population can be described as elderly, it is doubly important that test materials are as natural and familiar as possible to avoid or at least reduce the influence of confounding factors.

Other studies have noted differences in performance of non-North American populations when compared to North American norms. Marriage, King and Lutman (2001) found that British children performed poorly compared to North American norms when tested using a screening test for Auditory Processing Disorder (SCAN; Keith, 1986) recorded by an AmE speaker. Analysis of word errors indicated accent and word familiarity effects. The authors of this study recommended that the SCAN test material be reviewed, replacing high error-rate target words with more appropriate vocabulary items, then re-recorded by a British English speaker. Finally, they recommended that normative data be collected for the new test material. A study looking at the performance of Australian and American children on the SCAN-A: Test for Auditory Processing Disorders in Adolescents and Adults (Keith, 1994), showed significant differences in the performance of Australian children on some sections of the test, suggesting that it is not adequate to simply use North American norms when testing other populations (Sockalingam et al., 2004). However, there was no statistical difference between the performance of Australian and American children on the section of the test which specifically looked at speech understanding in noise.

Due to differences in vocabulary and dialect between the AmE HINT test materials and NZE as spoken in New Zealand, there may be differences in intelligibility

between sentences and thus significant differences in list equivalency. This would suggest that it may not be appropriate to test a New Zealand population on an unmodified American test.

This study aims to develop and evaluate a speech-in-noise test for the New Zealand adult population and to collect preliminary normative data on this test. A secondary aim is to compare the performance of native NZE speakers on both the AmE HINT and the NZE HINT. Using an adaptive test procedure will help ensure that the resulting test is reliable, efficient, and clinically useful. The collaboration with House Ear Institute will enable a New Zealand version of the HINT test to be recorded, developed and normed, and subsequently included as part of the HINT software that is distributed internationally. The benefits of making a New Zealand version of an already-established test is that scores on the NZHINT can be translated into a standardized HINT score, the H-score, which can then be used to compare HINT scores directly across languages (Soli & Wong, 2008). There would also be the potential for developing the NZHINT further as a screening tool for workers in hearing critical jobs in New Zealand as has already been done in other countries (Giguere, Laroche, Soli, & Vaillancourt, 2008).

Developing a corpus of NZHINT sentences will be of benefit to research. A significant corpora of sentences recorded by a New Zealand speaker that demonstrate equivalent intelligibility and for which normative data exists will be able to be used in many different types of research that require a measure of sentence intelligibility in quiet or noise.

2 Methodology: Developing the NZHINT

This method closely follows the protocols developed by the House Ear Institute, Los Angeles, for use in the development of the HINT in a new language (see Appendix 1). A large number of short sentences are prepared and then recorded by a native speaker of the language. A masking noise with the same long term average spectrum as the speech signal is generated. The sentences are then equalised for intelligibility and formed into phonemically balanced lists of equivalent difficulty. Preliminary norms can then be established by obtaining speech reception thresholds (SRTs) via an adaptive test method in a number of different SNR conditions. This chapter will describe the general methodology and provide specific details relevant to developing the NZHINT.

2.1 Participants

Participants who were used to evaluate the stimuli at each stage of the process were recruited mainly from the student population of the University of Canterbury. Participation was voluntary and based on availability. All participants met the common inclusion criteria, as all 1) were born in New Zealand, were native speakers of NZE, and were educated in New Zealand until age 17; 2) were between 18 and 50 years of age; 3) had hearing thresholds equal to, or better than, 25 dB HL from 250 Hz to 8000 Hz; 4) presented no asymmetry between the two ears (defined as a difference of 30 dB or more at one frequency, 20 dB at two frequencies or 10 dB at three frequencies); 5) had no abnormalities noted on otoscopy; 6) showed normal tympanograms (Type A/As/Ad); and 7) reported no otologic history. The upper age limit was set at 50 in an attempt to reduce the effect of any age-related deterioration in speech intelligibility (Hanks & Johnson, 1998; Weinsten, 2002). Different groups of participants fulfilling the inclusion criteria were recruited for each subsection of the study, leading to a total of 61 participants. Ethics approval was granted by the Department of Communication Disorders, University of Canterbury. The consent form, information sheets and participant questionnaire are in Appendix 2.

2.2 Equipment

Puretone hearing thresholds were measured using a GSI-61 audiometer and TDH-39P headphones (calibration due date 5 February 2010). The hearing threshold tests were carried out using the modified Hughson-Westlake technique (Carhart & Jerger, 1959) in a sound-treated booth. Impedance data was obtained on a GSI Tympstar (calibration due date: 26 August 2010). All testing involving the HINT materials was

carried out with sentences being presented through Sennheiser HD215 headphones driven by an InSync Buddy USB 6G soundcard attached to an Acer TravelMate 210TER laptop computer.

The calibration for the first stages of HINT development up until the final testing of the 10-sentence lists (see *2.3.6 Creating phonemically-matched lists*) is as follows. After setting the soundcard output settings to maximum, the headphones were placed on a Brüel & Kjær Type 4128 Head and Torso Simulator (HATS) connected to a Brüel & Kjær 7539 5/1-ch. Input/Output Controller Module. The average A-weighted sound level of the noise was measured using Brüel & Kjær PULSE 11.1 noise and vibration analysis platform. As the masking noise output was measured at 102 dBA, a programme was written that attenuated the wave output of the headphones by 37 dB, ensuring that the masking noise would always be presented at 65 dBA. Before each testing session, the volume controls were turned up to maximum and the programme was then run to set the appropriate levels for the presentation of the materials. A biological check was conducted before each testing session as a subjective loudness test to ensure that the levels were consistent between testing sessions.

For the final testing stages using the adaptive HINT software, the output levels from the Buddy USB soundcard were measured on the HATS through the right and left headphones separately. The first set of data from four participants was collected with the soundcard volume turned up to maximum. For all subsequent testing, a Rolls MX28 Mini-Mix VI mixer was used to attenuate the maximum output of a 1000Hz puretone (110 dB SPL) until an output of 90 dB SPL was recorded. These values (110 dB SPL for the first set of data and 90 dB SPL for the second set) were entered into the HINT software for both left and right outputs and used to set the presentation level of both speech and noise. For further discussion as to why two different calibration procedures were used, see section *3.6 Limitations*.

2.3 Procedure

At the beginning of the process, the protocols for the development of the HINT in other languages were received. While this was a useful guide, it was often necessary to go back to the original AmHINT article (Nilson et al., 1994) and the procedure described by Vallaincourt et al. (2005) in the development of the French-Canadian HINT for more detailed information. Further changes were made in consultation with Andy Vermiglio and Sigfrid Soli at HEI at various stages of the process.

2.3.1 Preparation of written sentences

The objective of this step was to produce a set of 500-600 short, simple, natural-sounding sentences that would be recorded for possible use in the new version of the HINT. These sentences needed to be short enough to minimise any undue strain on memory. The "naturalness" of the sentences in terms of vocabulary, syntax, and usage were rated by native speakers of the language.

Sentences were gathered from a wide selection of New Zealand children's books. The sentences were modified to be as natural as possible while of only 5–7 syllables in length. Sentences were chosen to reflect common NZE vocabulary and sentence structure, but proper names, NZE slang, and unusual words were avoided. While some Maori names of flora and fauna were included (e.g. *tui*, *weka*), it was felt that incorporating names where there is a Maori pronunciation and a common anglicised pronunciation could be counterproductive. For example, would *kowhai* be pronounced /kɔfai/ or /koʊwai/? It was felt there could be an age related difference in the perception of these words, with older people being more familiar with the anglicised pronunciation. For these reasons few Maori words were included.

507 sentences were prepared and distributed to five native speakers of NZE. They were asked to rate the sentences on a scale of 1–5, with 1 being "very unnatural; I would never say this" to 5 being "very natural; I would quite happily say this". They were also asked to provide alternatives or suggestions as to how the naturalness of the lower rated sentences could be improved. The suggestions provided by the raters were then used to increase the naturalness of any sentences with an average rating of less than 4.5. The resulting sentences were reviewed by a linguist and an audiologist. During this process, 41 sentences were altered and 5 eliminated, to give a total of 502 sentences (see Appendix 3).

