Refinement and Normalisation of the University of Canterbury Auditory-Visual Matrix Sentence Test

---

A thesis submitted in partial fulfilment of the

requirements for the Degree

of Master of Audiology

in the University of Canterbury

by Amber D. McClelland

University of Canterbury

2015

Abstract

Developed by O'Beirne and Trounson (Trounson, 2012), the UC Auditory-Visual Matrix Sentence Test (UCAMST) is an auditory-visual speech test in NZ English where sentences are assembled from 50 words arranged into 5 columns (name, verb, quantity, adjective, object). Generation of sentence materials involved cutting and re-assembling 100 naturally spoken "original" sentences to create a large repertoire of 100,000 unique "synthesised" sentences.

The process of synthesising sentences from video fragments resulted in occasional artifactual image jerks ("judders")—quantified by an unusually large change in the "pixel difference value" of consecutive frames—at the edited transitions between video fragments. To preserve the naturalness of materials, Study 1 aimed to select transitions with the least "noticeable" judders.

Normal-hearing participants (n = 18) assigned a 10-point noticeability rating score to 100 sentences comprising unedited "no judder" sentences (n = 28), and "synthesised" sentences (n = 72) that varied in the severity (i.e. pixel difference value), number, and position of judders. The judders were found to be significantly noticeable compared to no judder controls, and based on mean rating score, 2,494 sentences with "minimal noticeable judder" were included

in the auditory-visual UCAMST. Follow-on work should establish equivalent lists using these sentences. The average pixel difference value was found to be a significant predictor of rating score, therefore may be used as a guide in future development of auditory-visual speech tests assembled from video fragments.

The aim of Study 2 was to normalise the auditory-alone UCAMST to make each audio fragment equally intelligible in noise. In Part I, individuals with normal hearing (n = 17) assessed 400 sentences containing each file fragment presented at four different SNRs (-18.5, -15, -11.5, and -8 dB) in both constant speech-shaped noise (n = 9) and six-talker babble (n = 8). An intelligibility function was fitted to word-specific data, and the midpoint (L_{mid}, intelligibility at 50%) of each function was adjusted to equal the mean pre-normalisation midpoint across fragments. In Part II, 30 lists of 20 sentences were generated with relatively homogeneous frequency of matrix word use. The predicted parameters in constant noise (L_{mid} = -14.0 dB SNR; slope = 13.9%/dB ± 0.0%/dB) are comparable with published equivalents. The babble noise condition was, conversely, less sensitive (L_{mid} = -14.9 dB SNR; slope = 10.3%/dB ± 0.1%/dB), possibly due to a smaller sample size (n = 8). Overall, this research constituted an important first step in establishing the UCAMST as a reliable measure of speech recognition; follow-on work will validate the normalisation procedure carried out in this project.

Acknowledgments

I would like to express sincere appreciation to my supervisor, Associate Professor Greg O'Beirne, for his guidance and expertise. I also thank my co-supervisor Dr Don Sinex for his valued support and input during the writing phase.

I would like to thank the participants for assisting with my research. Without you, this would not have been possible.

Thank you to the staff of the Audiology department at UC for sharing their talents. I thoroughly enjoyed my time on this course.

Thank you to my awesome class mates who made me feel very welcome in Christchurch. I have no doubt that you will all make excellent audiologists and I hope to work with you in the future.

Thank you to my mother, Denise, and father, Artie, for their love and support over 7 years of tertiary education. I also thank my brother Liam for his input and "brotherly wisdom", and my friends for all the Skype dates and long text conversations while I was in Christchurch.

Last but not least I thank Laura for her unwavering support, patience, and warmth throughout my post-graduate career.

Table of Contents

List of Figures

Abbreviations

BM	basilar membrane
CVC	consonant-vowel-consonant
dB	decibel
dB HL	decibels hearing level
dB SPL	decibels sound pressure level
HINT	Hearing In Noise Test
Hz	Hertz
IHC	inner hair cell
kHz	kilohertz
LTSS	long term speech spectrum
NZ	New Zealand
NZHINT	New Zealand Hearing In Noise Test
OHC	outer hair cell
PI	performance-intensity
RGB	red green blue
SNHL	sensorineural hearing loss
SRT	speech reception threshold
WHO	World Health Organisation

Chapter 1

Introduction

1.1 Background

1.1.1 Hearing impairment in New Zealand (NZ)

Hearing impairment disrupts oral communication and presents numerous difficulties on a daily basis. Conversations are fatiguing, as greater effort is required to understand speech, and uncertainty regarding subject matter may lead to social withdrawal due to diminished confidence (Arlinger, 2003). Hearing impairment has been shown to negatively impact quality of life (Dalton et al., 2003). More concerning is the effect on mental health: adults with hearing impairment are more likely to experience depression—particularly women and individuals under 70 years of age (Li et al., 2014). Not only does hearing impairment affect the individual, but also their loved ones. Scarinci, Worrall, and Hickson (2008) found that spouses of affected individuals reported feeling exhausted at having to provide numerous repetitions. In some cases, the spouses reported fewer attempts at initiating conversation.

However, for a number of reasons, a hearing impairment will often go undiagnosed (Dalton et al., 2003). First of all, it is not part of usual practice for

primary care professionals (i.e. general practitioners) to perform hearing checks. Most often, affected individuals complain of an inability to understand speech in noisy environments (Hochmuth et al., 2012). As a consequence, in quiet situations, such as the doctor's office, hearing impairment tends to go unnoticed. In the elderly population, a hearing impairment may be acknowledged, but is secondary to other more serious ailments (Newman & Sandridge, 2004). In other cases, the individual acknowledges their hearing impairment, but does not seek treatment (Dalton et al., 2003). In 2005, approximately 10.3% of the NZ population reported a hearing impairment; however, only 29% of this group reported the use of hearing aids (Greville, 2005). The reluctance to seek treatment may be the product of a passive attitude towards healthcare; for example, the perception that hearing impairment is an inescapable consequence of ageing (Dalton et al., 2003). However, affected individuals have also cited financial constraints (64%) and the perceived stigma of hearing aid use (48%) as reasons for not adopting hearing aids (Kochkin, 2007). Despite these barriers, the diagnosis and treatment of hearing impairment has pervasive benefit to quality of life and relationships. Hearing aids were found to alleviate symptoms of depression and improve cognitive functioning in the elderly (Acar, Yurekli, Babademez, Karabulut, & Karasen, 2011). Furthermore, the negative impact on the spousal relationship was reduced if the affected spouse accepted their hearing impairment (Scarinci et al., 2008). These outcomes illustrate the value of

diagnosing and treating hearing loss. A reliable measure of speech recognition is crucial to this process.

1.1.2 The structure of this project

The University of Canterbury Auditory-visual Matrix Sentence Test (UCAMST) was developed by Trounson and O'Beirne (Trounson, 2012) to provide an assessment of speech recognition with an extensive repertoire of test sentences. The UCAMST allows a number of test conditions through choice of masking noise (i.e. six-talker babble, constant speech-shaped noise, and quiet), set presentation (open vs. closed) and presentation modality (i.e. auditory-alone, visual-alone, or auditory-visual) for different diagnostic and rehabilitative needs. The aim of the current project was to further the development of the UCAMST with two necessary studies. Study 1 investigated how noticeable video "judders"—artifactual image jerks at the edited transitions between video file fragments—were with the use of a subjective rating scale. The relationship between this rating score and the "pixel difference value"—an objective measure describing the change in head position across each transition—was also probed. Combined, this data informed selection of sentences with the least noticeable judders for inclusion in the auditory-visual version of the UCAMST. Study 2 normalised (i.e. equalised the difficulty of) matrix words in the auditory-alone condition, from

which lists of equivalent sentences were selected. Study 2 was carried out in two types of masking noise, constant speech-shaped noise and six-talker babble, and was an important step towards establishing the UCAMST as a reliable audiological test.

1.2 The anatomy of hearing

Before discussing the importance of speech recognition testing, we will first examine the auditory anatomy and physiology fundamental to understanding speech. Numerous key anatomical features of the peripheral auditory system are involved in the perception of sound. The ear is divisible into three anatomical segments: outer, middle, and inner. The outer ear comprises the pinna (the visible ear); the external auditory meatus, or ear canal; and the outer layer of the tympanic membrane, or eardrum. The middle ear consists of the tympanic cavity and ossicles, as well as the inner layer of the tympanic membrane. “Ossicles” is a collective term for the three bones individually referred to as the malleus, incus, and stapes (Donkelaar & Kaga, 2011) . The inner ear consists of the vestibular system and the cochlea or organ of hearing, which are responsible for balance and hearing, respectively (Ko, 2010). Within the cochlea is the organ of Corti, which sits on the basilar membrane (BM) and possesses sensory hair cells crucial to the perception of

sound (Donkelaar & Kaga, 2011; Pickles, 2012). Cochlear hair cells and their function will be discussed in further detail later in this section.

Each anatomical segment of the ear has an important role in audition. Sound, a pressure wave propagating through the air, will first encounter the listener's pinna (Pickles, 2012). The two main functions of the pinna are 1) to maximise sound pressure at the tympanic membrane using the resonant properties of the pinna, in particular the concha bowl, and 2) to provide cues for localising sound sources (Pickles, 2012). The sound wave is channelled down the ear canal by the pinna, which causes vibration of the tympanic membrane and ossicles (Patuzzi, 2009; Pickles, 2012). The head of the malleus connects to the tympanic membrane (Donkelaar & Kaga, 2011), whereas the stapes connects to the oval window of the cochlea (Pickles, 2012). The main function of the middle ear is to match the impedances of the air medium of the ear canal to the fluid medium inside the cochlea (Puria, Fay, & Popper, 2013).To achieve this, ossicular vibration—in particular, the "piston-like" action of the stapes—generates pressure waves in the cochlear fluids that travel from the stapes in an apical (i.e. towards the cochlear apex) direction (Gates & Mills, 2005; Patuzzi, 2009). If the input is a sinusoidal pure tone, a displacement wave forms on the BM that reaches maximum amplitude at the location of peak resonance before subsiding sharply (Patuzzi, 2009). The arrangement of the BM is tonotopic: low and high frequency inputs peak at the

apex and base of the cochlea, respectively. The location of peak resonance is referred to as the characteristic frequency (Patuzzi, 2009; Pickles, 2012).

What has been described thus far is the passive mechanism by which an incoming sound wave vibrates the BM; however, this passive mechanism is not sufficient for the perception of low-level sounds, particularly at high frequencies (Patuzzi, 2009). From this point onwards, the transduction of sound is dependent on two classes of hair cell located in the organ of Corti: outer hair cells (OHCs) and inner hair cells (IHCs) (Pickles, 2012). A healthy cochlea houses approximately 11,000 OHCs organised in rows of three or four, and 3,500 IHCs organised in single rows (Ashmore, 2008). The stereocilia of the OHCs are entrenched in the tectorial membrane—a large, "jelly-like" roof over the organ of Corti—whereas the IHCs are believed not to make contact with this membrane (Pickles, 2012). The role of the OHCs is to amplify the vibrations of the BM, reducing friction. This is achieved with an active process involving cyclical contraction of the motor protein "prestin", which is found in the baso-lateral wall of the OHCs (Donkelaar & Kaga, 2011; Patuzzi, 2009; Pickles, 2012). The IHCs sense the vibrations of the OHCs and release neurotransmitters to stimulate afferent nerves, which send signals to the brain for higher level processing (Gates & Mills, 2005; Patuzzi, 2009). If auditory function is preserved, the result is the perception of sound.

1.3 The anatomy of hearing loss

Abnormalities in the peripheral auditory system may cause hearing loss, the degree and origin of which can be ascertained by audiological testing. Adult hearing thresholds are typically determined with pure tone audiometry. This test involves presenting pure tones to a client who is instructed to respond (e.g., with a button press) when they hear a tone (Katz, 2009). The frequencies presented encompass those most important for understanding speech, usually 250 to 8000 Hz. An audiogram provides a graphical representation of a client's hearing sensitivity, with the threshold in dB HL (decibels hearing level) depicted as a function of frequency in kilohertz (kHz) or hertz (Hz). Normal hearing is defined by the World Health Organisation (WHO) as a threshold of 25 dB HL or less. Conversely, hearing impairment is denoted by a threshold of 26 dB HL or greater, and is graded by severity from mild to profound (Mathers, Smith, & Concha, 2000).

Based on the origin of the hearing loss, it is termed either "conductive" or "sensorineural". A third term, "mixed", is used if a hearing loss is comprises both conductive and sensorineural components (Patuzzi, 2009). A conductive hearing loss typically results from an irregularity in the outer or middle portions of the ear that disrupts the transmission of sound to the cochlea (Pickles, 2012). By comparison, sensorineural hearing loss (SNHL) originates from damage to the cochlea or the auditory nerve (Pickles, 2012).

This is the most common type of hearing loss in adults and is typically permanent (Newman & Sandridge, 2004). A common type of SNHL, known as "presbycusis", is age-related, and may first be first noticed at approximately 65 to 75 years of age. Characteristically, presbycusis first manifests as a high-frequency hearing loss—the result of hair cell loss at the basal end of the cochlea (Donkelaar & Kaga, 2011). With time, the hearing loss spreads to encompass the lower frequencies (Gates & Mills, 2005), causing the audiogram to flatten.

SNHL can be further subcategorised by origin. For example, the presbycusis described above would be termed a "cochlear" SNHL. A SNHL may also be defined as "retrocochlear" (i.e. beyond the cochlea) in origin (Patuzzi, 2009). A cochlear hearing loss arises from disruption to "motor" or "sensory" processes. The former relates to OHC function and the active process, and the latter relates to IHC function (Patuzzi, 2009). The effect of a motor hearing loss is twofold: a decrease in hearing sensitivity and a decrease in frequency specificity. The loss of hearing sensitivity is limited to approximately 60 dB HL as the active process only contributes to BM stimulation with low- to mid- intensity sounds. For high-intensity sounds, the cochlea is passive: the spatial selectivity of the BM is less sharp, and hence, a large area of the BM may be stimulated by a pure tone (Patuzzi, 2009; Pickles, 2012). Consequently, damage to the OHCs and lack of active process means

the cochlea is reliant on passive stimulation. The result is the irreversible loss of sharp tuning, and therefore, a loss of frequency specificity (Patuzzi, 2009).

The loss of frequency specificity can be defined through Plomp's framework of hearing loss as having both "attenuation" and "distortion" components (Plomp, 1978). Attenuation corresponds to an increase in thresholds (i.e. decreased hearing sensitivity), while distortion corresponds to an impairment in the ability to understand speech (i.e. loss of the clarity of speech). The frequency specificity provided by the OHCs is essential for the intelligibility of speech (Patuzzi, 2009). The loss of this function has profound implications for rehabilitation, such as whether a client will benefit from amplification. Speech audiometry is, therefore, a unique and crucial part of the audiological test battery as it assesses the "distortion" component of hearing loss (Plomp, 1978).