2.3.2 Selection of speaker

The objective of this step was to find a native speaker of the language who displayed sufficiently typical features of the language while remaining broadly intelligible with no unusual regional or dialectal features. The speaker was required to be a professional voice actor or otherwise have a clear voice suitable for recording purposes.

In conjunction with the New Zealand embassy, a 56-year-old male NZE speaker living in Los Angeles was recruited to record the 500 NZHINT sentences. Although he had lived in Los Angeles for almost 20 years, he had strong links to New Zealand through his wife (also a New Zealander) and family in New Zealand, encouraging

regular visits home. In contrast to most speakers found by HEI for the development of the HINT in other languages, this speaker was not a trained voice actor, instead working in real estate. Nevertheless, the speaker's voice was judged to be typically NZE-sounding and suitable for recording purposes by HEI members and two New Zealanders, a linguist and an audiologist.

This was confirmed through post hoc analysis of the speaker's vowels. For each of 11 monophthongs and 5 diphthongs, six words from the 500 sentences were chosen to be good examples of these phonemes in stressed positions in the sentences. Acoustic analysis of an individual speaker is often done from a recording of a list of so-called /h_d/ words (eg *hid, head, had*), which feature each vowel as stressed within known constant consonants. As this was not done, the words chosen for analysis displayed the vowel in any appropriate CVC frame, avoiding the consonants /r/, /l/, /w/ and /j/. This was to limit the possibility of these approximants unduly influencing the vowel formants. For each phoneme, the six examples were analysed using Praat, a free acoustic analysis program available under GNU license (version 4.8.09 Boersma and Weenink http://www.fon.hum.uva.nl/praat/). The first and second formants were measured at a point judged to be maximally stable, called the target. If there was no steady state, formant readings were taken at the F2 maximum (and F1 minimum) for front vowels, the F1 maximum (and F2 minimum) for central vowels and the F2 minimum (and F1 minimum) for back vowels. For the diphthongs, a measurement was taken at each of the two targets corresponding to the two vowels of the diphthong.

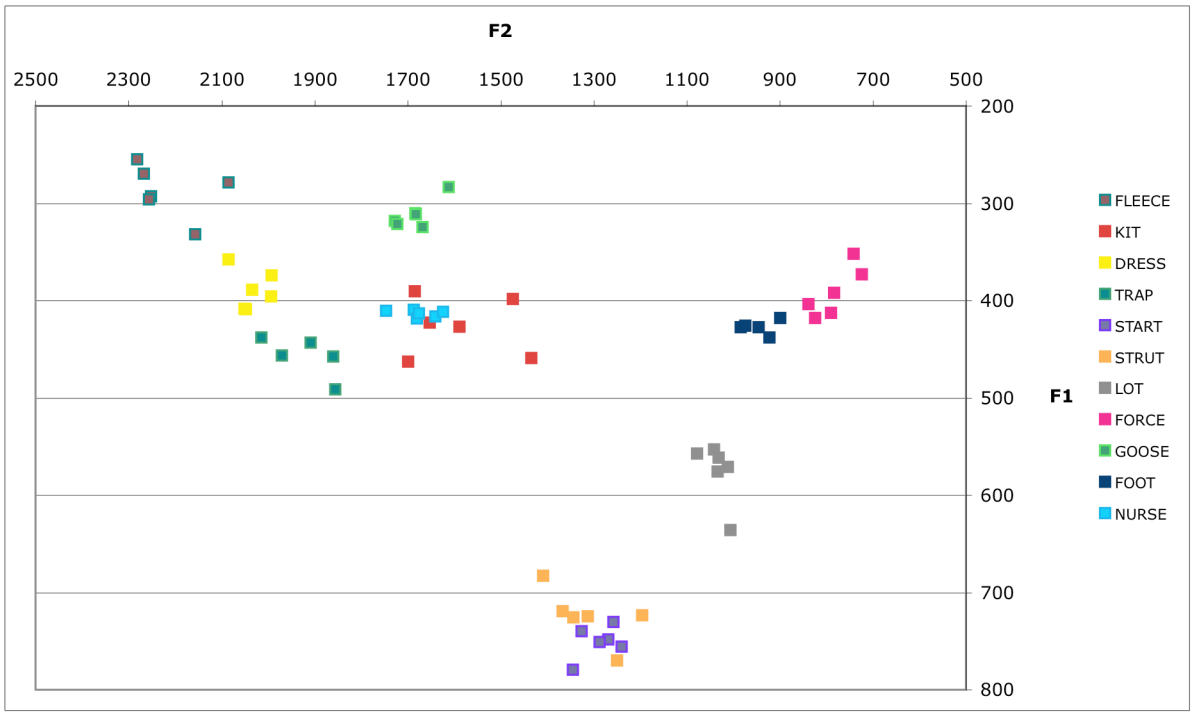

Figure 3. Acoustic plot of the first (F1) and second (F2) formants of all tokens of the vowels of the NZHINT speaker.

The measurements from each of the six tokens are shown on the graph in figure 3. These measurements were averaged to give the vowel plot shown in figure 4. This could then be compared to average data on NZE vowel placement collected from the Origins of New Zealand English (ONZE) project (http://www.ling.canterbury.ac.nz/onze/).

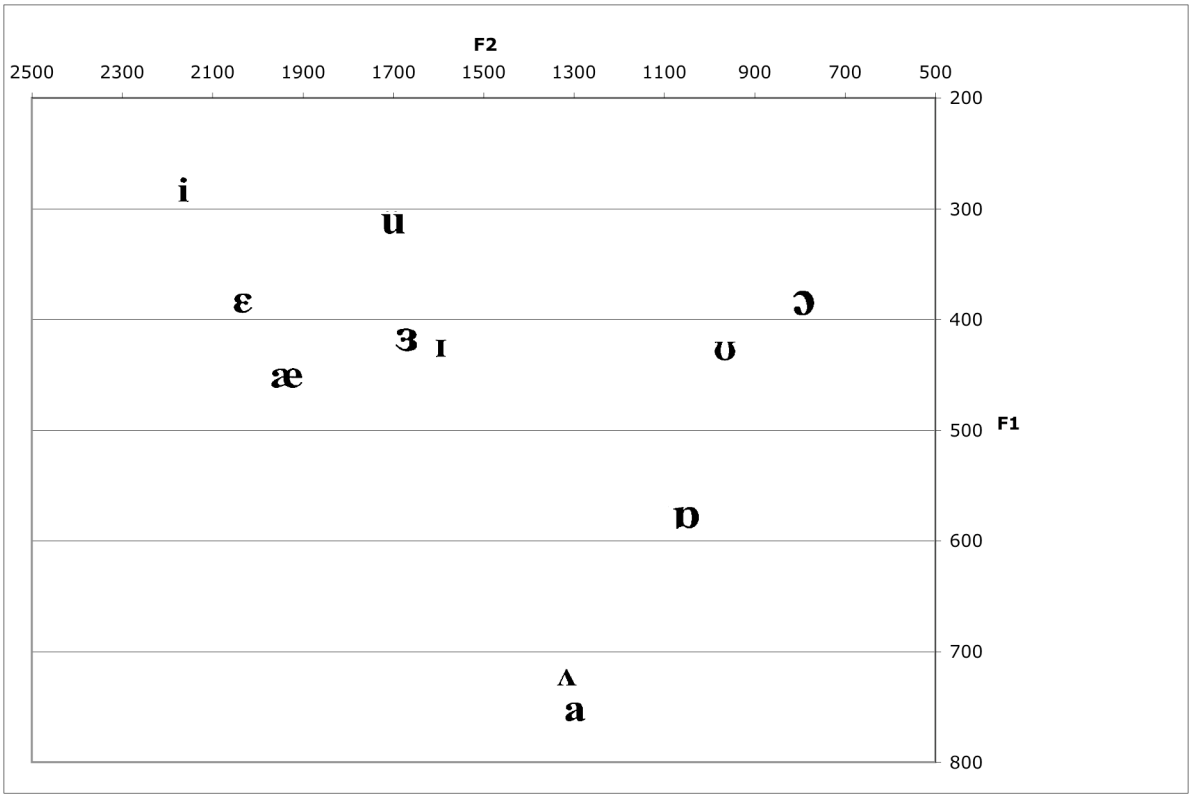

Figure 4. Acoustic plot of the average first (F1) and second (F2) formants of the vowels of the NZHINT speaker.

When compared with the data from the ONZE corpora, the vowel plot from this speaker looks very similar to that of other New Zealanders born in the 1950s (figure 1, reproduced below).

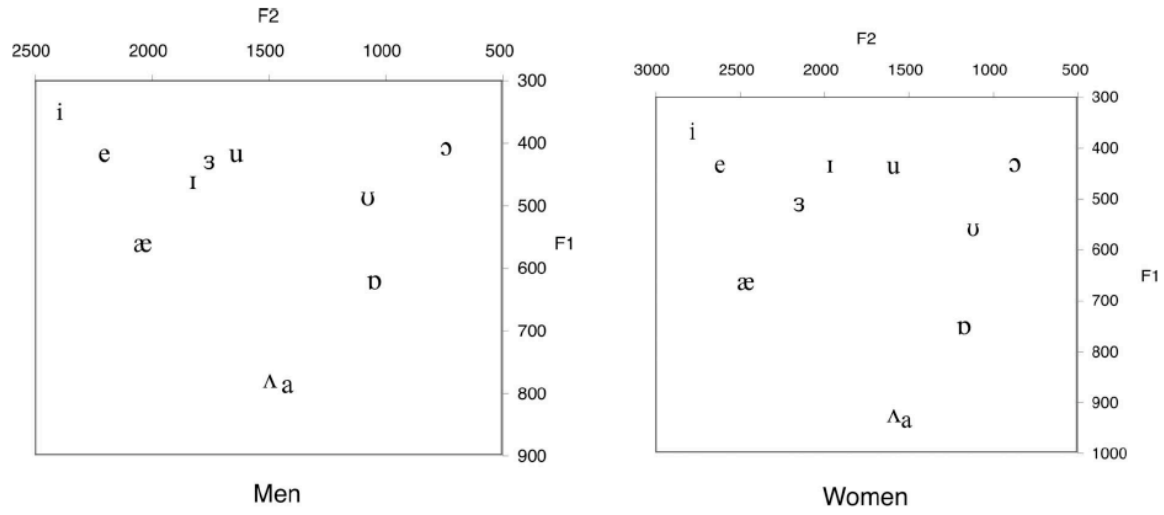

Figure 5. Acoustic plot of the first and second formants of the vowels of 25 men (left) and 25 women (right) born about 1950. Data from Maclagan (1982).

However, the speaker's vowels are slightly more advanced than the average vowels shown here, especially in the case of the /ɪ/ vowel, which has lowered more. To compare further, here is the vowel plot for speakers born c 1970 (figure 1, reproduced below for convenience).

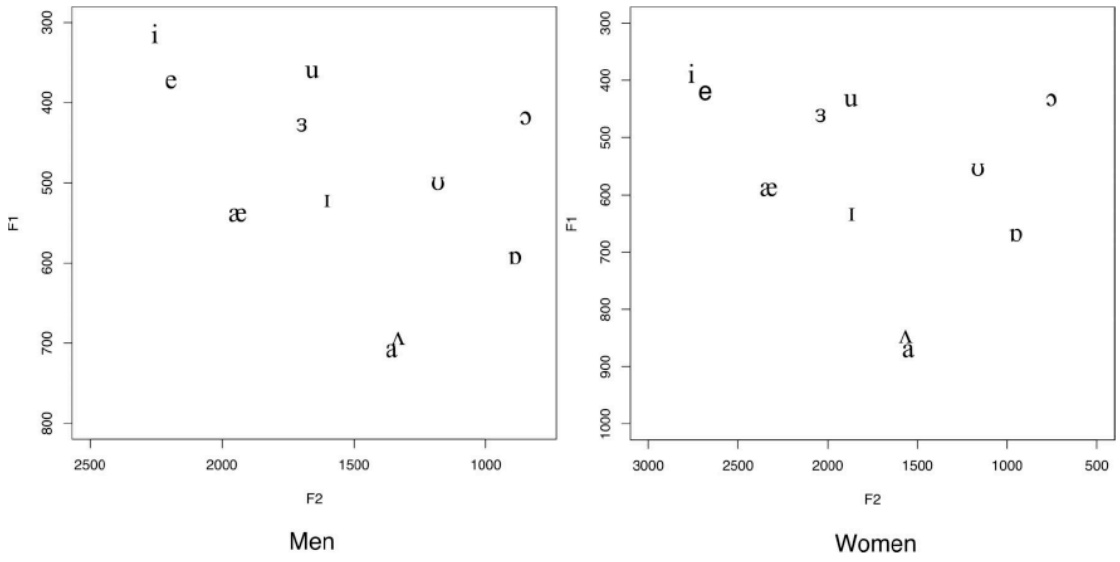

Figure 1. Acoustic plot of the first and second formants of the vowels of 20 men (left) and 20 women (right) born about 1970. Data from Maclagan and Hay (2007).

The most recent acoustic data recorded and analysed for Mid West American speakers can be seen in figure 2, reproduced below for convenience. It can be seen that the vowels of the NZHINT speaker are much more aligned with the New Zealand equivalents than the American vowel plot. This is most obvious in the raised DRESS vowel, the centralised KIT vowel, and the fronted GOOSE vowel, all hallmarks of the New Zealand accent.

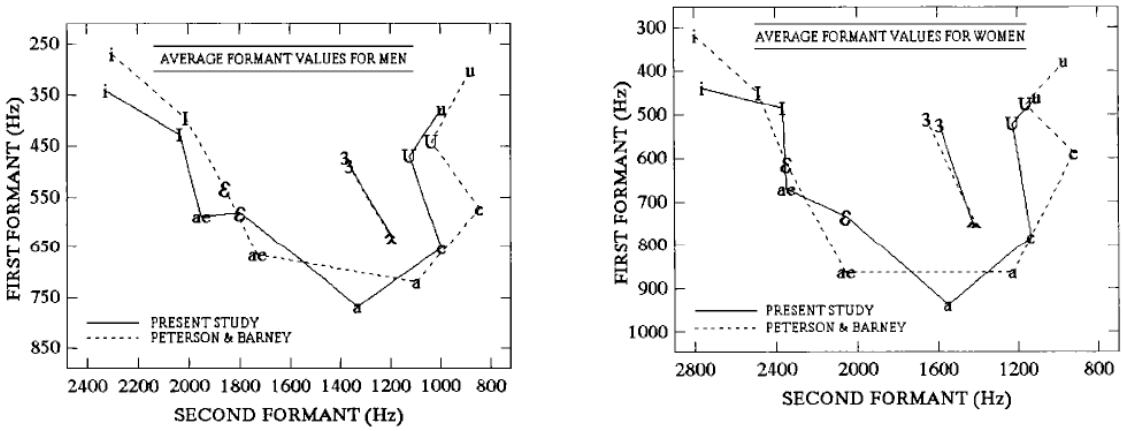

Figure 2. Acoustic vowel plot showing average formant frequencies for men from both Hillenbrand et al. (1995) and Peterson & Barney (1952). Figure taken from Hillenbrand et al. (1995).

The diphthongs recorded from the NZHINT speaker in figure 8 are similar to the NZE data from ONZE in figure 9, although two are slightly more conservative. The first target in FACE is in a more raised position than would be expected. This indicates that the diphthong is more conservative than for most NZE speakers and relatively more conservative than the speaker's monophthongs. The targets of MOUTH are also different, as the second target in this diphthong of the NZHINT speaker is fronted and raised, in contrast to the equivalent target in the ONZE data, which backs and lowers, again producing a somewhat more conservative diphthong. This is unlikely to have any impact on intelligibility.