1.4 Speech audiometry in NZ

In NZ clinics, speech audiometry is typically carried out using the Consonant-Vowel-Consonant (CVC) meaningful word lists. Each list consists of 10 monosyllabic (single syllable) words comprising the same 10 first consonants, 10 vowels, and 10 final consonants (Boothroyd & Nittrouer, 1988). Three different lists are presented auditory-alone in quiet (i.e. with no accompanying masking noise), each at a different presentation level (dB HL).

The test adopts phoneme scoring, which is based on the correct repetition of vowels and consonants as opposed to whole words (Boothroyd, 2008). The percentage correct (%) scored at each presentation level is plotted as a Performance-Intensity (PI) function (Katz, 2009). The listener's Speech Reception Threshold (SRT) can be derived from this function: the presentation level at which the listener scores 50% correct (Boothroyd, 2008). Additionally, one may obtain the PB_{max}, which denotes the best possible score that a client can achieve (Katz, 2009).

The PI function, SRT, and PB_{max} have a number of uses in diagnostic audiology. The PI function is usually compared to a normative curve to gauge the client's level of performance relative to individuals with normal hearing. Furthermore, the SRT provides a value by which the pure tone thresholds can be cross-checked (Boothroyd, 2008; Mendel, 2008). Nonetheless, an experienced clinician can often predict speech recognition performance based on the severity of hearing loss on the audiogram (Gulya, Glasscock, Minor, & Poe, 2010). If the speech results and pure tone thresholds are inconsistent, a non-organic hearing loss (i.e. false hearing loss) may be indicated. However, poor speech performance, beyond what would be predicted given the audiogram, may be a "red flag" for a vestibular schwannoma (Gulya et al., 2010) or auditory neuropathy (Starr, Picton, Sininger, Hood, & Berlin, 1996). The morphology of the PI function may be consistent with the origin of the hearing loss. A PI function with the typical shape but shifted to an area of

increased presentation level is consistent with a conductive hearing loss, as increasing stimulus loudness overcomes the conductive component. A client with a high-frequency hearing loss may notice reduction in the clarity of consonants, which can result in word recognition errors (Garstecki & Erler, 2009). In this case, speech audiometry may exhibit a pattern of consonant confusion or omission. On the other hand, a client with a cochlear impairment may not be able to perceive speech sounds—regardless of intensity—due to poor frequency discrimination (Walsh, 1953). This would manifest as a maximum score of less than 100% correct (i.e. a low PB_{max}) at high intensity levels. Based on such outcomes, speech audiometry is pivotal to rehabilitation, and can be used to guide hearing aid fitting. Continuing the example above, if poor frequency discrimination in spite of presentation level is noted, amplification may not improve the client's speech intelligibility, and may instead be detrimental to comfort.

Overall, the current practice of speech audiometry in NZ is a valuable component of the audiological test battery in terms of diagnosing hearing impairment and informing client rehabilitation. Despite the pervasive use of CVC word lists in NZ, this test is disadvantaged by 1) the lack of masking noise, 2) the use of monosyllabic words, and 3) the limited amount of materials available for testing. Each of these disadvantages will be discussed over the following sections.

1.5 Speech testing in noise

1.5.1 Psychophysical parameters

Similar to speech tests in quiet, performance is typically indicated by the listener's SRT: the presentation level that results in a performance of 50% intelligibility (Brand & Kollmeier, 2002). In noise, however, the SRT is derived from a psychometric function that represents performance—the proportion of correct responses, or intelligibility (%)—as a function of the signal-to-noise ratio (SNR). Psychometric functions of this nature are typically sigmoid ('s'-shaped), as the probability of a correct response monotonically increases with the intensity of the stimulus (Gilchrist, Jerwood, & Ismaiel, 2005). Figure 1 exemplifies the typical morphology of a psychometric function.

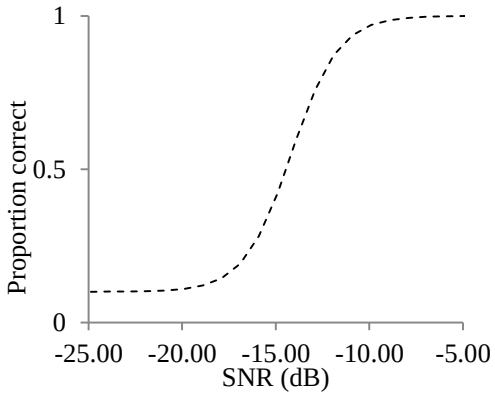

Figure 1. Typical sigmoid shape associated with psychometric functions measuring proportion of correct responses, or p(c), against SNR (dB).

Regarding speech-in-noise testing, the accuracy of the SRT is denoted by the slope of the psychometric function at the point of the SRT (Ozimek, Warzybok, & Kutzner, 2010). The slope determines the "sensitivity" of the test: it represents the percentage increase in intelligibility for every 1 dB increase in the SNR (%/dB). A highly sensitive test will see a small change in stimulus value (i.e. dB SNR) yield a large change in the measured value (i.e. % correct response) (Brand & Kollmeier, 2002). Figure 2 compares the morphology of psychometric functions with steep and shallow slopes.

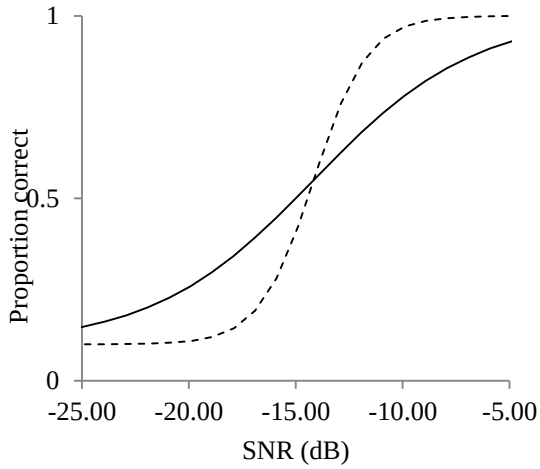

Figure 2. Comparison of intelligibility function with steep (dashed line) and shallow (solid line) slopes.

Tests with a high degree of sensitivity are desirable as they better discriminate between variables of interest, and also allow for greater precision in determining SRT within a smaller number of trials (Francart, 2011). In short,

the slope of the psychometric function is related to the reliability and efficiency of SRT measurement. We will return to the concept of reliability later in relation to the normalisation process carried out in Study 2.

1.5.2 Advantages of masking noise

Speech-in-noise tests are increasingly necessitated by evidence-based audiological practice for a number of reasons. The presentation of stimuli in quiet, as is typically practised in NZ, does not assess a client's ability to understand speech in everyday situations where speech signals are usually masked by background noises (i.e. air-conditioning, traffic, and other talkers). A further advantage, the use of masking noise also improves the sensitivity of a speech test. In an experimental study, McArdle, Wilson, and Burks (2005) evinced that presenting speech at various SNRs in multi-talker babble produced a steeper psychometric function slope than when speech materials were presented in quiet. Compared to in quiet, the use of masking noise better distinguished the hearing-impaired group from the normal-hearing group, with a separation in the group SRTs of approximately 8 dB. The result is perhaps unsurprising, as difficulty understanding speech in noisy environments is a frequent complaint of individuals with hearing impairment (Hochmuth et al., 2012). This difficulty may be the consequence of poor frequency discrimination which, as discussed previously, is commonly associated with a

cochlear hearing loss (Patuzzi, 2009). Frequency information (e.g., formant frequency) is crucial in distinguishing the requisite signal in situations with multiple talkers (Darwin & Hukin, 2000). Compounded with a lower hearing sensitivity, noisy environments present a challenge for individuals with hearing impairment.

In a similar vein, speech-in-noise tests can also identify pathologies associated with impaired temporal processing. Poor speech discrimination in quiet—despite a normal pure tone audiogram—may support the diagnosis of auditory neuropathy (Starr et al., 1996). However, performance will typically deteriorate further in noise (Zeng & Liu, 2006). Similarly, older adults who perform well in quiet, but poorly in noise, may have what is referred to as "central presbycusis" (Gates & Mills, 2005). The specific temporal deficit is theorised to be an inability to utilise drops in background noise to enhance speech recognition (Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006). This is relevant when selecting the type of masking noise (to be discussed in section 1.5.3). In summary, masking noise would alert the clinician to deficits in both frequency specificity and temporal processing.

Speech-in-noise tests are also useful for rehabilitation, such as when assessing amplification benefit. With development of digital hearing aids came new features designed to improve speech recognition in noise; for example, noise cancellation algorithms (Katz, 2009). For this reason, speech-in-noise tests are useful in counselling clients, as they may indicate which technology

provides the largest benefit to intelligibility in noisy environments (Wilson, McArdle, & Smith, 2007). If a speech-in-noise test is sound field capable, it may have utility in determining whether hearing aid technology is beneficial to the client.

1.5.3 Selection of masking noise

The selection of a masking noise has been a contentious issue in the literature. One viewpoint posits that everyday masking noise is typically speech; therefore, a masking noise that simulates this (i.e. multi-talker babble) has higher face validity (Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004; Plomp, 1978). However, a disadvantage of babble is that it causes fluctuations in SNR (Killion et al., 2004), which may affect the accuracy of the SRT measurement. An alternative to babble is a masking noise that has the same spectral content as the target signal (i.e. a "speech-shaped" noise). Wagener and Brand (2005) found that a speech-shaped noise resulted in a higher sensitivity speech test than fluctuating multi-talker babble noise. A similar result was found elsewhere with a Dutch speech-in-noise test (Francart, 2011). Specifically, the stationary noises were found to produce steeper intelligibility functions than fluctuating babble noise. The babble noise, however, better discerned between levels of hearing impairment (by SRT) than the stationary noise maskers (Francart, 2011)—a result that may be due to

differences in temporal processing. As babble noise fluctuates, listeners are able to take advantage of amplitude variations in the envelope (i.e. gradual changes in amplitude) to detect the signal (Moore, 2008; Wagener & Brand, 2005). The relative SNR is much more favourable (i.e. higher) in these temporal dips, allowing the listener a brief "glance" at the signal—a phenomenon known as "masking release" (Hopkins & Moore, 2009). By the same process, individuals with normal hearing also perform better in modulated (i.e. fluctuating) noise compared to steady noise (Peters, Moore, & Baer, 1998).

Taken as a whole, the aforementioned studies illustrate the importance of selecting masking noise to complement the goals of the speech test. Speech-shaped noise may be preferable in a research context, where high sensitivity test is desirable to discern between two variables. High sensitivity is also beneficial when measurements are repeated over time, as differences in performance can be attributed to the variable being manipulated. Conversely, the use of multi-talker babble may be suited to clinical testing where simulation of an everyday noisy situation is desired, or as a means of distinguishing between levels of hearing loss (Francart, 2011).

1.6 Sentence tests

Another important consideration in speech audiometry is whether monosyllabic or sentence stimuli should be used. This decision should be based on the purpose of the test and the cognitive capabilities of the listener. Monosyllabic word stimuli are suited to situations where contextual information may overtly influence the listener's response. They are also advantageous in that the listener does not have to repeat or recall an entire sentence, and therefore, memory does not constrain auditory performance (Wilson et al., 2007). Thus, short stimuli should be used when testing populations with impaired memory function. However, a number of arguments exist in favour of the use of sentence stimuli over monosyllabic words. Firstly, as everyday speech often consists of sentences, they have higher face validity as stimuli in speech testing. Secondly, the use of sentence stimuli would test the listener's ability to perceive multiple speech sounds in a single trial, enhancing the time-efficiency of the test (Hochmuth et al., 2012). In spite of these advantages, and the ready availability of sentence tests, they are rarely used in NZ clinics.

There are a number of sentence tests available for use, and these can be divided into two distinct groups. The first group, referred to as "Plomp-type" sentences, are based on everyday conversation. The sentence lists are phonemically balanced; however, there is no consistent grammatical structure

across sentences (Plomp & Mimpen, 1979). An example of a test that uses Plomp-type sentences is the Hearing in Noise Test (HINT). Developed by Nilsson, Soli, and Sullivan (1994), the HINT consists of 25 phonemically balanced lists of 10 sentences spoken by a male speaker in the presence of spectrally matched masking noise. Since its development, the HINT has gained international popularity, and is available in a number of different languages, such as Cantonese (Wong & Soli, 2005), NZ English (Hope, 2010), and Swedish (Hällgren, Larsby, & Arlinger, 2006).

The format of the second group, known as Matrix Sentence Tests (MSTs), was originally pioneered by Hagerman (1982) in Swedish. Hagerman's goal was to create a standardised speech-in-noise test with ample speech material. The matrix consisted of 10 identically structured five word sentences (name, verb, number, adjective, object). For example (translated into English):

"Karin gave two old buttons."

The original Swedish version contained 13 phonetically balanced sentence lists dictated by a female speaker. The co-articulation between words was avoided in the recording process so that the words could be cut individually and synthesised to generate semantically dissimilar, but syntactically identical, sentences. Word files with an unnatural sound, such as long silences before or

after the words, were re-edited. The result was an essentially unlimited repertoire of speech material, which consisted of 10^5 or 100,000 possible unique sentences.

1.6.1 The advantages of MSTs

The MST format has numerous advantages over existing speech tests. It is low redundancy and semantically unpredictable; thus, it prevents contextual information from influencing a listener's response (Hochmuth et al., 2012). In addition, the uniform grammatical structure of matrix sentences permits the generation of an essentially unlimited repertoire of test sentences (Hagerman, 1982; Hochmuth et al., 2012). This is an advantage over the CVC meaningful word lists, which only has 10 lists of 10 words available for clinical use, and may be easily memorised in the case of repeat testing (Boothroyd & Nittrouer, 1988). Due to these advantages, MSTs have been developed for a number of languages, including Danish, Polish, Spanish, Dutch, Finnish and Italian (Acar et al., 2011; Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek et al., 2010; Puglisi et al., 2014; Wagener, Josvassen, & Ardenkjær, 2003). The uniform format of the MST allows for comparability across languages, with experimental evidence showing similar reference intelligibility functions across the Danish, Dutch, French, and Polish languages. This effectively allows listener performance to be compared internationally (Zokoll et al., 2013). However, listener performance is

dependent on the specific dialect used, as this performance may be confounded by the speaker's pronunciation (Hochmuth et al., 2012). For that reason, the British English version would not be valid within a NZ context, which necessitated the development of a MST in NZ English.

1.7 Development of the UCAMST

1.7.1 Background

The UCAMST was developed by Trounson and O' Beirne (Trounson, 2012) as a speech test specifically for use in NZ clinics. The sentences were read by an actress with a verified NZ English accent, and the materials were selected and adjusted to ensure the distribution of phonemes matched those used in the NZ Hearing In Noise Test (NZHINT) (Hope, 2010). Table 1 shows the UCAMST word matrix.