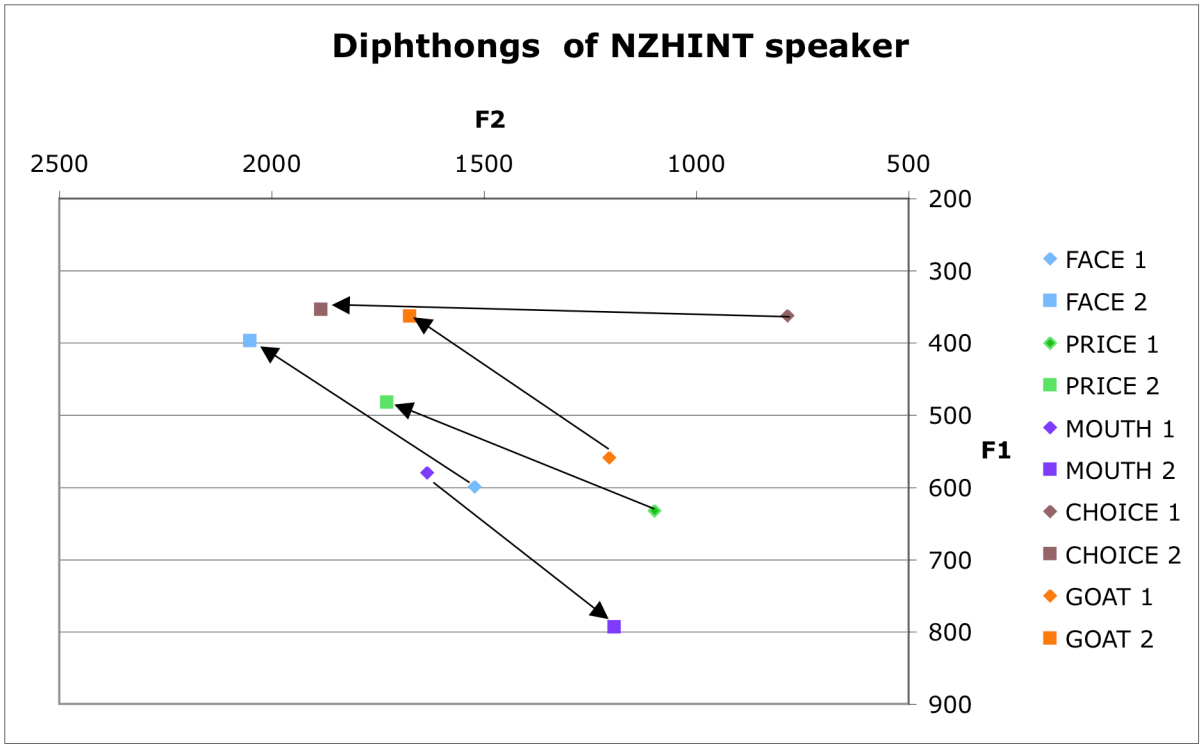

Figure 6. Acoustic plot showing average formant frequencies of diphthongs from NZHINT speaker.

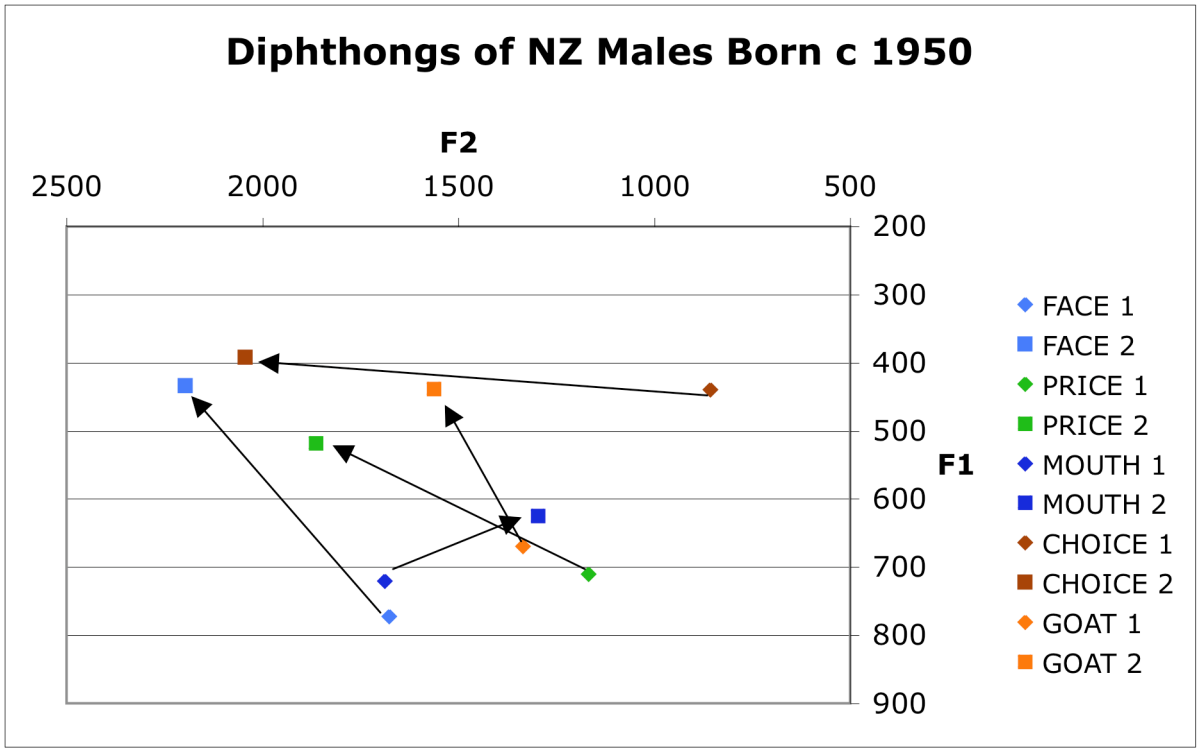

Figure 7. Acoustic plot of first and second formants of diphthongs from 20 NZE males born c1950. Data from Maclagan (1982).

As can be seen from figures 1-7, the acoustic data supports the assertion that the NZHINT speaker has vowels that are typical of NZE.

2.3.3 Recording of the sentences

The objective of this step was to make high quality digital recordings of the set of sentences developed in the first stage (section 2.3.1). These recordings will be edited and processed for subsequent use in the listening tests to be performed in the country where the HINT will be used.

The Excel file containing the approved sentences was sent to the House Ear Institute (HEI) in Los Angeles for recording. The sentence lists were recorded at HEI by the native speaker of NZE. All recording and editing of the sentences, including generation of the masking noise, was carried out by HEI according to their standard procedures. This was to ensure that the New Zealand material would be comparable in quality to the material developed and recorded for the HINT in other languages. All recordings were carried out in a sound-treated booth. The speaker was instructed to say each sentence slowly and clearly at a similar volume, in their normal accent, without unduly emphasising any words. The speech was recorded with a Larson Davis 2575 mic, B&K ZC0020 mic preamp, B&K 2609 amplifier, and Panasonic SV-2700 DAT deck at 44.1 kHz.

The signal was digitally transferred from the DAT deck to the PC with an M-Audio Transit audio interface. The recording was then hand-edited into individual sentences before being downsampled to 24 kHz. Any infrasonic ambient noise (building rumble) was filtered out using a finite impulse response (FIR) highpass filter with stopband from 0 to 20 Hz, passband above 50 Hz, and 60 dB stopband attenuation.

Each sentence was equalised to the same A-weighted root-mean-square (RMS) level. The long-term average spectrum of all sentences was calculated. A masking noise was generated by synthesising white noise, filtering it through another FIR filter, and then scaling it to the same RMS amplitude as the sentences. This ensured that the resulting speech-spectrum noise was matched to the speaker, so that the SNR at any frequency is approximately the same. A peak-limiting algorithm was applied to the recordings to reduce the crest factor (peak/RMS ratio) of each sentence. The sentences and noise were placed in .wav format for use with the testing software.

The recorded sentences (see Appendix 1) were assessed auditorily by a linguist and an audiologist to check that they matched expected NZE pronunciations. From the 502 sentences recorded, 48 sentences were dropped from further participation in the

development of the NZHINT due to unnatural intonation and/or non-typical NZE pronunciation. In particular, examples of *–ing* pronounced as /ɪn/ rather than /ɪŋ/ were removed.

2.3.4 Determining the Performance-Intensity function

The objective of this step was to estimate the relationship between SNR and percent word intelligibility. This relationship can be defined by the performance intensity (PI) function, which was to be used in the remainder of the project to determine a) the SNR at which sentences were first presented, before any adjustments to the RMS level were made and b) a general rule for how to adjust the SNR of individual sentences to equalise their intelligibility in noise. From similar experiments, it was expected that the slope of the PI function would show that an increase of 1 dB SN resulted in approximately a 10% increase in speech intelligibility.

The PI function was determined by recording the speech intelligibility scores achieved by nine participants at three different SNRs (-7 dB, -5 dB and -2 dB). A subset of the recorded sentences consisting of the first 150 sentences divided into three 50-sentence lists was used in this study, and each subject was tested with all three lists. List order was counterbalanced between participants.