Table 1

The UCAMST word matrix.

Name	*Verb*	*Number*	*Adjective*	*Object*
Amy	bought	two	big	bikes
David	gives	three	cheap	books
Hannah	got	four	dark	coats
Oscar	has	six	good	hats
Kathy	kept	eight	green	mugs
Peter	likes	nine	large	ships
Rachel	sees	ten	new	shirts
Sophie	sold	twelve	old	shoes
Thomas	wants	some	red	spoons
William	wins	those	small	toys

The UCAMST, unlike any known MST, included a visual component. The rationale for this was that visual feedback increases the perceived naturalness of speech (Mattheyses, Latacz, & Verhelst, 2009), and enhances speech intelligibility—particularly in situations with an unfavourable (i.e. low) SNR (Sumby & Pollack, 1954). The ability to switch between auditory-alone, auditory-visual, and visual-alone presentation modes would permit customisation of the speech test to complement clinical or research goals. For example, a visual-alone condition could be used to test lip-reading ability. However, the process of including a visual component presented some challenge in the development of the UCAMST, as outlined in the following section.

1.7.2 Recording and editing UCAMST sentences

The UCAMST sentences followed the typical matrix sentence format: a five word sentence consisting of a name, verb, number, adjective, object. For example:

"Amy bought two big bikes."

The method used to record UCAMST sentences was based on that of Wagener et al. (2003), who developed the Danish MST. In that study, one-hundred five word sentences (henceforth termed "original" sentences) were recorded in a manner that all words from one column were spoken in conjunction with all words from the following column. Figure 3 illustrates this method in the Danish version (English translation).

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

Figure 3. Sentence recording method used to ensure each word had 10 co-articulation specific realisations. Figure from Wagener et al., 2003.

This method ensured that each word had 10 co-articulation-specific occurrences—in other words, each word had 10 different realisations. In the editing phase, the files were cut in a manner that best preserved the natural prosody of the sentences by accounting for the co-articulations of word. This resulted in smooth and natural transitions between words in the sentence, as opposed to the original Swedish version, where isolated words were synthesised with no co-articulation (Hagerman, 1982). The word fragments were then able to be re-combined and synthesised to generate 100,000 unique sentences.

The method described above has since been widely adopted as standard in international MSTs (Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek et al., 2010; Wagener et al., 2003). The process of synthesising sentences, at times, resulted in unnatural sounding audio artefacts. In such instances, the affected sentences were excluded from the final sentence lists (Hochmuth et al., 2012; Houben et al., 2014). The UCAMST had the unique challenge of ensuring both audio *and* video components were perceived as natural at the edited transitions between words. A mismatch in the position of the actress's head across image transitions in the video component would result in a perceivable image jerk artefact referred to as a "judder".

Numerous precautions were taken to minimise judder in the development of the UCAMST. Using the recording method described above,

the actress read the 100 original sentences in full. A microphone and a camera with an autocue system were used to capture both the audio and video components (mp4 format). Back and neck supports stabilised the actress's head position, and post-recording algorithms were applied to improve video stability (Trounson, 2012). The original sentences, recorded in 720p resolution at 50 frames/second, were then cut and edited to generate 400 file fragments containing unique word *pairs*. A series of complex cutting rules were used to preserve the uniformity of facial expression across the edited transitions between file fragments. Figure 4 depicts how file fragments containing word pairs were edited together to create a complete sentence.

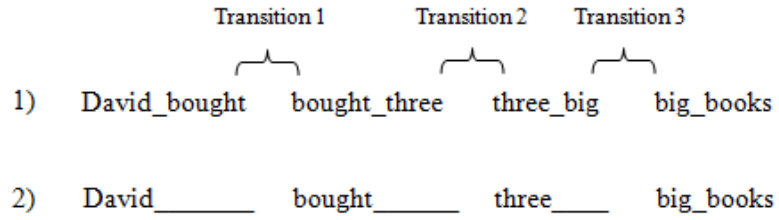

Figure 4 . Sentence 1) shows the file fragments involved used to create the sentence, and sentence 2) shows the specific audio content used from each fragment.

As shown, a sentence contains three "transitions", each of which marks the point of interchange between two consecutive file fragments, which is where judder may occur. In Figure 4, the file fragments were cut at the *beginning* of the word; however, in a number of cases, this was not an ideal in terms of the smoothness of the transition between images (Trounson, 2012). In such cases,

file fragments were cut to provide an optimal transition which, at times, resulted in only parts of a word being used, as exemplified in Figure 5.

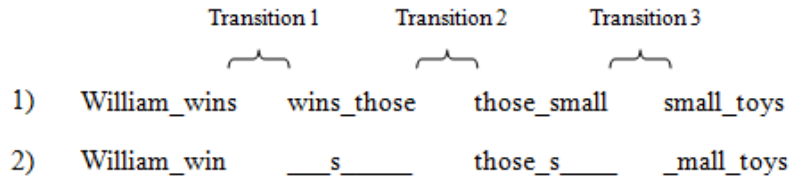

Figure 5. Sentence 1) shows the file fragments involved used to create the sentence, and sentence 2) shows the specific audio content used from each fragment.

A total of 3,000 unique transitions were generated; 300 were natural transitions spoken in full during the recording phase (e.g., "Amy gives two"), whereas 2,700 were *edited* transitions—word combinations not actually spoken by the actress during the recording phase. For example, "Amy gives six", which was formed by synthesising the 'Amy_gives' fragment from the original sentence "Amy gives two cheap bikes", and the 'gives_six' fragment from the original sentence "Kathy gives six cheap hats". Sentences formed using at least one edited transition are henceforth referred to as "synthesised" sentences to distinguish them from the 100 original sentences.

1.8 Study 1: Noticeability of video judders

1.8.1 Video judders

Although the audio component of the UCAMST was noted to have a natural sound, unfortunately a large proportion of the edited transitions had an evident judder. Judder was largely attributed to changes in the actress's head position, which affected the positioning of facial features (i.e. eyes, mouth, and nose) across the edited transitions (Trounson, 2012). As a means of quantifying the severity of the judders, the absolute difference between red, green, and blue (RGB) colour channels was calculated across the transitions—henceforth known as the "pixel difference value" (Trounson, 2012). The mouth area was excluded from this calculation to focus on changes in head position across transitions. The pixel difference value was calculated as the difference between the RGB colour channels of the last image frame prior to the transition and the first image frame post-transition. For example, for transition 1 in Figure 4, it would be calculated as the difference between the last frame of 'David_____', and the first frame of 'bought_____'. The pixel difference value was an objective measure of "judder severity": the larger the pixel difference value, the larger the difference between subsequent images, and the more severe the resultant judder. As the resized video contained 640 horizontal pixels by 480 vertical pixels by 256 colours/channel, the largest possible pixel difference value was 78,643,200 (which would occur changing

from a completely black screen to a completely white one). The 100 original sentences had values in the range of 178,761 to 249,157 (M = 217708 ±13923), and therefore, this describes the range of values for the *natural* image transitions (n = 300). For the purposes of this project, transitions with values of less than 300,000 were termed "no judder", while a "judder" was used to describe a transition with a value over 300,000. It is important to note that of the total number of edited transitions (n = 2700), a large proportion met the "no judder" criterion (44%). However, in contrast, over half (56%) of the edited transitions had pixel difference values of over 300,000 (i.e. were termed "judders"). Figure 6 shows the relevant proportion of transitions belonging to each of the described conditions.

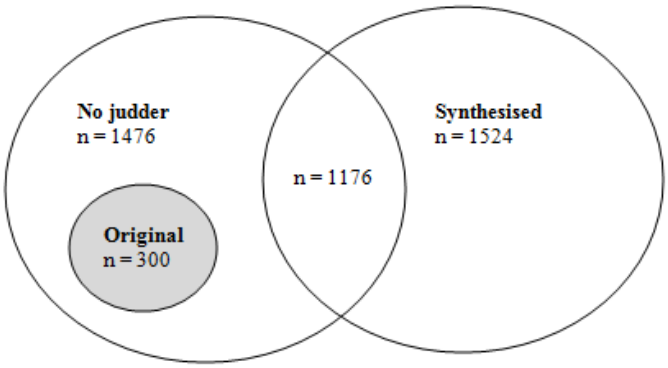

Figure 6. A breakdown of the total number of transitions (n = 3000) into "no judder" (n = 1476) and "synthesised" (i.e. edited) transitions (n = 1524). The original (i.e. natural) transitions (n = 300) met the "no judder" criterion (i.e. pixel difference value < 300,000), therefore make up a small proportion of the no judder sentences. The intersection of no judder and synthesised groups represents the synthesised sentences that meet the no judder criterion (n = 1176).

We were interested in determining the utility of the transitions in the judder group (i.e. pixel difference value > 300,000) in the auditory-visual UCAMST, as these transitions comprised over half (51%) of the available materials from which sentences could be generated.

1.8.2 Study 1 rationale

Although pixel difference value provided an objective index of judder severity, it was uncertain whether this would reflect how noticeable human observers would perceive the judders to be. Therefore, a subjective measure of how *noticeable* the judders were was necessitated to inform sentence selection for the auditory-visual UCAMST. This comprised the main goal of Study 1 of this project. Listeners with normal hearing were asked to assign a rating score based on how noticeable they found judders (from 0, "no noticeable judder", to 10, "highly noticeable judder") in both synthesised sentences and sentences with no judder, the latter of which acted as controls. Three variables were proposed to influence the "noticeability" of judders: 1) "judder severity" (i.e. the pixel difference value), 2) the number of juddered transitions within the sentence ("judder number"), and 3) the position of these judders within the sentence ("judder position"). The aims of Study 1 were twofold and are outlined below.

1) *Aim 1: To determine whether synthesised sentences were acceptable for inclusion in the auditory-visual UCAMST based on a comparison of judder noticeability rating scores with no judder control sentences.*

   The synthesised sentences were broken down into 18 sub-conditions based on judder severity, judder number, and judder position (sentence stimuli are further detailed in Chapter 2). Each of the 18 sub-conditions was then compared with the no judder sentences in a series of 18 paired *t*-test analyses. We predicted that synthesised sentences would have significantly higher mean rating scores than the no judder sentences, due to the higher pixel difference value at one or more transition within these sentences. Based on rating score, sub-conditions of synthesised sentences with an acceptable level of noticeable judder were eligible for inclusion in the auditory-visual UCAMST.

2) *Aim 2: To determine the validity of pixel difference value as a measure of noticeable judder.*

   The relationship between pixel difference value and rating score was determined by a linear regression analysis. The pixel difference value was hypothesised to be a significant predictor of rating score.

The method and results of Study 1 are contained within Chapter 2 of this thesis.

1.9 Normalisation of speech materials

1.9.1 The purpose of normalisation

The second goal of this project was to normalise the auditory-alone condition of the UCAMST, which was part of the process referred to as "optimisation" by various established MSTs (i.e., Hochmuth et al., 2012; Ozimek et al., 2010). In general, normalisation involves equalising the difficulty (% correct) of test materials (Wagener et al., 2003). This is typically achieved by manipulating word presentation level (dB) so that the psychometric function midpoint (L_{mid}, or intelligibility at 50%) equals the mean pre-normalisation midpoint across words (i.e., Ozimek et al., 2010). Figure 7 provides an example of pre- and post-normalisation psychometric functions.

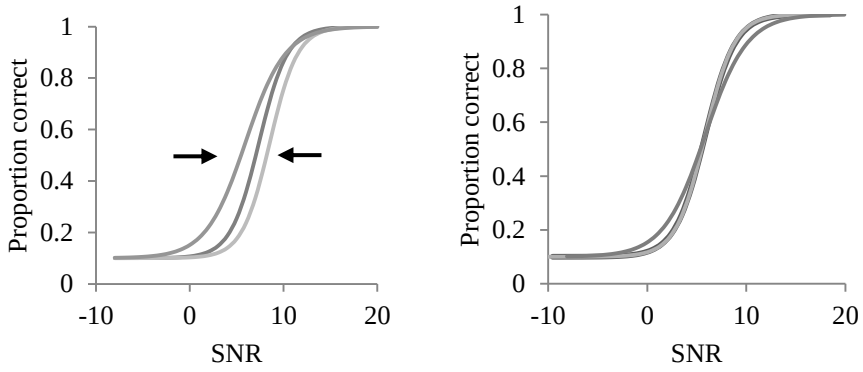

Figure 7. Pre-normalisation (left graph) and post-normalisation (right graph) word-specific psychometric functions. The arrows indicate the direction of the adjustment to the L_{mid}.

Normalisation has the effect of decreasing the standard deviation—that is, the distribution—of the word-specific L_{mid} measures (σ_{Lmid}), thus increasing the homogeneity of the data. The homogeneity of the data can be determined by the slope of the psychometric function, which has an inverse relationship with standard deviation. In short, a steep slope is indicative of a small standard deviation—that is, a small *distribution* of word-specific intelligibilities—and hence a homogeneous data set (Brand & Kollmeier, 2002).

Statistically, the efficacy of the normalisation process can be evaluated by calculation of the test-specific slope using a probabilistic model described by Kollmeier (1990), and included as equation 3 in Chapter 3 of this thesis. According to Kollmeier's model, the test-specific intelligibility function equates to the convolution of the mean word-specific slope and standard deviation of L_{mid} measures. The slope of the test-specific function ($s50_{test}$) can be determined using measured pre-normalisation data, and based on the required adjustments, predicted post-normalisation data and measured post-normalisation data. The increase to the test-specific slope reflects the efficacy of the normalisation process; specifically, the adjustments made to the pre-normalisation L_{mid} measure for each word.

Once the data is normalised, so-called "equivalent lists" can be generated for clinical use (Tye-Murray, 2014). The use of homogeneous lists ensures that differences in performance between test lists are attributed to the

test variable—rather than inherent differences in difficulty between speech materials—as variations in the intelligibility of different lists are smaller than the effect to be assessed (Wagener et al., 2003). In other words, the test has high "test re-test" reliability. Normalisation has a crucial role in the development of a speech test, increasing its sensitivity—and hence, reliability—as a measure of SRT (Brand & Kollmeier, 2002). To emphasise the key methodological differences between the UCAMST and published MSTs, the normalisation method used by published MSTs will be discussed over the following sections.

1.9.2 The Swedish MST

In the Swedish version (Hagerman, 1982), an equal difficulty (% correct) of test lists was achieved by manipulating the presentation level of matrix words, which increased the slope of the list-specific psychometric functions. Six subjects initially assessed the sentences. The presentation level of the matrix words (dB), and hence the overall SNR, was subsequently manipulated based on the shallowness of the initial intelligibility function slope (11.0%/dB). The level correction was limited to ± 4 dB to maintain a natural intonation; however, the maximum correction required was ± 1.3 dB. The adjustments were evaluated by 20 listeners with normal hearing in spectrally matched noise. Only five lists were tested, as the authors argued the same sounds were used in each list. The finalised lists were judged to be

homogeneous in terms of difficulty based on the similarity of mean intelligibility scores across lists.

1.9.3 The Danish MST

The method of normalisation has become more sophisticated since the original Swedish version, introducing the use of probabilistic modelling (Kollmeier, 1990) to ensure homogeneous data. This approach was used in the development of the Danish MST by Wagener et al. (2003). Here, the authors created 25 lists of 10 sentences for the optimisation phase, with each list containing one use of each matrix word. Listeners with normal hearing listened to sentence lists presented at 10 different SNRs, ranging from -18 dB to 0 dB SNR (2 dB increments). Spectrally matched stationary noise, created from superimposition of test sentences, was presented at a constant level of 65 dB SPL, and the word presentation level was varied. The listeners were required to respond verbally in a supervised (i.e. scored by an experimenter) open-set format (i.e. without the word matrix visible). A maximum likelihood function was used to fit a psychometric function to each word-specific realisation. The pre-normalisation standard deviation of word-specific SRTs was 3.8 dB, with a mean slope of 16.1%/dB. Application of the probabilistic model (Kollmeier, 1990) to pre-normalisation data produced a test-specific slope of 8.7%/dB. A level adjustment of ± 4 dB was applied to the pre-

normalisation word-specific SRTs, based on the difference between each word-specific SRT and the mean pre-normalisation SRT across all words. The predicted standard deviation was 1.8 dB, denoting a 2 dB decrement from the measured pre-normalisation value (3.8 dB). This resulted in a predicted post-normalisation test-specific slope of 13.2%/dB, which constituted a 4.9%/dB increase in slope from the measured pre-normalisation equivalent.

To obtain measured post-normalisation data, the sentences were re-evaluated by normal hearing listeners using a procedure developed by Brand and Kollmeier (2002), which will be briefly detailed here. This procedure involves the presentation of the sentences at two alternating SNRs, which approximate the "pair of compromise" (p_1 = .19 and p_2 = .81), to estimate the SRT and function in a quick and efficient manner. Test lists are presented in accordance with an adaptive procedure: the preceding stimulus and listener's response determine the presentation level of the following trial (Levitt, 1971). If the listener responds incorrectly to one trial, the stimulus level will be increased for the subsequent trial, and vice-versa for a correct response. Over the course of the test, the presentation level will start to converge around the listener's SRT, at which point the presentation levels are averaged to produce the final SRT. Using this method, Wagener et al. obtained measured post-normalisation values. The measured post-normalisation test-specific slope was 13.2%/dB, consistent with the predicted value. The mean list-specific SRT

was -8.4 dB SNR (± 0.2 dB), with an accompanying slope of 12.6%/dB (± 0.8%/dB).

1.9.4 The Polish MST

A Polish MST was developed by Ozimek et al. (2010). To establish measured pre-normalisation data, sentences were presented to 30 listeners in a multi-talker babble. Each of the 500 words was presented at 11 different SNRs, ranging from -16.5 to -1.5 dB, and listeners responded verbally in a supervised open-set format. The range of SNRs was selected to approximate a speech intelligibility range of 10% to 90%. The pre-normalisation mean word-specific slope was measured as 18.6%/dB, and the standard deviation of word-specific SRTs was 1.9 dB. A probabilistic model (Kollmeier, 1990) was used to estimate the slope of the test-specific intelligibility function, which was calculated as 13.9%/dB. A correction factor, limited to ± 3 dB, was applied. The standard deviation of the mean word-specific SRTs was predicted to be 0.4 dB, resulting in a predicted test-specific slope of 18.2%/dB. This denoted a 4.3%/dB increase from the test-specific slope measured in the pre-normalisation phase.

To obtain measured post-normalisation data, 10 lists of 10 sentences were re-evaluated in alternating SNRs of -7 and -11 dB by the same 30 listeners. The two SNRs were chosen as they approximated the “pair of

compromise" (Brand & Kollmeier, 2002). Post-optimisation, the measured mean test-specific slope was 17.1%/dB—a 3.2%/dB increase from the measured pre-normalisation equivalent. The mean list-specific SRT was -9.6 dB SNR (± 0.2 dB) with a slope of 17.1%/dB (± 1.5%/dB).

1.9.5 The Spanish MST

The Spanish MST was developed by Hochmuth et al. (2012). Spectrally matched stationary noise, created from the superposition of test sentences, was used as the masker. Listeners completed 12 "triple lists" (i.e. lists of 30 sentences), at SNRs that ranged from -15 to -2 dB, to determine pre-normalisation SRTs. The test was carried out in open-set format (i.e. in the absence of the matrix display), and verbal responses were scored by an experimenter. Based on pre-normalisation data, 8 word realisations (out of 500) were removed due to an inability to fit word-specific intelligibility functions. The resulting mean SRT across words was -8.1 dB SNR, with a slope of 27.0%/dB. A level correction of ± 3 dB was applied to each word realisation, which was judged as appropriate by a phonetician and a group of native Spanish speakers. Word realisations that exceeded this correction limit were excluded from the lists. The process of optimisation saw the standard deviation of word-specific SRTs decrease from 2.8 dB (measured pre-normalisation) to 1.1 dB (predicted post-normalisation). Using a probabilistic

model (Kollmeier, 1990), the test-specific slope was predicted to increase from 10.9%/dB (measured pre-normalisation) to 16.0%/dB (predicted post-normalisation).

To obtain post-normalisation measured data, 12 lists of 10 sentences were evaluated in an open-set format using the adaptive procedure described by Brand and Kollmeier (2002). The test-specific slope was 13.2%/dB, a 2.3%/dB increase from the pre-normalisation value. The resultant mean SRT and test-specific slope measured post-normalisation were -6.8 dB SNR (± 0.2 dB) and 13.2%/dB (standard deviation not supplied), respectively, for open-set measurement, and -7.7 dB SNR (standard deviation not supplied) and 14.0%/dB, respectively, for closed-set measurement.

1.9.6 The Dutch MST

A Dutch MST was developed recently by Houben et al. (2014). Word-specific intelligibility functions were created by assessing 360 sentences at five different SNRs (-12, -9, -6, -3, and 0 dB) in spectrally matched stationary noise created by the superimposition of sentence materials. Listeners with normal hearing responded by selecting words from a visible display containing the 50 matrix words—that is, in an unsupervised, closed-set format. Psychometric functions were fitted to the pre-normalisation data (not provided) using a logistic model that accounted for a 10% chance level. To

equalise word difficulty, an adjustment limit of ± 3 dB was applied to each word realisation. Sentences that were deemed too unnatural were discarded, and 14 lists of 20 sentences were created from the remaining 311 sentences. A further 15 listeners validated the pre-normalisation parameters (not provided) by assessing lists at three SNRs (-5, -7, and -9 dB), with noise presented at a constant level of 70 dB SPL. The resultant mean list-specific SRT was -8.4 dB SNR (± 0.2 dB), with an average slope of 10.2%/dB (± 0.9%/dB). As pre-normalisation data were not provided, the efficacy of the normalisation process for the Dutch MST cannot be commented on.

1.9.7 The Finnish MST

Dietz (2014) developed the Finnish version of the MST. The test materials comprised 30 lists of 10 sentences, which were homogeneous in terms of word and transition occurrences, and therefore, could be combined freely. Spectrally matched stationary noise was presented at 65 dB SPL. Twenty-one Finnish speakers participated; three were presented with the test lists at 10 different SNRs, which ranged from -5 to -14 dB SNR. After analysing the initial data, the range of presented SNRs was adjusted; the following 18 listeners listened to the test items at 15 different SNRs between -2 and -20 dB SNR. Responses were elicited verbally in a supervised open-set format. To establish pre-normalisation data, a logistic function was fitted to the

raw data. A mean word-specific SRT of -10.4 dB SNR (± 2.3 dB) with a mean slope of 18.9%/dB (± 7.9%/dB) was established. A level correction of ± 3 dB was applied to each word realisation. Based on the pre-normalisation data, words could be excluded for three reasons: 1) they exceed adjustment limit by 2 dB, 2) they have an abnormally shallow slope, and 3) they show unreliable SRT and slope parameters. Fifteen words met the first criterion, and no words met the latter two. The predicted post-normalisation mean word-specific SRT was -10.4 dB SNR (± 0.6 dB). To obtain post-normalisation data, 14 lists of 10 sentences were constructed. The lists were presented at three SNRs that approximated intelligibilities of 20%, 50%, and 80%. Measured data showed the lists to have a mean SRT of -10.1 dB SNR (± 0.7 dB) and a slope of 16.7%/dB (± 1.2%/dB).

1.9.8 The Italian MST

An Italian MST was developed recently by Puglisi et al. (2014). The test materials consisted of 30 ten item lists, which contained all 50 matrix words. To obtain pre-normalisation data, intelligibility was measured at SNRs between 2 and -18 dB (2 dB increments) in spectrally matched stationary noise. A logistic model was used to fit the word-specific intelligibility functions to raw data. The pre-normalisation mean SRT was -8.3 dB SNR (± 3.7 dB), with a median slope of 17.7%/dB over the 500 word realisations.

Level adjustments (limit ± 3 dB) were applied to the pre-normalisation data to equal the pre-normalisation mean word-specific SRT. The post-normalisation word-specific SRT was predicted to be -8.2 dB SNR (± 1.4 dB). In terms of the test-specific slope (Kollmeier, 1990), the adjustments were predicted to increase the test-specific slope from 9.2%/dB to 15.2%/dB (i.e. a 6.0%/dB increase). The equivalence of the compiled lists was evaluated using an adaptive procedure (Brand & Kollmeier, 2002), where six double-lists were presented at SNRs of -4.5, -7, and -9.5 dB (corresponding to 80%, 50%, and 20% intelligibility, respectively). The mean list-specific SRT and slope were calculated to be -7.3 dB SNR (± 0.2 dB) and 13.3%/dB (± 1.2%/dB), respectively. The actual increase between pre-normalisation and post-normalisation measured test-specific slopes was therefore 4.1%/dB.

1.9.9 Set presentation format

Before proceeding, it is necessary to discuss the importance of whether speech materials are presented in closed- or open-set format—that is, with or without a visible word matrix. An advantage of the uniform format MSTs is that it allows for closed-set testing (i.e. via a touch-screen), which in turn negates the need for a supervisor (i.e. an experimenter or audiologist) to score the responses of the listeners. Published MSTs have used both an open-set format (i.e., Dietz, 2014; Hochmuth et al., 2012; Ozimek et al., 2010; Wagener et al., 2003) and a closed-set format (i.e., Houben et al., 2014); whether this

presentation format affects listener performance is a contentious subject in the literature. In the Polish version, listener performance was unaffected by whether tests were performed verbally in open-set format under experimenter supervision, or carried out in a closed-set format without supervision (Ozimek et al., 2010). Elsewhere, Hochmuth et al. (2012) found that the use of a closed-set format, where listeners responded via touch-screen, resulted in significantly better listener SRTs compared with an equivalent open-set format. According to the authors, this may result from the lack of training provided to the listeners in comparison to Ozimek et al. (2010), where listeners would have been more familiar with the test materials prior to open-set testing. In this manner, a closed-set format is advantageous in that it improves the equivalency of list difficulty due to the provision of visual cues (Tye-Murray, 2014). Furthermore, removing the need for supervision is advantageous in that scoring is not susceptible to experimenter inattention or error. It would also be more time-efficient when testing large populations (Ozimek et al., 2010). However, as the literature tentatively suggests, training may be necessary to ensure listener performance is similar between the two formats.

1.10 Study 2: Normalisation of the auditory-alone UCAMST

1.10.1 Rationale for auditory-alone normalisation

The aim of Study 2 was to normalise the auditory-alone UCAMST, which would establish its reliability as a measure of a listener's SRT. The auditory-alone condition was normalised, specifically, as pilot testing with the Malay version of the UCAMST by Jamaluddin (2013) found presenting auditory-visual sentences at poor SNRs was essentially equivalent to a visual-alone condition, as listeners were reliant on lip-reading. This made it difficult to obtain the data required for a full auditory-visual psychometric function. Therefore, Study 2 normalised the auditory-alone condition of the UCAMST to exclude the effect of lip-reading, which significantly lifted the floor value of the measurement range. Study 2 comprises two sequential parts, which will be described below.

1.10.2 Part I: Normalisation of UCAMST

As a summary, the normalisation method used by published MSTs was as follows. Average listener performance on each word-specific realisation (n = 500) was measured across a range of SNRs, and an intelligibility function was fitted to each word realisation. Level adjustments were made to equal the SRT of each word realisation to the mean pre-normalisation SRT across all words. As each word realisation was contained within a single file fragment,

these published MSTs normalised both the words and fragments, simultaneously (Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek et al., 2010; Wagener et al., 2003). In contrast, the UCAMST presented a unique challenge in that the audio of a specific word often mapped onto more than one file fragment. Therefore, level adjustments could be applied to equalise the difficulty of either the file fragments ("fragment normalisation") *or* the matrix words ("word normalisation"). The latter involved averaging listener performance across fragments that contained the audio component of the specific word and fitting intelligibility functions to this data, resulting in 50 word-specific intelligibility functions. The former, similar to published MSTs, involved fitting intelligibility functions to each fragment, resulting in the generation of 400 fragment-specific intelligibility functions. In theory, word normalisation is justified by the fact that listeners are responding to words, not fragments. Normalisation by fragment does not consider the manner in which words map onto fragments; for example, a single word that maps onto two different fragments may be adjusted in two different directions (i.e. one increased and one decreased in level), resulting in an unnatural "level jump" within the sentence. Conversely, unlike fragment normalisation, the word normalisation technique described here does not account for the effect of context. As a brief example, the word "bought" may be more difficult to recognise when preceded by "Amy", but easier when preceded by "Thomas". In practice, the software used in this study acquired normalisation data at both

the word and fragment level; however, as listeners respond to words, and to prevent unnatural level jumps, the data from word normalisation was used to generate base lists (Part II). The efficacy of this technique will be the subject of further commentary in the discussion section. Furthermore, due to the benefits of different types of masking noise (see section 1.5.3), the normalisation process was carried out in both constant speech-shaped noise ("constant noise") and a six-talker babble ("babble noise"); the effect of each noise type on performance of the UCAMST was also the subject of investigation. Relevant aims and hypotheses for Part I are outlined below.

Aim 1: To determine the effect of noise type on the sensitivity and difficulty of the UCAMST through a comparison of pre-normalisation test-specific slope ($s50_{test}$) and mean L_{mid} (midpoint of the function, or intelligibility at 50%).