The sentences were presented to the nine participants via headphones in the presence of 65 dBA masking noise. The participants were instructed to listen carefully to each sentence and repeat aloud as much of the sentence as possible. Guessing was encouraged if participants were not entirely sure of the sentence. The responses were scored by word. The average percent intelligibility at each SNR was calculated from the number of words repeated correctly. A linear function was fitted to the three data points (one for each SNR) giving a PI function with a slope of 7.5% per dB.[1]

2.3.5 Equalizing sentence intelligibility

For this adaptive test to be valid, the intelligibility of sentences must be approximately similar, so that different lists will yield the same SRT. Sentence intelligibility is affected not only by the overall RMS level but by the phonemic content, familiarity of words and variations in level or intonation (Nilson, 1994). The specific aim of this phase of testing was to adjust the RMS level of each individual sentence so that participants scored an average of 65% correct when scored by word,

[1]The function was calculated by HEI using the data provided to them. Information on how this was done (e.g.the parameters used) could not be obtained.

in the presence of masking noise. A target word intelligibility score of 65% was chosen in order to avoid floor and ceiling effects, and because that was the average intelligibility score over all sentences from Round 1. It was also close to the 70% word intelligibility score that has been associated with 50% sentence intelligibility (the SRT for sentences) (Nilsson, Sullivan, & Soli, 1991a, b). Several rounds of testing were anticipated until the average intelligibility of each sentence was 65%.

Sentences were scored by word, and the scores entered into an Excel spreadsheet. The results were sent to HEI so the appropriate adjustments could be made to the SNRs of the sentences that deviated from 65% intelligibility. The SNR of easy sentences was lowered, making them quieter in relation to the noise and therefore more difficult. The SNR of difficult sentences was raised, making them louder in relation to the noise and therefore easier. The amount by which sentence SNRs levels were raised or lowered was based on the slope of the PI function determined in the previous study. Very easy and very difficult sentences were discarded, as well as sentences that did not show the expected changes in intelligibility after adjustment of their SNR. After HEI had adjusted the intensity levels, a new round of testing was carried out with a new set of ten participants. Testing, adjustment and discarding was continued until all the remaining sentences reached an average of 65% intelligibility. This process required three rounds of testing and produced a set of 244 HINT sentences of approximately equal intelligibility.

a) Round 1

The 455 NZHINT sentences that remained after the auditory checking of the recordings were divided into two sets, Set A and Set B. These contained four lists of 50 sentences and a fifth list of 27 and 28 sentences respectively. Each set was presented to one of two groups of 10 participants at an SNR of –6.8 which corresponded to an average of 65% intelligibility in the previous testing phase. The lists were presented in counterbalanced order, with the same set-up and instructions as in the previous testing phase. The average intelligibility of each sentence was calculated. The mean intelligibility across all the sentences was 65% (see figure 8).

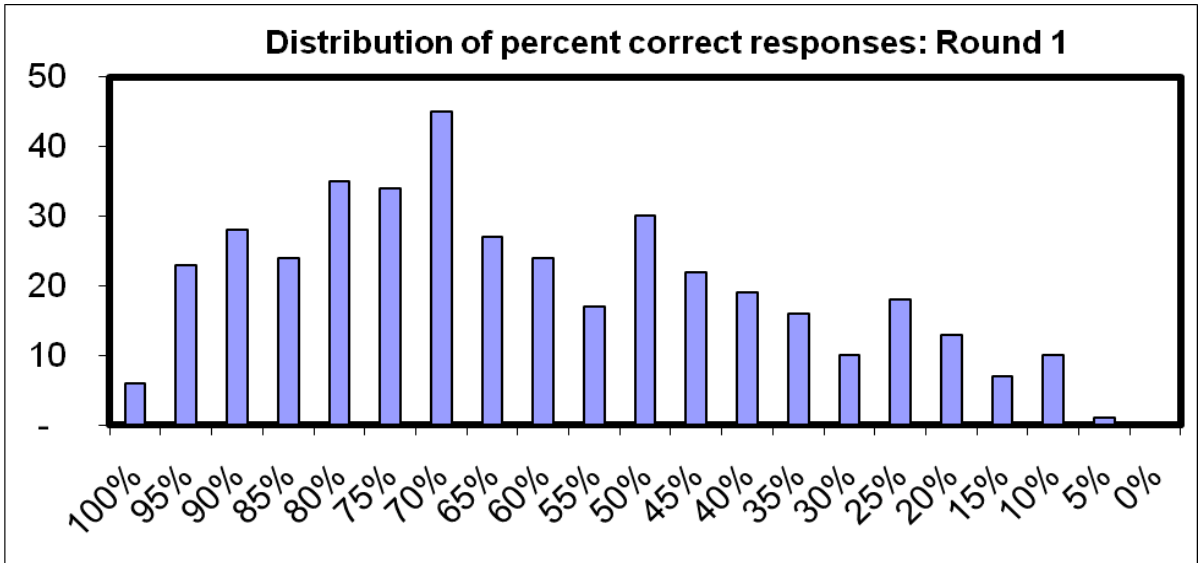

Figure 8. Distribution of percent correct responses

Any sentence with a lower or higher intelligibility score had an 'SNR correction' applied, which was the change in SNR predicted to achieve a change in intelligibility so that the sentence intelligibility would be approximately 65%.

A total of 46 sentences were dropped because at least 9 out of 10 participants scored 0% *or* 100%. For 80 sentences, the required adjustment exceeded 3 dB SNR, which would lead to inappropriately large jumps in volume between sentences. These sentences were also discarded, leaving a total of 326 sentences that were adjusted and used in the following round.

b) Round 2

New text files and score sheets were generated for Round 2. This round consisted of 326 sentences divided into six lists of 50 and one list of 26. Ten participants listened to all sentences and their intelligibility scores were averaged for each sentence.

Rather than decreasing, the subject variability increased after this round. This was unexpected. On closer analysis of the data, it was decided that the speaker variability at the beginning in the PI function stage may have contributed to the increased variability in this round. Unwanted variability in subject responses resulted in a PI function whose slope was somewhat shallower (7.5%/dB) than what had been observed for all other languages (typically about 10%/dB). The slope from the PI study was used to adjust the S/N ratios after Round 1 of the EQ study. As the slope was underestimated (i.e., too shallow), the S/N ratios tended to be over-corrected. More than 200 sentences were over-corrected for Round 2 testing. This meant that the adjustments to make them easier made them too easy, and adjustments to make them harder made them too hard. Only about 65 sentences appeared to have been

properly corrected. For the adjustments for Round 3, a 10% slope was used to calculate the new SNRs.

Contributing to the variability was the performance of four of the ten participants. They did significantly worse than the other participants, scoring below 60%. They also achieved considerably more 0% scores (over 50) than the other participants, who scored 0 on fewer than 30 sentences. Subject 4 scored 0 on over 100 sentences. The data from these four participants were not considered in the adjustments for this round.

c) Round 3

Ten new participants listened to the 326 sentences that had been adjusted following the previous round, and the intelligibility scores were averaged. The average percent correct achieved in this round was 70%. The measurement error of the HINT is generally ±1.5 dB, which equates to 15% intelligibility. Due to the variability of the participants in the preceding rounds, it was felt that this could be extended to encompass sentences that were within ±20% intelligibility of the target. This process yielded 244 sentences that showed 70% ±20% intelligibility. These sentences formed the materials for the NZHINT.

2.3.6 Creating phonemically matched lists

The objective of this stage of the project was to form 24 phonemically matched 10-sentence lists from the 244 sentences. After further testing, pairs of these 10-sentence lists were combined to make 12 20-sentence lists that form the materials for the NZHINT.

A phonemic transcription of each of the 244 sentences was made with the object of forming 24 10-sentence lists whose phoneme distribution matched that of the overall pool of sentences (see Appendix 4). The transcriptions from 50 randomly selected sentences were compared to transcriptions of these sentences done by a linguist unconnected with the project. The inter-rater reliability was 96%. The phonemic distribution of all phonemes is shown in figure 12. The percentage occurrence in figure 12 represents the percentage occurrence of the phonemes in this set of sentences. It does not claim to be representative for NZE as a whole. There is currently no indication of the approximate phoneme distribution in spoken NZE against which these percentages could be compared. Because they are based on simple sentences rather than conversation the phoneme distribution may, in fact, be somewhat different from normal spoken NZE.