Previously, speech-shaped noises were shown to produce steeper intelligibility functions compared to when a fluctuating babble noise was used (McArdle et al., 2005). However, fluctuating noise was found to have a lower average L_{mid} (i.e. 50% intelligibility at a poorer SNR), and therefore, words were more easily detected in babble noise than speech-shaped noises. This was attributed to the phenomenon of "masking release" (Peters et al., 1998; Wagener & Brand, 2005). Based on this previous research, it was hypothesised the constant noise would result in a steeper intelligibility function and a higher mean L_{mid} than babble noise.

Aim 2: To normalise the difficulty of matrix words (L_{mid}) by adjusting word presentation level to equal the measured pre-normalisation data (mean word-specific L_{mid}).

As mentioned previously, the efficacy of the normalisation process described here can be observed with application of a probabilistic model (Kollmeier, 1990). Therefore, the predicted post-normalisation test-specific slope is hypothesised to be steeper than the measured pre-normalisation test-specific slope due to a reduction in the standard deviation of word-specific L_{mid} measures (σ_{Lmid}). The increase to test-specific slope, based on predicted post-normalisation data (in %/dB), is hypothesised to be comparable with existing MSTs (Hochmuth et al., 2012; Ozimek et al., 2010; Wagener et al., 2003).

1.10.3 Part II: Generation of test lists

Part II involved the selection of sentences to create 30 lists of 20 sentences for each noise type that were homogeneous in terms of sensitivity. The purpose of generating lists was 1) to obtain post-normalisation data, 2) to evaluate test re-test reliability of lists, and 3) to evaluate scoring methods in

determining SRT[1]. A summary of the aims and relevant hypotheses for Study 2 are detailed below.

Aim 3: To produce 30 base lists of 20 sentences that are predicted to be homogeneous based on predicted post-normalisation list-specific slopes.

Homogeneity was based on the definition used in previous MSTs (Ozimek et al., 2010; Wagener et al., 2003), in terms of the distribution of the predicted list-specific slopes across lists, as calculated using a probabilistic model (Kollmeier, 1990). List composition was adjusted to ensure minimal within-list variation in sentence-specific slopes (based on measured pre-normalisation data).

The method and results of Study 2 are located in Chapter 3 of this thesis.

1.11 Summary of project rationale

Taken as a whole, the information provided by Studies 1 and 2 would allow both 1) the selection of sentences with minimal noticeable judder for inclusion in the auditory-visual version of the UCAMST, and 2) generation of

[1] Unfortunately, due to time constraints, these lists could not be evaluated in the current project. Measured pre-normalisation and predicted post-normalisation are, however, provided here. Evaluation of these lists will take place in follow-on research.

base lists predicted to have minimal variation in list-specific slopes, and minimal within-list variation in sentence-specific slopes. This project constituted a stepping stone towards inclusion of the UCAMST the University of Canterbury Adaptive Speech Test (UCAST) platform, which will comprise a battery of audiological speech tests for clinical and research use (O'Beirne, McGaffin, & Rickard, 2012).

Chapter 2

Study 1: Judder Noticeability Rating Task

2.1 Method

2.1.1 Design

A range of synthesised sentences was included in Study 1. As mentioned previously, the noticeability of a judder may be influenced by the number of judder transitions in a sentence ("judder number"), as well as the position of the judder transitions within a sentence ("judder position"). As such, these were incorporated as variables, which are described in detail below.

The pixel difference values (i.e. "judder severity") were categorised into different brackets ("tier groups") using software custom-written in LabVIEW by Associate Professor O'Beirne. Table 2 displays the tier boundaries and the number of transitions within each tier group. The categories included in Study 1 were no judder, Tier 2, Tier 3, and Tier 4 transitions, as these groups represented the majority of the unique transitions (97%).

Table 2

Transitions organised according to tier.

Tier	Pixel difference value		*n*	Transition label
	Lower limit	Upper limit		
0	0	199,999	30	"No judder"
1	200,000	299,999	1446	
2	300,000	399,999	806	"Judder"
3	400,000	499,999	449	
4	500,000	599,999	186	
5	600,000	699.999	72	
6	700,000	799,999	10	
Total			**3000**	

To investigate the effect of judder number and judder position, the sentences were divided into those with one judder (at either transition 1, 2, or 3) and those with two judders (at either transition 1 and 2, 2 and 3, or 1 and 3). The transitions of sentences with two judders were selected so as to have pixel difference values within ± 2% of each other in the respective tier, preventing one more severe judder transition from influencing the rating score. The sentences were coded by judder number (*J1* or *J2*), judder position (*Tr01*, *Tr02*, *Tr03*, *Tr12*, *Tr13*, *Tr23*) and judder severity (*Ti2*, *Ti3*, *Ti4*). For example, a sentence with two judders at transitions 1 and 2 with Tier 3 pixel difference value would be labelled J2Tr12Ti3. The design for the synthesised sentences was 2 x 3 x 3 (judder number x judder position x judder severity), resulting in 18 sub-conditions, with four sentences per condition (n = 72). Sentences were selected at random from the pool of sentences fitting the

description of that condition. A further 28 sentences were sentences without judder randomly selected from the pool of 100 original sentences. This ensured that selected sentences with no judder had the lowest pixel difference values.

During the process of assembling sentence lists, the custom software did not generate an adequate number of sentences for three of the synthesised sentence sub-conditions. For example, only three sentences were produced for *J2Tr13Ti4*. To rectify this, the three sentences were re-used in each of the 10 lists, and the fourth sentence was randomly selected from the pool of sentences of equivalent judder number and tier (i.e. *J2Tr12Ti4* and *J2Tr23Ti4*). Sentence lists were checked to ensure the randomly selected fourth sentence did not appear more than once within a list. The software did not produce any sentences for the other two conditions (*J1Tr03Ti3* and *J1Tr03Ti4*). In these cases, the sentences were manually assembled by substituting the final word of an original sentence to create a transition of the desired tier. For example, the original sentence "David gives three cheap books" had the ending altered to "David gives three cheap shirts", resulting in a transition 3 pixel difference value of 414,435 (Tier 3). The initial two transitions would have pixel difference values within tiers 0 and 1, as these were natural transitions derived from an original sentence. Through this process, 94 (*J1Tr03Ti3*) and 29 (*J1Tr03Ti4*) sentences were produced for incorporation into the sentence lists. In sum, 10 unique lists of 100 sentences

were generated, composed of synthesised sentences (n = 72) and sentences with no judder (n = 28).

2.1.2 Participants

Study 1 participants were 18 adults (2 males; 16 females). Ages ranged from 21 to 28 years of age (M = 23.6y ±1.7y). Participants were required to 1) be native speakers of NZ English; 2) have hearing within normal limits (as shown by a hearing test); and 3) have no chronic issues of dexterity (as the procedure involved selection of words on a touch-screen). Both Study 1 and 2 gained approval from the University of Canterbury Human Ethics Committee (Reference: HEC 2014/49), and support from the Maori Research Advisory Group at the University of Canterbury; written informed consent was obtained from all participants for both studies. These documents can be found in Appendix A.

To determine hearing status, a hearing test was carried out in a soundproof booth at the audiology clinic at the Department of Communication Disorders, University of Canterbury. The test comprised questions on ear health, an otoscopic examination of the ear, and pure tone audiometry. Air-conduction pure tone thresholds were required to be within the normal range (≤ 20 dB HL) across the frequencies 250-8000 Hz; this encompasses what are considered the main speech frequencies (500-4000 Hz). Participants were

informed of the results of their hearing test: if normal hearing was confirmed, they were eligible to proceed with Study 1.

2.1.3 Equipment set-up

The UCAMST software, written in LabVIEW, was created by Associate Professor Greg O'Beirne, and was run from a laptop computer (HP Elitebook Revolve 810). This software presented the stimuli (described below) and provided a graphical user interface for data entry. The experimenter selected the parameters of the test, including which experiment to perform ['rating' (Study 1) or 'normalisation' (Study 2)], the language ('NZ English' in this case), and whether the sentence presentation was 'monaural' or 'binaural'. For Study 1, the binaural presentation mode, the 'rating' task, and NZ English options were selected, and the sentences were presented via headphones (Sennheiser HD280 pro, 64 Ω impedance) attached to an external Sound Blaster X-Fi Surround 5.1 Pro soundcard (Creative Technologies, Singapore). The laptop volume control was set to maximum and the participant name and number were entered into the software. The participant number determined which of the 10 lists of 100 sentences was used. The list allocation repeated modulo 10, such that participants 1 and 11 would receive the same list (list 1). The data collected data was imported into SPSS statistics v. 20 (IBM) for

statistical analyses, while graphics were produced with RStudio v. 0.98.1081 and Microsoft Excel.

2.1.4 Procedure

Participants were tested individually in a quiet room in the Department of Communication Disorders, University of Canterbury. The participant was seated in front of the laptop and verbally instructed that they would wear headphones and watch a video of a female speaker reading short sentences in quiet on the laptop screen. At the end of each sentence, they were required to rate how noticeable the "judders" at the transitions in the sentences were on 10-point scale, from 0 ("no noticeable judder") to 10 ("highly noticeable judder"), thus providing a noticeability rating score for each sentence. The verbal instructions were accompanied by on screen instructions, as shown in Figure 8.

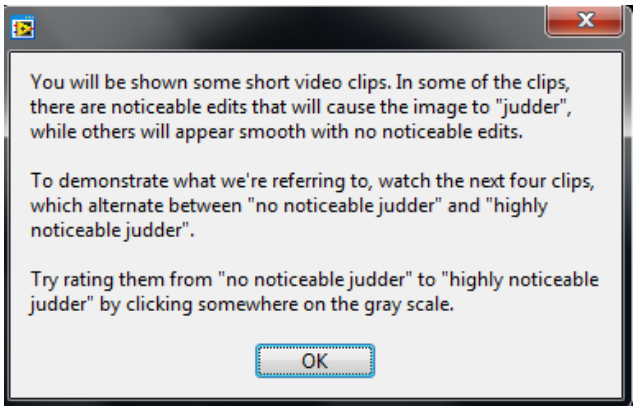

Figure 8. On-screen instructions presented prior to commencing practice phase.

To assign a rating score, the participant had to adjust an indicator on a sliding scale using the mouse or touch-screen, as shown in Figure 9. The rating score was based on the position of the slider on the scale.

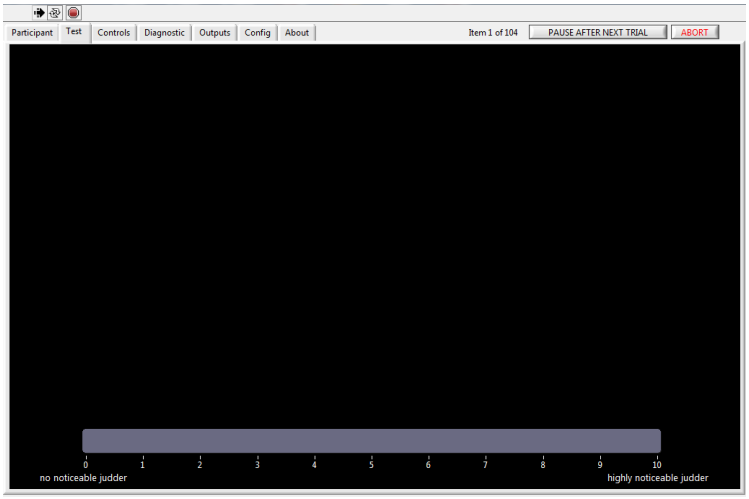

Figure 9. Response screen after each sentence presentation, showing a 10-point sliding scale from “no noticeable judder” at 0 to “highly noticeable judder” at 10.

To ensure full comprehension of the task the participant assessed four practice sentences consisting of two sentences with no judders alternated with two sentences which had numerous severe judders (three judder transitions of tiers 4, 4 and 3 respectively). On-screen instructions were provided again prior to commencing data collection, as shown in Figure 10.

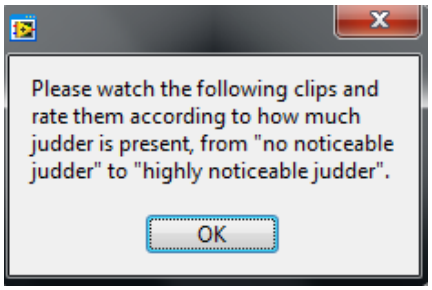

Figure 10. Second set of on-screen instructions presented prior to data collection phase.

When presented with the second set of instructions, the participant was asked if they had any questions about the task; the data collection phase commenced after any such questions were answered. The participant then evaluated a single list of 100 sentences presented in a random order. The rating score allocated to each sentence, along with the sentence details (i.e. condition, pixel difference value, and word composition) and the time stamp of each response, were exported as a text file (.txt) to a specified folder. On average, the rating task took 18 minutes to complete.

2.2 Results

2.2.1 Comparison of rating score between conditions

The first goal of Study 1 was to compare the rating score of no judder and synthesised sentences; the latter of which was divided into 18 sub-conditions based on judder severity, judder number, and judder position. For each participant, the rating score was averaged for each condition. Before proceeding to statistical analyses, the emergent trends in mean rating score will be discussed. The resulting mean rating score for each sub-condition[2], as compared to sentences with no judder, is shown in Figure 11.

[2] Key: J1Tr03Ti4 = one Tier 4 judder at transition 3; J2Tr12Ti3 = two Tier 3 judders at transitions 1 and 2; etc.

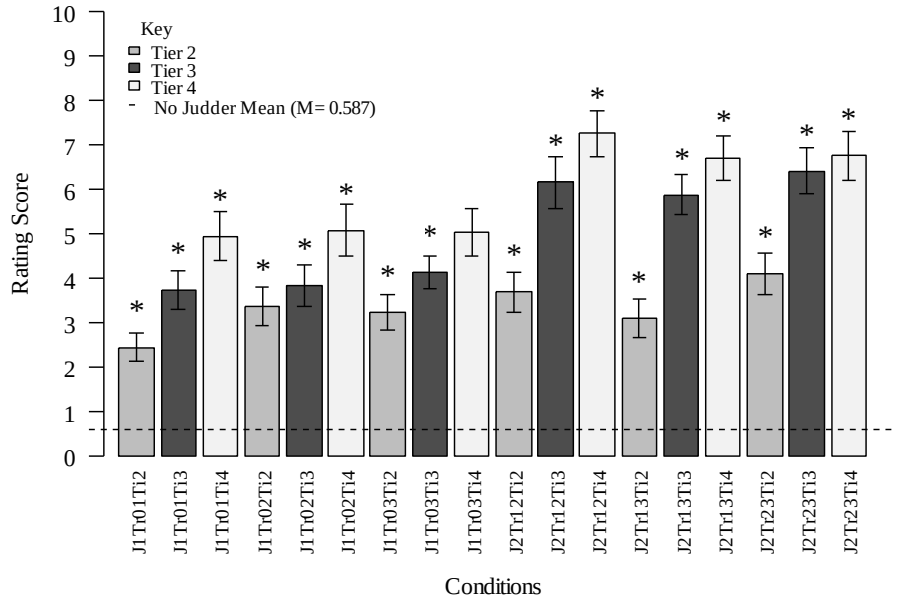

Figure 11. Histogram depicting the mean rating score for each sub-condition of the synthesised sentences. The mean rating score for sentences with no judder is represented by the dashed line. Error bars represent the standard error of the mean.

Figure 11 suggests two main trends: 1) sentences with two judders had higher rating score than sentences with one judder across the three tier groups (Tier 2, Tier 3, and Tier 4), and 2) judder position (Tr01, Tr02, Tr03, Tr12, Tr13, Tr23) did not appear to affect the overall mean rating score. A closer observation of these trends is provided in Figure 12 and Figure 13, respectively.

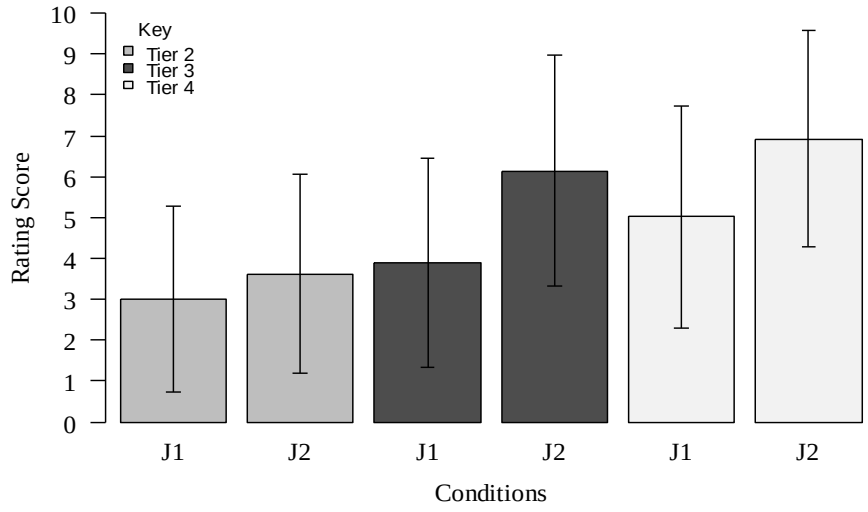

Figure 12. Mean rating score of one judder (J1) and two judder (J2) sentences within each tier group (Tier 2, Tier 3, Tier 4). Error bars represent standard deviation of each sub-condition.

Figure 12 further suggests a large difference in mean rating score between one judder and two judder sentences in the Tier 3 and Tier 4 groups. On the other hand, the difference between one and two judder sentences in the Tier 2 group was comparatively small.

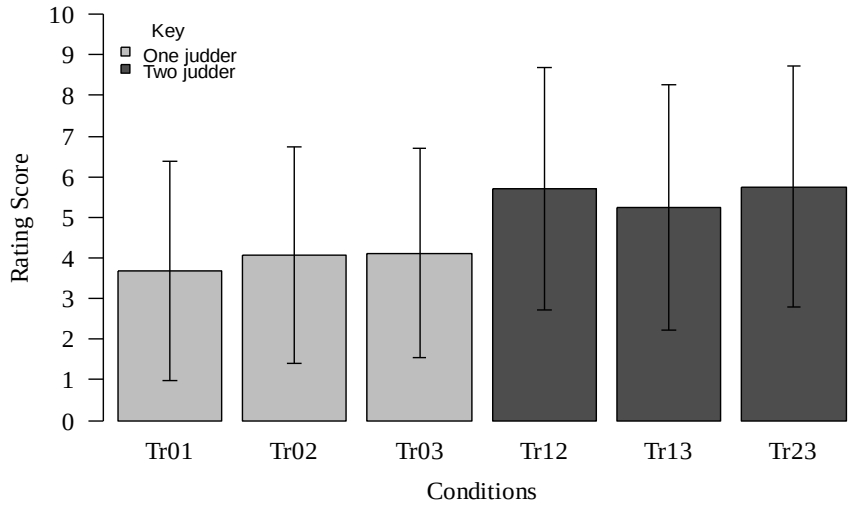

Figure 13. Mean rating score of sentences based on judder position. One judder sentences have a single judder transition (Tr01, Tr02, or Tr03); whereas two judder sentences have two judder transitions (Tr12, Tr13, or Tr13). Error bars represent the standard deviation of each sub-condition.

Figure 13 shows little difference in mean rating score with the manipulation of judder position in both one judder (Tr01, Tr02, and Tr03) and two judder (Tr12, Tr13, and Tr23) sentences.

The mean rating score of the 18 sub-conditions was compared to the mean rating score of no judder sentences using a series of 18 paired *t*-tests. To account for the multiple comparisons, a Bonferroni correction was applied by dividing the original alpha ($\alpha = .05$) by the number of paired comparisons

made (n = 18), resulting in an alpha of .00278 (Napierala, 2012). The multiple comparisons of group means are summarised in Table 3.

Table 3

Multiple paired comparisons between synthesised and no judder sentences.

Comparison Group	*Group Mean*	*SD*	*t-stat*	*p-value*
J1Tr01Ti2	2.44	1.10	-7.12	<.001*
J1Tr01Ti3	3.72	1.71	-7.76	<.001*
J1Tr01Ti4	4.94	2.25	-8.20	<.001*
J1Tr02Ti2	3.35	1.80	-6.51	<.001*
J1Tr02Ti3	3.83	1.97	-6.97	<.001*
J1Tr02Ti4	5.08	2.58	-7.40	<.001*
J1Tr03Ti2	3.21	1.49	-7.48	<.001*
J1Tr03Ti3	4.12	1.57	-9.53	<.001*
J1Tr03Ti4	5.03	2.24	-8.42	<.001*
J2Tr12Ti2	3.68	1.71	-7.69	<.001*
J2Tr12Ti3	6.15	2.36	-10.00	<.001*
J2Tr12Ti4	7.25	2.15	-13.16	<.001*
J2Tr13Ti2	3.09	1.59	-6.67	<.001*
J2Tr13Ti3	5.88	1.81	-12.38	<.001*
J2Tr13Ti4	6.71	2.18	-11.90	<.001*
J2Tr23Ti2	4.11	1.96	-7.63	<.001*
J2Tr23Ti3	6.41	2.20	-11.26	<.001*
J2Tr23Ti4	6.75	2.33	-11.22	<.001*

*Significant difference ($p < .00278$) between specified condition and sentences with no judder ($M = 0.59$, $SD = 0.66$).

According to the results, all 18 sub-conditions of the synthesised sentences had significantly higher average rating score than sentences with no judder. In short, there was on average a statistical difference between synthesised sentences and no judder sentences in terms of rating score.

2.2.2 Relationship between rating score and average pixel difference value

The second aim of Study 1 was to elucidate the relationship between the subjective rating score and the objective pixel difference value, so as to determine whether the calculated value could be used to predict the perceived level of judder in a sentence. As the "two judder" sentences were selected to contain two judder transitions with pixel difference values within ± 2%, the pixel difference value was averaged across the two transitions creating a variable called the "average pixel difference value". Outlined below is the method of calculating the average pixel difference value for each judder number condition.

1) One judder sentences: the pixel difference value of the single edited transition.

2) Two judder sentences: the average difference pixel value for the two edited transitions.

3) No Judder sentences: the average pixel difference value of all three original no judder transitions.

A linear regression was carried out using average pixel difference value as a predictor of rating score. Figure *14* depicts the relationship between average pixel difference value and rating score, including the resulting regression equation.

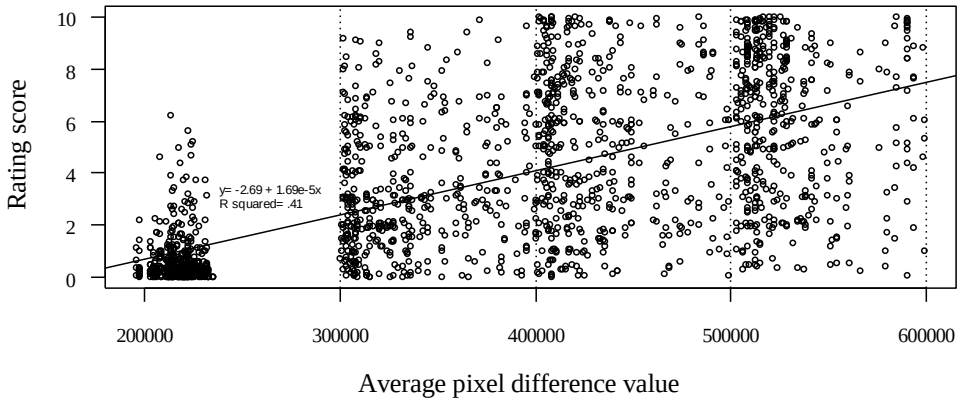

Figure 14. Scatter plot depicting the relationship between average pixel difference value and rating score. The solid line represents the model equation.

The figure shows that a small number of participants were assigning '0' rating scores to large judders, which may suggest that judders were missed in these sentences.

The regression model was statistically significant ($F_{(1, 1798)} = 1251.67$, $p = < .001$). Average pixel difference value accounted for approximately 41% of the variance in rating score ($R^2 = 0.41$, Pearson's $r = 0.64$). This correlation

coefficient (Pearson's r = 0.64) suggests a large effect size in accordance with the rule of thumb provided by Cohen (2003). On average, a one unit increase in average pixel difference value will result in a significant increase in rating score by 1.69×10^{-5}. The equation presented in Figure *14* can be used to predict rating score based on average pixel difference value. It is important to note that this is based on averages, and therefore, only provides a rough approximation of rating score. Nevertheless, this measure may be of utility in the selection of sentences for inclusion in the auditory-visual UCAMST. This selection process will be discussed in section 2.3.

To investigate the additional effect of judder number on rating score, the data was subjected to a stepwise multiple regression. This analysis informs on whether subsequent additions into the regression model would predict a unique proportion of the variance in rating score. Judder position was excluded from the analysis as it did not appear to affect rating score. In step 1, the average pixel difference value was entered, in step 2 both average pixel difference value and judder number were entered. The results showed that adding judder number significantly improved the model ($F_{(1,1797)} = 246.26$, $p < .001$, $R^2 = 0.48$), accounting for 7% of the variance in rating score over and above what is explained by average pixel difference value (41%). When reversed, the average pixel difference value significantly improved the model ($F_{(1,1797)} = 349.15$, $p < .001$), accounting for 10.1% unique variance in rating score over and above what is explained by judder number (38.1%).Although

conflation of the two predictors explains the greatest proportion of variance in rating score (48%), individually, the average pixel difference value predicts a greater proportion of variance than judder number (41% vs. 38.1%).

2.3 Selection of sentences for auditory-visual UCAMST

The result of the multiple comparisons carried out in section 2.2.1 revealed that synthesised sentences were rated as having significantly more noticeable judder than control no judder sentences. With the no judder sentences alone, the number of sentences that could be included in the auditory-visual UCAMST would be limited to 1,233. Thus, to maximise the number of sentences that could be included, it was decided that the two synthesised sentence sub-conditions with the lowest rating scores would be selected to provide a larger overall repertoire of sentences for auditory-visual testing. Table 4 displays the mean rating scores by judder number and judder severity. Judder position, which did not appear to affect the rating score, was excluded to simplify the selection process.

Table 4

Mean rating score across Tiers and Judders

Tier	*Judder number*	*Mean*	*SD*
Tier 2	1	3.00*	2.28
	2	3.63*	2.45
	1 & 2	3.31	2.38
Tier 3	1	3.89	2.56
	2	6.15	2.84
	1 & 2	5.02	2.93
Tier 4	1	5.02	2.72
	2	6.93	2.64
	1 & 2	5.97	2.84

* Acceptable conditions for inclusion in auditory-visual UCAMST.

The two lowest mean rating scores in the synthesised sentences were the Tier 2 sentences with one judder (M= 3.00, SD = 2.28) and two judders (M = 3.63, SD = 2.45). It was decided that these Tier 2 (M = 3.31, SD = 2.38) sentences had an acceptable level of noticeable judder. Based on this "Tier 2" criterion, usable sentences were selected from the original pool of sentences from which Study 1 sentence lists were generated. As the judder criteria become stricter, the number of transitions that can be used to form sentences diminishes, reducing the total number of available sentences. Table 5 provides a summary of the number of available sentences for auditory-visual presentation, including those with no judder.

Table 5

Synthesised and no judder sentences available for inclusion in auditory-visual UCAMST.

Judder number	*Tier*	*Number of sentences*
0 ('No Judder')	0	27
	1	1206
1	2	551
2	2	710
Total		2494

The inclusion criterion provided a large sample of acceptable sentences (n = 2494). Figure 15 compares the number of acceptable transitions with the number of unacceptable transitions based on the position within the sentence.

The results showed that a smaller proportion of transition 2 (52.1%) were acceptable compared with transition 1 (89%) and transition 3 (87.2%). The number of sentences is, therefore, reduced by smaller number of acceptable transition 2 fragment pairs.

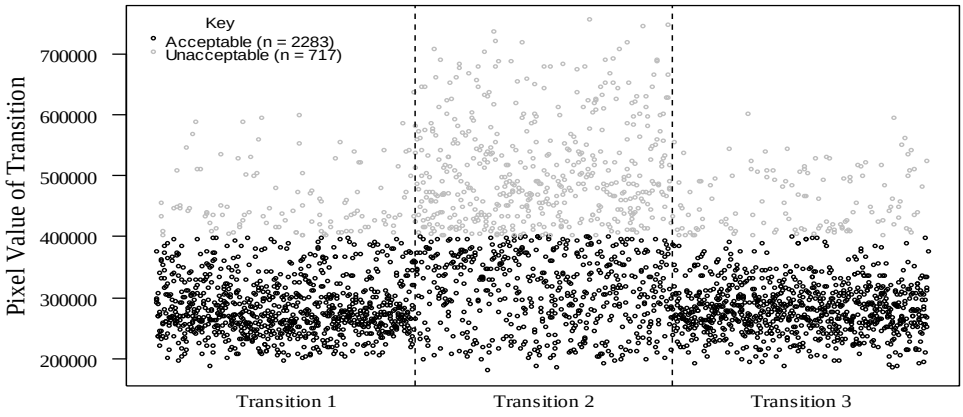

Figure 15. The 3,000 unique transitions labelled as "Acceptable" or "Unacceptable" based on the pixel difference value of each. Dashed lines illustrate transition boundaries. The position of each data point on the *x*-axis is random.

Chapter 3

Study 2: Normalisation of auditory-alone UCAMST

5.1 Part I. Normalisation

5.1.1 Participants

Participants were 17 adults (2 males; 15 females) from 21 to 28 years of age (M = 23.54 y ±1.64 y). The same inclusion criteria applied as in Study 1 (see Chapter 2); those who had not participated in that study (n = 2) were required first to provide written informed consent before undergoing a hearing test. Each participant received a $20 petrol voucher as compensation for their time and effort.