Table 1. Phonemic distribution of all phonemes in NZHINT sentence set

Consonants		Vowels	
Phoneme	Frequency of Occurrence (%)	Phoneme	Frequency of Occurrence (%)
t	7.6%	ə	12.7%
ð	6.2%	i	4.3%
n	4.9%	ei	2.8%
d	4.8%	ɒ	2.3%
s	4.2%	æ	1.9%
l	4.1%	ɪ	1.9%
k	4.1%	e	1.7%
w	3.3%	ʌ	1.7%
z	2.8%	ai	1.4%
m	2.7%	ɔ	1.3%
h	2.5%	oʊ	1.0%
r	2.5%	u	1.0%
p	2.1%	au	0.9%
f	1.8%	ɜ	0.8%
b	1.8%	a	0.8%
ʃ	1.6%	ʊ	0.6%
v	1.2%	iə	0.6%
g	1.1%	ɑʊ	0.4%
ŋ	0.9%	ɔi	0.2%
tʃ	0.6%		
θ	0.4%		
dʒ	0.3%		
j	0.3%		

By a process of trial-and-error, the sentences were shuffled into lists of 10 whose phonemic distribution showed the closest match to the overall phonemic distribution.

The success of the phonemic matching was judged by calculating the percentage of all phonemes in each list which fell within 2.5% of the overall phoneme distribution. The NZHINT achieved over 99% of phonemes falling within this range. This compares favourably to more recent versions of the HINT, e.g. Brazilian Portuguese, Norwegian and Korean (all 99%), Latin American Spanish (98%), Turkish (95%) and Malay (92%). Four sentences that were surplus to requirements were discarded.

Once these lists were formed, adaptive HINT thresholds were measured for each list in a sample of 12 participants who met the same selection criteria as specified above. This was achieved using the BiLateral IMPlant (BLIMP) testing system software to present all 24 lists to the participants via headphones, accompanied by 65 dBA masking noise. Participants were asked to repeat aloud as much of each sentence as they heard. In this phase of testing, responses were not scored by word, but based on whether the subject got the entire sentence correct. For example, participants could use either "a/the", "is/was", and "are/were" and still be scored correct. If the sentence was 'They saw a tui in the bush' and the subject answered 'They saw the tui in a bush' then this would still be considered correct. Appropriate substitutions are shown in Appendix 4. Participants were tested under three conditions, facilitated by the application of head related transfer functions (HRTFs): noise front (NF), with the noise and speech both at 0° azimuth; noise right (NR), where the noise is at 90° azimuth and the speech at 0°; and noise left (NL) with the noise at 270° azimuth and the speech at 0°. For four of the participants, all 24 lists were presented in the NF condition. For the remaining participants, eight lists were tested in each of three conditions, NF, NR and NL. List order was selected at random by the BLIMP system.

As the participants had normal hearing, the first sentence in the NF condition was presented in 65 dBA noise at -3 dB SNR, with the first sentence in the NR and NL conditions presented at -8 dB SNR. If the subject answered incorrectly, the same sentence was presented at a SNR 4 dB higher, until the correct answer was given. From then on, or if the subject answered correctly on the first presentation, the SNR of the following sentence was increased when the participants repeated the sentence correctly, and decreased when the answer was incorrect. The first four sentences were followed by a change in level (whether increase or decrease) of 4 dB. The step size for all subsequent presentations in each list was 2 dB. At the end of each list the SRT was calculated. The mean SRT for each list and the overall mean SRT across all participants and all (NF) lists can be seen in table 2. The mean SRT across all subjects and lists for the NR condition was -14.1 dB and for the NL condition was -14.3.

Table 2. Individual participant thresholds on all 24 NZHINT lists, with list means and overall mean (in dB).

	Participants													
List	A	B	C	D	E	F	G	H	I	J	K	L	Mean	SD
1	-6.3	-5.7	-6.8	-7.8	-5.8				-7.1		-5.2		-6.39	0.90
2	-6.7	-6.3	-8.1	-3.3	-6.5						-5.9		-6.13	1.58
3	-4.1	-5.7	-6.3	-5.7	-4.6		-5.2		-4.2				-5.11	0.84
4	-7.3	-7.5	-7.4	-7.4	-4.5				-5		-5.8		-6.41	1.29
5	-7.9	-6.1	-6.8	-5	-5.2		-7.1						-6.35	1.13
6	-6.3	-6.2	-5.1	-2.2	-4.2		-4.9		-6.6				-5.07	1.54
7	-6.3	-6.8	-5.8	-5.7	-6.9		-6.8				-6.5		-6.40	0.49
8	-4.6	-7.4	-7.3	-6.9	-5.9		-5.5				-5.3		-6.13	1.08
9	-7.9	-6.8	-4.5	-6.2	-9.2		-5.6		-5.9		-6.3		-6.55	1.45
10	-6.7	-7.9	-5.4	-6.3	-8.2		-7.1		-5.7		-5.2		-6.56	1.12
11	-5.7	-5.7	-5.7	-5.7	-6.9		-7.6		-6.5				-6.26	0.77
12	-5.1	-5.8	-5.8	-6.1	-5.7				-4.8		-5.5		-5.54	0.45
13	-5.5	-5.5	-6.2	-6.8		-7.1		-5.5		-4.3		-6.5	-5.93	0.91
14	-5.7	*	-5.1	-6.2				-2.4				-5.9	-5.06	1.54
15	-5.7	-4.1	-6.3	-4.7		-4.8				-5.5		-6	-5.30	0.79
16	-6.7	-6.7	-6.8	-7.5		-7.5		-7.2				-8.1	-7.21	0.52
17	-5.2	-5.7	-5.1	-4.6		-4.8		-5.9		-6.5			-5.40	0.67
18	-5.7	-5.2	-6.7	-6.1		-6.5				-6.5			-6.12	0.57
19	-6.2	-5.1	-5.2	-5.2		-5.9		-6.5				-5.3	-5.63	0.56
20	-7.8	-4.1	-6.8	-5.7						-5.8		-5.5	-5.95	1.25
21	-8.4	-5.7	-6.8	-6.3				-6.5				-7.5	-6.87	0.96
22	-5.5	-3.9	-5.1	-3.5		-6.5		-3.7		-6.4			-4.94	1.26
23	-6.8	-5.2	-4.4	-6.8		-8.2		-6		-6.6		-8.1	-6.51	1.31
24	-5.2	-6.1	-5.1	-6.2						-4.1			-5.34	0.86
												Mean	-5.97	1.1

*data not collected due to equipment error

The lists with the highest and lowest thresholds were combined to make a 20-sentence list. Next, the 10-sentence lists with the highest and lowest thresholds from the remaining lists were combined to make another 20-sentence list. This procedure was repeated until 12 20-sentence lists were formed. Three lists had to be matched

with lists outside this order in order to prevent repetition of vocabulary items within the same 20-sentence list. The phoneme distributions of these combined lists were calculated and compared with the overall distribution to ensure that none of the pairings had produced discrepant phoneme distributions. This process yielded the final 12 20-sentence phonemically matched lists of equal difficulty for use in adaptive HINT threshold measurements (see Appendix 5). The 20-sentence lists will subsequently be installed as a new language module in the HINT for Windows system.

2.3.7 Collecting preliminary norms

Completion of all previous steps was necessary in order to develop a valid speech-in-noise test, which was the primary aim of the study. The final step would have been to collect some preliminary normative data and determine the reliability and measurement error of the HINT. This step was unable to be completed because of time constraints.

3 Discussion

3.1 Features of the test

3.1.1 Scoring

The scoring system for the PI function and the three rounds of the EQ study required the tester to record how many words of each sentence the subject got right. Scoring was very strict, as the subject was only marked correct if they said exactly the same word as the one in the recording. Adding a plural "-s" or changing grammatical endings were counted as errors, but adding other words were not. This meant that for the sentence 'They listened to the news', the response 'They were listening to the news' was scored as 4 out of 5 words correct.