5.1.2 Generation of masking noise

Two types of noise were generated for use in the UCAMST: constant speech-shaped noise ("constant noise") and six-talker babble ("babble noise"). Briefly, the constant noise was generated by randomly superimposing the audio recordings from the actress 10,000 times in an automated process. This

meant that the signal and the noise had an almost identical spectral content (i.e. it was "spectrally matched"). A full description can be found in King (2010). The six-talker babble noise that was used in this study was generated for a previous study in the department (for details, see Spencer, 2011). In brief, the noise consisted of semantically anomalous sentences read by three male and three female speakers with NZ English accents.

5.1.3 Initial pilot of SNRs

Study 2 was piloted in constant noise with four SNRs (-15, -11.5, -8 and -4 dB) in constant noise presented binaurally. The desired SNRs were achieved in by varying signal level with a constant noise level of 65 dB SPL. The averaged performance across SNRs (data not shown) did not produce an adequate psychometric function, with the two highest SNRs producing a proportion of correct response scores, or p(c), of over .90, and the lowest SNR condition, producing a p(c) of .57. In other words, the measurement range was overall too easy. The -4 dB SNR condition was therefore replaced with a more difficult -18.5 dB SNR condition. Average performance across these adjusted SNRs (-8, -11.5, -15, and -18.5 dB) was found to produce an adequate psychometric function.

5.1.4 Procedure

Study 2 was carried out in a soundproof booth at the Department of Communication Disorders or the Rutherford building at the University of Canterbury in Christchurch. Participants were tested individually using the same equipment and software as Study 1 (see Chapter 2). In the software, the 'normalisation' task was selected with 'binaural'[3] presentation and either 'constant' or 'babble' noise (see Chapter 1). As 17 participants were recruited in total, the first nine participants received constant noise, while the final eight participants received babble noise.

First, the system volume control was set to maximum and participant name and number was entered into the software. The participant was then seated in front of a laptop and was verbally instructed that: 1) they would hear short sentences in noise, in which the words would change in loudness and may be difficult to hear; and 2) they were required to choose the sentence they heard by selecting the words on the touch-screen or using the mouse (i.e. it was a closed-set format). The words could be selected in any order and the

[3] Binaural listening (i.e. listening with both ears), compared to monaural (i.e. listening with one ear), has been shown to improve speech intelligibility in noise in a range of situations; the central auditory system can take advantage of acoustic differences between the two ears to diminish the masking effect of noise (Moncur & Dirks, 1967; Porter, Grantham, Ashmead, & Tharpe, 2014). We therefore expect the L_{mid} measures in this study to be at lower SNRs than published work that used monaural presentation.

participant was advised to guess a word when uncertain. Breaks were encouraged, and the task could be suspended by selecting a button that read "PAUSE AFTER NEXT TRIAL". Figure 16 depicts the layout of the response panel.

Figure 16. Matrix layout of response panel after each sentence presentation. A closed-set format was used with the 50 matrix words visible. Responses were entered by touching the desired word from each column.

After instructions were delivered and any questions were answered, the participant put on headphones and commenced the practise phase. The participant assessed 20 sentences with SNRs that ranged from -10 to -12 dB SNR. The purpose of this phase was to allow the participant to become familiar with the user interface and task, and therefore minimise learning

effects (Tye-Murray, 2014)[4]. After the practise phase was completed, the data collection phase commenced. In this phase the participant assessed 400 sentences, in which all 400 file fragments containing word pairs (i.e. 'Amy_bought') were randomly presented at four different SNRs (-8, -11.5, -15, and -18.5 dB). On completion of this phase, the participant's data was saved to a .txt file in a specified folder for data analysis. On average Study 2 took 90 minutes to complete, excluding breaks.

5.1.5 UCAMST scoring

The responses collected for each fragment were scored as shown in Figure 17 and used to create intelligibility functions for both fragments and words. Information about which word parts were contained in each fragment sample was used in the scoring calculations. Each word was divided into two parts, which could be within one fragment or across two adjacent fragments.

[4] Additionally, in the Polish version, a training session equalised participant performance on closed- and open-set formats (Ozimek et al., 2010).

Scoring procedure for UCAMST sentences.

Actual	*Amy*		*bought*		*some*		*red*		*coats*		**Fragment scoring**		
Selected	*Amy*		*bought*		*some*		*red*		*coats*		**Pt1**	**Pt2**	**Total**
amy_bought	A	my	_	_							2/2	0/0	1
bought_some			bou	ght	s	_					2/2	1/1	1
some_red					_	ome	re	_			1/1	1/1	1
red_coats							_	d	co	ats	1/1	2/2	1
Word scoring	1		1		1		1		1				

Actual	*William*		*kept*		*ten*		*good*		*toys*		**Fragment scoring**		
Selected	*William*		*kept*		*ten*		*green*		*toys*		**Pt1**	**Pt2**	**Total**
william_kept	Will	iam	_	_							2/2	0/0	1
kept_ten			ke	pt	_	_					2/2	0/0	1
ten_good					te	n	_	_			2/2	0/0	1
good_toys							go	od	to	ys	0/2	2/2	0.5
Word scoring	1		1		1		0		1				

Actual	*Peter*		*has*		*six*		*good*		*mugs*		**Fragment scoring**		
Selected	*Peter*		*has*		*some*		*good*		*mugs*		**Pt1**	**Pt2**	**Total**
peter_has	Pe	ter	_	_							2/2	0/0	1
has_six			ha	s	s	_					2/2	0/1	0.667
six_good					_	ix	_	_			0/1	0/0	0
good_mugs							go	od	mu	gs	2/2	2/2	1
Word scoring	1		1		0		1		1				

Actual	*Oscar*		*got*		*four*		*red*		*books*		**Fragment scoring**		
Selected	*Oscar*		*got*		*four*		*red*		*books*		**Pt1**	**Pt2**	**Total**
oscar_got	Os	car	_	_							2/2	0/0	1
got_four			go	t	f	_					2/2	1/1	1
four_red					_	our	re	_			1/1	1/1	1
red_books							_	d	bo	oks	1/1	0/2	0.333
Word scoring	1		1		1		1		0				

Actual	*Hannah*		*sees*		*nine*		*large*		*toys*		**Fragment scoring**		
Selected	*Amy*		*sees*		*eight*		*cheap*		*spoons*		**Pt1**	**Pt2**	**Total**
hannah_sees	Han	nah	s	_							0/2	1/1	0.333
sees_nine			_	ees	_	_					1/1	0/0	1
nine_large					ni	ne	lar	_			0/2	0/1	0
large_toys							_	ge	to	ys	0/1	0/2	0
Word scoring	0		1		0		0		0				

Figure 17. Scoring procedure for the matrix sentences illustrated with five examples.

5.1.6 Normalisation by fragment

Normalisation by fragment was carried out separately for each noise condition. The mean intelligibility (%) for each fragment was first calculated across the four SNRs. These mean scores were used to construct fragment-specific intelligibility functions with intelligibility (%) on the *y*-axis and SNR on the *x*-axis. A logistic model was used to fit these functions, as shown in Equation (1):

(1)

$$SI(L) = \frac{1}{A}\left(\frac{(1 + SI\max) \cdot (A - 1)}{1 + \exp(-4.S.[Lmid - L])}\right)$$

Equation adapted from Green and Swets (1966), Kollmeier and Wesselkamp (1997) and Wagener et al. (2003).

where *SI* is Speech Intelligibility (%), *S* is slope (%/dB), *L* is level, *Lmid* is the midpoint of the psychometric function, *SImax* is the function ceiling, and *A* is the number of alternatives (with *1/A* being the function floor). When the number of alternatives is very large, a "0 floor" version of the equation may be used, as follows:

(2)

$$SI(L) = \frac{1}{1 + \exp(-4 \cdot S \cdot [Lmid - L])}$$

Zero floor version of Equation 1.

The L_{mid} for each fragment was derived from this function, and the mean L_{mid} across fragments was then calculated. The presentation level of each fragment was adjusted so that each individual L_{mid} was equal to the pre-normalisation mean L_{mid} (i.e. the mean intelligibility at 50%). The adjustment limit was ± 3 dB, consistent with the most conservative limit used in international versions of the matrix sentence test (Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek, Kutzner, & Libiszewski, 2012).

5.1.7 Normalisation by word

In order to apply level adjustments to each individual word, the fragments needed to have components from different words adjusted independently, as exemplified in Figure 18.

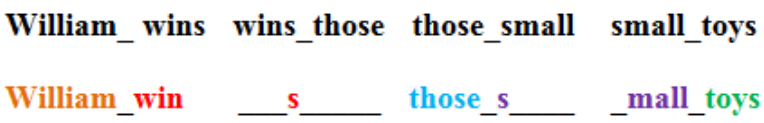

Figure 18. The top sentence displays constituent sentence fragments, whereas the bottom sentence shows components used to create the audio of this sentence. These audio components were adjusted independently and like colours represent an equal magnitude of adjustment.

Using custom written software, the point on each waveform at which the transition to a new word occurred was labelled (in ms) for each fragment. This data was stored in the UCAST software and used to adjust programmatically the individual words of each word pair independently.

The mean intelligibility (% correct) for each fragment was first calculated across the four presented SNRs. The mean intelligibility was then averaged across fragments that contained that specific word sound. As a brief example, the function for the word "four" was constructed by averaging the fragments containing "f___" (i.e. sees_four, kept_four, sold_four, etc.) and those containing "_our" (i.e. four_green, four_big, four_dark, etc.) across the four presented SNRs. The result was the mean intelligibility for each word at each of the four SNRs. These values were used to derive the word-specific L_{mid} using the logistic model presented in equation (1) fitted to a 0.1 floor to represent 10% chance rate (a 1 in 10 chance of correctly selecting the matrix word). Each word-specific L_{mid} was then adjusted for to match the pre-

normalisation mean fragment L_{mid} with an adjustment limit of ± 3 dB. The pre-normalisation mean fragment L_{mid} was used so that normalisation by word and by fragment produced sentences with the same mean SRT.

5.2 Part I. Results

5.2.1 Constant noise

The data were first normalised by fragment. Fifteen (4%) fragments were discarded due psychometric functions with an unusual morphology, ceiling or floor effects[5], which prevented an adequate fit with the logistic model. The most frequent reason for exclusion was ceiling effects (n = 12). Figure 19 provides an example of a fragment with a floor effect ('wins_ten'), ceiling effect ('nine_big'), and inconsistent performance across SNRs that resulted in an abnormal morphology ('sees_twelve'). The fragment ('green_hats') had a good function fit, and the raw data follows the expected sigmoid ('s-shaped') function.

[5] . A "floor effect" describes when performance reaches the lower limit of the measurement scale. Conversely, a "ceiling effect" describes when performance reaches the upper limit of the measurement scale (Twisk & Rijmen, 2009).

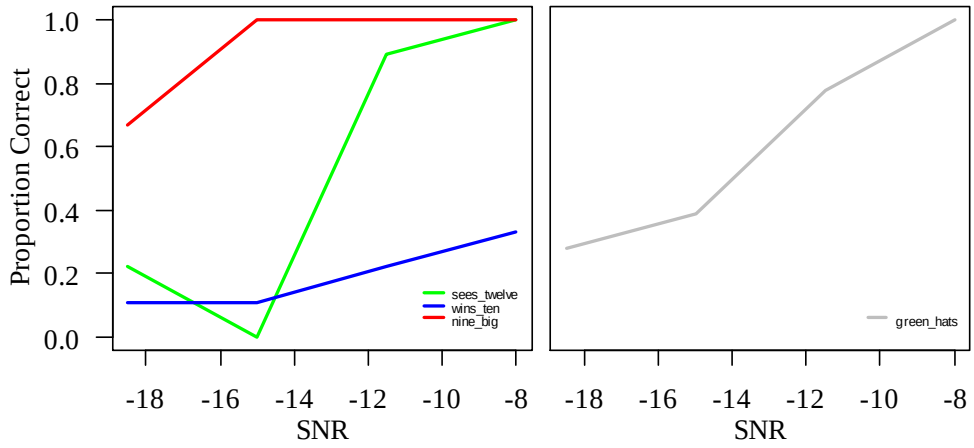

Figure 19. Fragments with poor (left graph) and good (right graph) function fits. The examples provided are based on raw data performance across four SNRs (-8, -11.5, -15, and -18.5 dB).

Some discarded fragments contained the same words; for example, four started with "nine" (eg 'nine_big', 'nine_old', 'nine_good', 'nine_dark') and three started with "Rachel" (e.g., 'Rachel_has', 'Rachel_bought', 'Rachel_got'). Both of these groups exhibited ceiling effects, and therefore these word pair combinations were avoided when assembling sentence lists for use in constant noise (see section 3.3).

The remaining 385 fragments produced a mean pre-normalisation L_{mid} of -14.2 dB SNR (± 2.1 dB). On average, the easiest fragment to detect was 'two_old' (L_{mid} = -20.5 dB SNR), whereas the hardest fragment to detect was 'gives_twelve' (L_{mid} = -7.3 dB SNR). The fragments were adjusted to the mean pre-normalisation L_{mid} of -14.2 dB SNR; based on this, the fragments were predicted to have a 50% chance of correct identification when presented

at -14.2 dB SNR. The average magnitude of adjustment required prior to applying the limit was 1.7 dB (± 1.0 dB). Eighty-one (20%) fragments had a required adjustment of greater than the ± 3 dB limit. The average magnitude of adjustment was 1.6 dB (± 1.0 dB) with the limit enforced.

The pre-normalisation word-specific functions were then fitted. According to this data, the easiest word to detect was "nine" (L_{mid} = -18.4 dB SNR) whereas the hardest word was "shirts" (L_{mid} = -8.0 dB SNR). The mean word-specific L_{mid} was -13.6 dB SNR (± 2.4 dB) and the mean slope was 14.4%/dB (± 3.2%/dB). The pre-normalisation (measured) descriptive statistics for each of the five word positions (name, verb, number, adjective, and object) are provided in Table 6.

Table 6

Measured (pre-normalisation) descriptive statistics of L_{mid} and slope based on word position in constant noise.

Position	*Mean L_{mid} (dB SNR)*	*SD*	*Mean slope (%/dB)*	*SD*
Name	-15.3	1.6	16.2	2.6
Verb	-11.3	1.8	13.8	2.5
Number	-15.2	1.9	16.1	2.8
Adjective	-13.7	2.0	13.7	1.8
Object	-12.6	2.2	12.2	4.2

When examining the data based on word position, the names (L_{mid} = -15.3 dB SNR ± 1.6 dB) and numbers (L_{mid} = -15.2 dB SNR ± 1.9 dB) were the easiest

word positions to detect and had the smallest distribution of L_{mid} measures. As a result, names (16.2%/dB ± 2.6%/dB) and numbers (16.1%/dB ± 2.8%/dB) had the steepest slopes at the midpoint due to the inverse relationship between standard deviation and slope (Brand & Kollmeier, 2002). The verb (L_{mid} = -11.3 dB SNR) and object (L_{mid} = -12.6 dB SNR) word positions were, on average, the most difficult to detect.