As sentences were chosen to have 5-7 syllables, this meant that sentences could have 3 -7 words. These sentences were then scored by word, not syllable. Later on in the process of list formation it is possible that a sentence could be judged correct or not on only two words (if the first word was an article). Therefore it may have preferable to have avoided 3-word sentences at the stage of collecting NZHINT materials. This consideration was not noted in the HINT Protocols or in published papers on HINT development.

3.1.2 Subject pronouns

This method of scoring and adjustment of the SNR based on these scores brought to light a number of issues. A number of speakers produced very short phonetic representations of the first subject pronoun or article, making them very difficult to hear and/or differentiate. The tester asked for clarification if unsure, but may have mis-recorded scores for some sentences due to 'swallowing' of the first word. The unstressed pronouns 'he', 'we' and 'they' at the beginning of the sentence were particularly hard to differentiate, both by the participants, scoring at only chance levels, and by the testers when recording the scores. It was agreed to allow any of these pronouns for the purposes of the collection of norms. The pronoun 'she' was generally able to be differentiated and was not usually substituted for other pronouns. This may be due to the increased intensity of the /ʃ/ compared to the /h/, /w/, and /ð/ which allowed it more salience against the masking noise.

As the first word of each sentence was generally very difficult to hear, several participants asked why an alerting beep was not used. Other tests such as the LISN-S have employed a beep or other such signal that can be heard above the noise in order

to signal the start of the sentence (Cameron & Dillon, 2007). HEI's protocol has the noise starting 0.5 seconds before the speech in all sentences. It was decided that this was sufficiently predictable to act as a cue for the subject and since comparability with other versions of the HINT was highly desirable, it was decided not to add beeps to the New Zealand version.

3.1.3 Consequential error

Often, lower scores were the result of consequential error from both grammatical agreement and pronoun/possessive agreement. This was due to the participant mishearing one word in the phrase or sentence which caused them to choose the wrong word, or the wrong form of the right word, further on in the sentence. For the sentence 'People had seen the horses', the subject's response and score depended on what the subject heard as the first word of the verb phrase. Responses including 'People could see/can see/were seeing the horses' scored 3 points, as the auxiliary of the verb phrase demanded a form other than 'seen' for the main verb. Consequential error also occurred with pronouns. In the sentence 'We brushed sand off our jeans', there could be essentially two marks for guessing the first pronoun correctly, as this would dictate the possessive pronoun used later on. However, participants would often use information in the latter half of the sentence to predict earlier information, and change their answer, e.g. "He took three bottles….no, WE took three bottles with us."

3.1.4 Vowels and consonants

One particular feature of this test was the salience of the vowels over the consonants. It was very common for participants to give a response that included all the correct stressed vowels but with different combinations of consonants. Some examples include: 'number four five' (The oven caught fire); 'she boarded her flight' (They recorded her height); 'they caught fire from the friction (They caught five little fish); 'an alright system' (They looked like sisters). However, there were also sentences where consonants had a greater influence over responses than did the vowels. Some examples of this include: 'the washing machine beeped/bleeped/leaked' (The washing machine was em**pt**y); 'she bought a toothpick' (She bought a **tick**et); 'she walked to hospital' (She walked to the bu**s st**o**p**). Some responses followed a pattern which was difficult to comprehend, such as 'they saw a penguin' (They saw a rainbow).

3.2 Rater reliability

Reliability measures were performed in Round 3. This consisted of an experienced linguist sitting in with the tester and scoring for 8 of the 11 participants. Inter-rater reliability ranged from 95% to 99% agreement when scored by sentence, with an average of 96.8%. The two scorers had to score the whole sentence correctly to be considered in agreement. Most disagreements stemmed from the difficulty in determining the presence or absence of the initial article or subject pronoun. Errors due to miscounting were very rare.

3.3 Possible reasons for variability

The first few rounds of testing were characterised by large inter-subject variability. Although all participants had what is classified as 'normal hearing', there was still some variation in puretone thresholds. This variation may have accounted for some of the differences in intelligibility scores. However, one of the participants in Round 2 who performed significantly below average had most thresholds better than 0 dB HL in the 250-8000 Hz. There was no information sought from the participants about the duration, frequency, or recency of their exposure to loud noise. There is some evidence that otoacoustic emissions (OAEs) are able to detect noise-induced cochlear damage before this affects puretone thresholds (Attias et al., 1995). Screening with OAEs may have given more information on the status of the cochlea.

Another factor contributing to speech intelligibility in noise is ultra high frequency thresholds. In traditional audiometry, hearing acuity is tested up to 8 kHz. However, it has been shown that ultra high frequency acuity can influence speech perception (Best, Carlile, Jin, & van Schaik, 2005). Perhaps testing hearing over 8 kHz may have given another indicator of speech perception in noise. Unfortunately, testing of high frequencies requires special equipment and careful calibration, which would not have been practical or feasible in this project.

It was considered that certain experience or attributes may have affected the performance of some participants. The possibility was discussed that musicians may have had a natural advantage through practice at stream segregation with different instruments and generally being more attuned to aural stimuli. This theory did not appear to be borne out by the data, as while one musician performed above average, two others who reported extensive experience with music were at the average-to-low end of the data range. Not all participants were asked about their musical history so few conclusions can be drawn from this. One subject in Round 3 who performed well above average was a speech language therapy student who had had many hours of

experience in transcribing interviews being used for linguistics research. It is likely that the extra practice this student had in transcribing speech over background noise and/or other speakers improved her performance on this task. However, it may also be the case that her above-average understanding in noise may have suited her to the role of transcribing interviews, rather than vice versa.

All participants were asked in the patient questionnaire (required by HEI) whether they considered they had difficulty understanding speech in noisy environments on a scale of 1-7. While this question was somewhat ambiguous (how noisy is noisy?), it was generally understood as a scale comparing their difficulty to how much difficulty they considered the average person would have in comprehending speech in noise. It was noted that only four out of 49 participants said they had any difficulty hearing in noise. Three participants scored themselves as a 2 out of 7, while one suggested they were a 4 out of 7. None of these participants did poorly in the testing. In contrast, none of the participants who scored below average indicated that they had difficulties hearing in noise. This suggests that the perception of one's hearing ability can be different to the reality (see *3.7 Directions for further research*).

The motivation of the participants may have also been a factor. Participants were tested at different times of the day in order to fit into their schedules, and therefore some may have been tested when tired or hungry or otherwise not in an optimum physical or mental state. In an attempt to combat this, short breaks of 1-2 minutes were offered between lists, where participants asked questions and talked with the tester(s). In Rounds 2 and 3, a longer break of 15 minutes (including a hot drink and something to eat) was taken when participants were approximately half way through. While this appeared to keep performance stable or even improved performance, it may not have been sufficient to counteract fatigue for some participants.

Different patterns of guessing may have affected results. Some participants were less inclined to guess, and therefore needed more encouragement to do so. Some participants guessed single words, whereas others tried to produce full sentences, which gave a higher probability that they would produce grammatical words such as 'a', 'the', 'in', 'on', and 'at' and achieve at least one point for that sentence. On the advice of consultants at HEI, the data from four participants were not included because of their reluctance to offer a response when they were not at least moderately sure. While the practice of excluding data that does not conform is generally frowned upon, this decision was made in order to streamline the equalisation process and reduce the need for many more rounds of testing. By not guessing, or only guessing

one or two words, the system of scoring by word and then adjusting until 65% intelligibility was reached would have been undermined.

There appears to be no linguistic reason for why some sentences were more difficult to identify correctly than others. The sentences that were dropped were from the beginning, middle and end of the original set of 502 sentences, suggesting that it was not merely because the sentences in the latter half were more or less intelligible than the earlier half. It is simply impossible to predict which sentences are more difficult to hear in noise, and this is why such a large pool of sentences is generated at the beginning.

3.4 Final set of NZHINT sentences

During the equalisation stage and its three rounds of testing, many sentences were eliminated, leaving only 244 sentences that met the criteria to be placed into phonetically matched lists. This process required 240 sentences, leaving very little room to manoeuvre if by this stage there were sentences that were no longer deemed appropriate, or were very similar in vocabulary to other sentences. The most obvious example is the pair 'the man was very strong' and 'the girl was very strong'. These were included as a matter of interest to see whether one would be easier/more intelligible than the other, due to assumptions about gender stereotypes. There were three sentences with 'washing' in them: 'the washing dried fast in the wind', 'the washing machine is empty' and 'sheets hang on the washing line'. There were also similar constructions such as 'the hot concrete burned her feet' and 'the hot soup burned her tongue'. It was assumed that there would be a larger pool of sentences to choose from when it came time to form the lists, ensuring that only one of these sentences would have to be included. However, given the lack of flexibility due to the small number of sentences that remained, all these sentences had to be retained. In light of this, special care was taken not to have two similar sentences in the same 20-sentence list.

3.5 Phonetic matching

The HEI protocol followed in developing the NZHINT required the 24 lists to be approximately matched in their phonemic distribution. Phonetically balanced lists have equivalent phonological composition which is representative of connected English discourse. This is a popular feature of speech tests, as it has good face validity. As the aim of speech testing is to test patients with the sorts of stimuli they would hear day to day, it would defeat the purpose to have lists with multiple

instances of /ŋ/ and /θ/ and very few /t/ and /n/. The HINT lists do not attempt full phonemic balance but approximate phonemic balance by arranging the sentences so that each list shows the same phonemic distribution as the entire set of sentences as a whole. This may introduce some distortion depending on the sentences that survive the equalisation process: the remaining sentences may not have the same phonemic distribution as spoken English discourse. For example, it is likely that there are fewer /ŋ/ phonemes in this set of sentences than there would be in a random sample of spoken English as a number of sentences containing this phoneme were discarded before equalisation testing began. However, it has been convincingly suggested that creating phonemically balanced lists may not actually be necessary or clinically relevant, as shown by Martin, Champlin and Perez (2000).

3.6 Limitations

The main limitation on this study was that norms and reliability tests were unable to be completed. Until these data have been collected, the NZHINT should not be used clinically. Although the collection of norms was intended from the beginning, a number of delays meant that there was insufficient time for this to be done. These delays were exacerbated by staff turnover at HEI, leaving much of the processing to be done by the director himself, who had many other demands on his time and spent many weeks of the year travelling.

There appeared to be more variability in the NZHINT data compared to the variability reported in the development of the HINT in other languages. From the PI function onwards, intra-subject variability was high. This was concerning, as it is unlikely that NZE is somehow harder or more variably intelligible than other languages. One way of dealing with this would have been to test more participants in order to neutralise some of this variability, and to continue with further rounds of equalisation testing until more stable average percent correct scores had been achieved. As a result of this, a wider range of average percent correct scores (i.e. 70% ±20% rather than ±15%) was allowed. Without relaxing this criterion, the NZHINT could not have been completed with 12 full lists within the time constraints of the thesis, as there would have been too few sentences (200, rather than the 240 required).

As indicated in the HINT protocols, calibration of the soundcard/headphone pair was done at a high level (110dB) and four participants were tested with NF. Following discussions with contacts at HEI, concern was expressed about the high calibration level, as the presentations at such low SNR during testing may have been too close to

the noise floor. For example, from the NF thresholds, the NR/NL could be conservatively predicted as being recorded at -10 dB S/N. In this case, the speech presentation level would be 55 dB below the maximum output capability of the soundcard/amp system (i.e. at 55 dB SPL). As most soundcards have a 16-bit, 96 dB dynamic range, this level would be at 41 dB above the minimum output capability of the soundcard. The sentences have an instantaneous dynamic range of +/- 15 dB about their RMS value, meaning that the lowest sentence samples would only be 25 dB above the lowest output level of the soundcard. This would bring the lowest required output level very close to the noise floor.

The soundcard was then recalibrated to 90 dBSPL (see *2.2 Equipment*). A further eight participants were tested in all three conditions, following suggestions from HEI that it would be useful to have some pre-norming data on the most appropriate SNR to start with for the NL/NR conditions. After further consultation, the sentences were checked at the higher calibration in quiet at 50 dB (as the average thresholds for the NL/NR conditions were approximately -14 dB SNR). Since there was no audible hiss, and the scores from both sets of data appeared to be very close, the results from all participants were combined. Although unlikely to have had any significant effect on thresholds, it is possible that further variability was introduced through the use of two different calibrations.

3.7 Directions for future research

The first step towards building on this research will involve establishing norms and reliability for the NZHINT for NH listeners under headphones and in the soundfield. Subsequent to this, norms should be collected for listeners with SNHL of varying degrees. From this point, the HINT could be used in research looking at differences in speech understanding in noise within and between many different populations, such as older vs younger listeners. Future research may show if the NZHINT sentences are suitable for testing children, or provide a basis for the development and standardisation of a test that is specific to a pediatric population.

Another interesting line of investigation would be to test whether in fact NZE speakers perform better on the NZHINT than they do on the original HINT. The intention was to do this at the same time as gathering the norms for the NZHINT, but this was unable to be completed. If significant differences were noted on performance on the NZHINT and the AmHINT then this gives further validity to the need for development of New Zealand-specific speech materials. However, even if clinically significant population differences are not observed, a subset of individuals for whom

there is a significant difference may be highlighted, and thus provide more information on the variables that would influence NZE speakers' performance on both tests. If few differences are noted, this will in no way diminish the face validity of having NZE materials for use in testing adult NZE speakers. It may simply lessen the urgency with which further NZE speech materials are developed.

Further direction for study could include the use of the NZHINT materials in measures of hearing aid benefit. Tools such as the Performance-Perceptual Test (PPT; Saunders & Cienkowski, 2002) can be used to gain further insight into hearing aid use, benefit and satisfaction. The PPT is an outcomes measurement tool which consists of an objective (performance) and subjective (perceptual) component, both using the same test format, materials, and unit of measurement. Its purpose is to measure any differences between subjective and objective assessment of hearing ability and hearing aid benefit. The Performance Speech Reception Threshold in Noise (SRTN) is measured by finding the listener's SNR for 50% correct speech in noise using the HINT sentences and adaptive protocol. The Perceptual SRTN is measured with the HINT materials and a SNR for 50% correct speech in noise is found, but it is calculated as the SNR at which the listener perceives that they can just understand all of the speech material. The difference between these two measures is called the Performance-Perceptual Discrepancy (PP-DIS) and can indicate if the listener is accurately, over- or underestimating their hearing ability.

Saunders, Forsline and Fausti (2004) found that the majority of their 33 NH and 74 HI participants had a small PP-DIS, suggesting that they were quite accurate in estimating their own hearing ability. In the same study, those who underestimated their ability to hear reported more handicap on the Hearing Handicap Inventory for the Elderly (HHIE)/ Hearing Handicap Inventory for Adults (HHIA), an outcome measure of hearing disability. Those who overestimated their ability reported less handicap on the HHIE/HHIA. While this may seem obvious, the PPT may assist in identifying those individuals who appear to gain less perceived benefit from their hearing aids in spite of good performance. The results of the PPT could be used as a counselling tool to encourage more realistic perception of hearing ability and adjust expectations of benefit from the use of amplification.

A further use of the NZHINT materials could be in conjunction with other tests that attempt to predict likely success with hearing aids by looking at other factors apart from speech intelligibility. The Acceptance of Noise Test (Nabelek, Tampas, & Burchfield, 2004) measures what level of background noise a listener is prepared to tolerate while following the words of a pre-recorded story. The difference between the

story (adjusted to the listener's most comfortable listening level) and the highest allowable background noise level is the listener's allowable SNR, now termed the acceptable noise level (ANL). Nabelek, Tucker and Letowski (1991) found that ANLs were related to the use of hearing aids: fulltime hearing aid users had significantly higher ANLs than did those who wore hearing aids part-time or who had rejected hearing aids. While no correlation was found between performance on the SPIN and ANL scores, it is an area of interest that may see measures of speech intelligibility and other measures combined to form a test battery that will predict likely benefit and satisfaction from hearing aids. It may also help to understand why some clients, after trials of hearing aids, do not seem to obtain benefit.

3.8 Summary

The aim of this study was to develop a version of the Hearing in Noise Test that was appropriate for clinical testing of a New Zealand population. Sentence materials using common New Zealand vocabulary were developed and evaluated for naturalness by native speakers. These sentences were adjusted to be approximately equally intelligible and formed into phonemically matched lists. The twelve lists were incorporated into the adaptive HINT software for clinical use. Norms and reliability measures were unable to be completed due to time constraints. However, once these data have been collected, the NZHINT will be a valuable clinical and research tool for use in testing NZE speakers.

Appendix 1