The data was then normalised; the pre-normalisation word-specific midpoints were then adjusted to equal the mean pre-normalisation mean fragment L_{mid} of -14.2 dB SNR (± 2.1 dB). The mean magnitude of adjustment required prior to applying the limit was 2.0 dB (± 1.5 dB); seven words (14%) required an adjustment in excess of the ± 3 dB limit. The mean magnitude of adjustment with the limit applied was 1.7 dB (± 1.0 dB). Figure 20 displays the pre-normalisation and predicted post-normalisation psychometric functions by each of the five word positions in the sentence (i.e. name, verb, number, adjective, object) for the 50 matrix words.

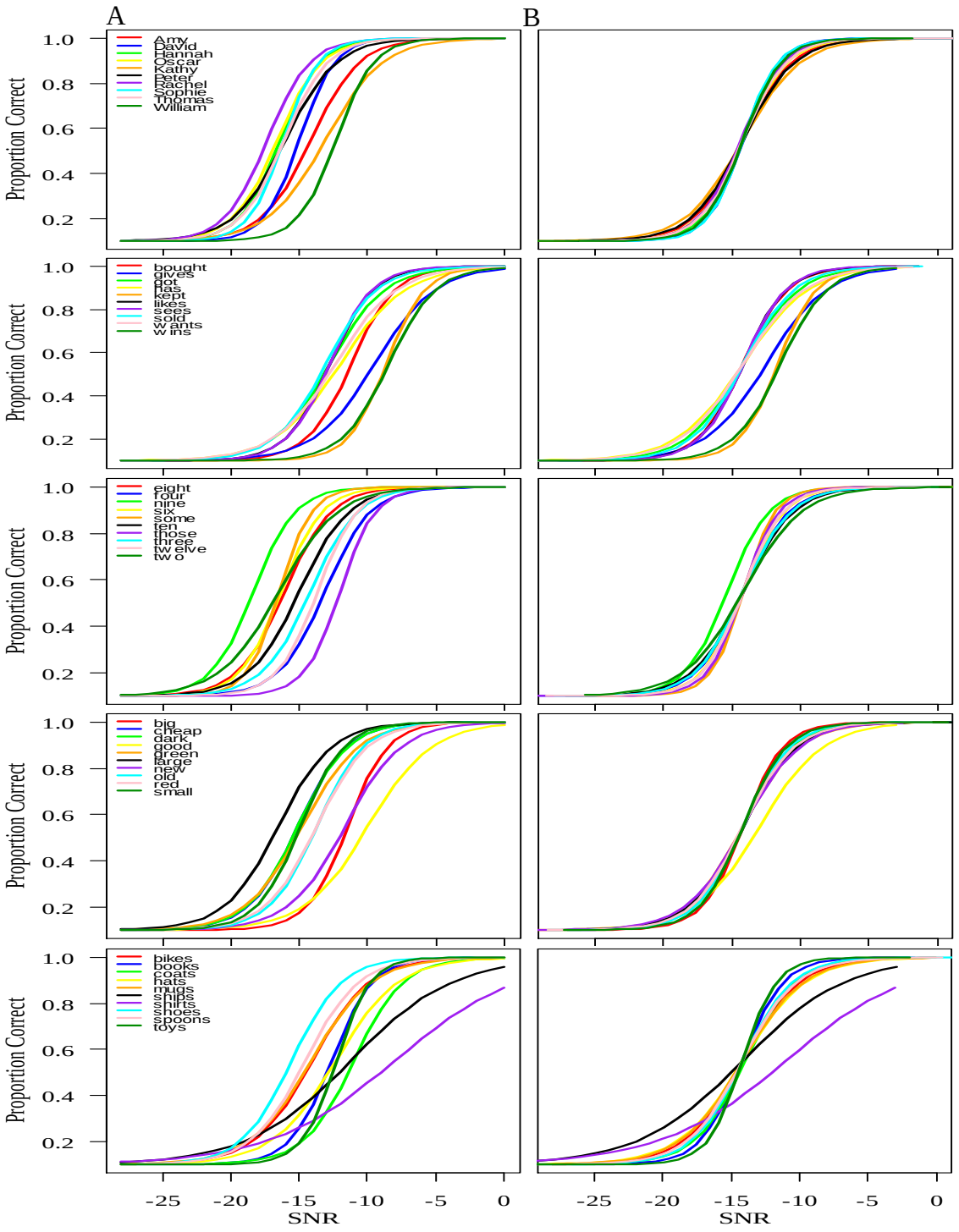

Figure 20. The pre-normalisation (Panel A) and predicted post-normalisation functions (Panel B) for the constant noise condition by word position.

According to Figure 20, the functions were adjusted by a maximum of ± 3dB to have an equivalent L_{mid} across words. The adjustment results in the aligning of the L_{mid}, and therefore, greater overlap in the post-normalisation functions than in the pre-normalisation functions. Those functions that do not align with other functions in the post-normalisation panel required an adjustment that exceeded the limit. This is particularly evident in the object words "shirts" and "ships". With the adjustments made, the post-normalisation mean word-specific L_{mid} is predicted to be -14.0 dB SNR ± 0.8 dB, denoting a 1.6 dB decrease in the standard deviation of word-specific L_{mid} measures.

5.2.2 Babble noise

Babble noise data was normalised first by fragment. Forty-seven (12%) fragments had to be excluded from the babble noise condition due to an inability to fit psychometric functions. The most common reason for excluding fragments was a ceiling effect (n = 24). Again, the discarded fragments contained certain word patterns. For example, "wins" (n = 8), "nine" (n = 7), "Thomas" (n = 4) and "some" (n = 4). As in the constant noise condition, fragments starting with "nine" displayed a ceiling effect. In contrast, fragments starting with "wins" were difficult to detect and thus had a floor effect. Fragments beginning with the word "Thomas" showed an unusual morphology, which resulted from inconsistent responses across the four

presented SNRs. Word pair combinations from discarded fragments were avoided in generating sentence lists in the babble noise condition to maximise the reliability of list functions (see section 3.3).

The mean L_{mid} across the remaining 353 fragments was -14.9 dB SNR (± 2.9 dB). Therefore, the UCAMST task was slightly easier in babble noise than in constant noise. 'Oscar_gives' was the easiest fragment to detect in this condition with an L_{mid} of -23.4 dB SNR). The most difficult fragment was 'kept_nine' (L_{mid} = -5.2 dB SNR). The fragments were adjusted to have an L_{mid} of -14.9 dB SNR, and the average magnitude of adjustment was 2.3 dB SNR (± 1.8 dB) prior to applying the limit. One-hundred and nine out of 400 fragments (37%) required adjustment in excess of the ± 3 dB limit. The average magnitude of adjustment with the limit applied was 1.6 dB (± 1.1 dB). The pre-normalisation word-specific functions were then fitted. The word "wins" was discarded and not used in the generation of sentence lists in babble noise, as it produced an abnormally steep word-specific function. The easiest word to detect was "nine" (L_{mid} = -27.7 dB SNR) and the hardest word was "shirts" (L_{mid} = 9.7 dB SNR). "Shirts" was also the hardest word in constant noise, although, not to quite the same extent (L_{mid} = -8.0 dB SNR). Owing to the difficulty of "shirts" in babble noise, which required a 24.6 dB adjustment to equalise the pre-normalisation mean L_{mid}, this was discarded from the test materials. The mean pre-normalisation L_{mid} across words was –14.5 dB SNR (± 3.6 dB) and the mean slope across words was 10.3%/dB SNR (± 2.8%/dB).

Table 7 shows the descriptive data based on word position in the sentence for babble noise.

Table 7

Pre-normalisation (measured) descriptive statistics of L_{mid} and slope based on word position in babble noise.

Position	*Mean L_{mid} (dB SNR)*	*SD*	*Mean slope (%/dB)*	*SD*
Name	-16.8	2.0	11.5	2.9
Verb	-10.4	2.5	8.6	2.2
Number	-18.1	3.6	10.7	2.6
Adjective	-14.0	1.9	11.5	2.7
Object	-13.2	2.8	8.9	2.1

As in the constant noise condition, the names (L_{mid} = -16.8 dB SNR ± 2.0 dB) and numbers (L_{mid} = -18.1 dB SNR ± 2.6 dB) were the easiest word positions to detect. The names (11.5%/dB ± 2.9%/dB) and adjectives (11.5%/dB ± 2.7%/dB) had the steepest mean slopes, a reflection of the small distribution of L_{mid} measures in these conditions (SD = 2.0 dB for names, SD = 1.9 dB for adjectives). Also as in the constant noise condition, the verb (L_{mid} = -10.4 dB SNR) and object (L_{mid} = -13.2 dB SNR) positions were the most difficult to detect. Thus, the same pattern of word difficulties was observed regardless of the masking noise used.

The word-specific midpoints were then adjusted to equal the mean pre-normalisation L_{mid} across fragments (L_{mid} = -14.9 dB SNR ± 2.9 dB). The

required adjustment prior to applying the limit was 3.4 dB (± 3.8 dB). Twenty words (out of 49, or 41%) required an adjustment in excess of the ± 3 dB limit. With the limit applied, the mean magnitude of adjustment was 2.2 dB (± 0.9 dB). Figure 21 shows the pre-normalised and predicted post-normalisation psychometric functions for each of the five word positions.

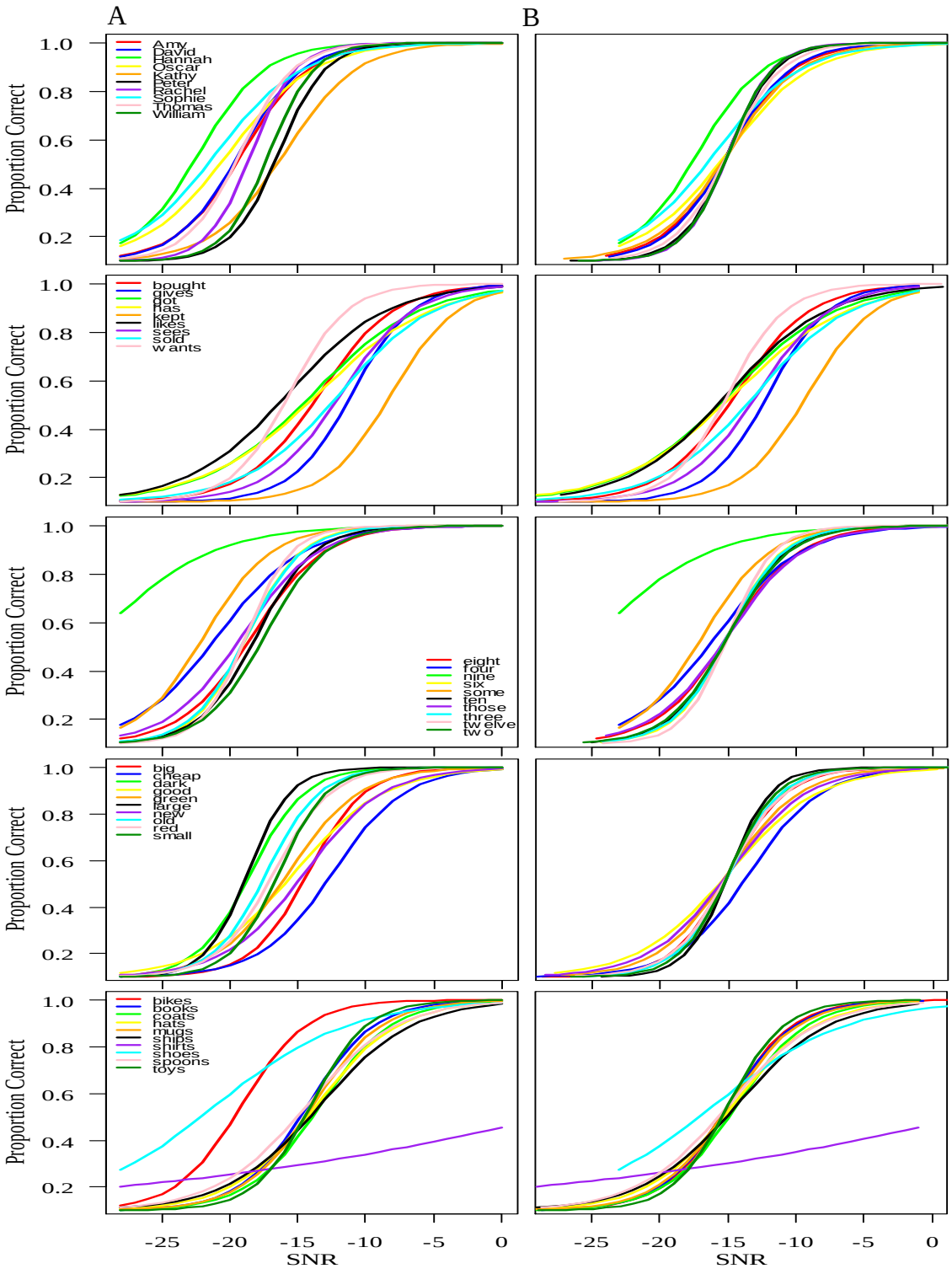

Figure 21. Babble noise pre-normalisation (Panel A) and post-normalisation (Panel B) psychometric functions.

In general, compared to constant noise (Figure 20), the predicted post-normalisation functions do not have as much overlap. This is likely due to the greater proportion of words (41%) that required adjustment in excess of the adjustment limit compared to the constant noise condition (14%); therefore, the L_{mid} of affected words are predicted to better approximate—but not equal—the mean pre-normalisation L_{mid}. The mean post-normalisation L_{mid} is predicted to be -14.9 dB SNR ± 1.9 dB, denoting a reduction in the standard deviation of word-specific L_{mid} of 1.7 dB.

5.3 Test-specific slope

The normalisation procedure can be evaluated by examining the slope of the test-specific intelligibility function ($s50_{test}$) based on the measured pre-normalisation data and the predicted post-normalisation data. The measured post-normalisation data will also be obtained for comparison; due to time constraints, this was not able to be achieved as part of the current project, and will be the subject of follow-on research.

The test-specific intelligibility function equates to the convolution of the mean word-specific functions and the standard deviation of SRTs (Kollmeier, 1990), as described in the probabilistic model presented in equation (3):

(3)

$$s50_{test} = \frac{s_{word}}{\sqrt{1 + \frac{16\, S_{word}^2 \times \sigma_{Lmid}^2}{\left(\ln\left(2e^{\frac{1}{2}} - 1 + 2e^{\frac{1}{4}}\right)\right)^2}}}$$

Equation adapted from Hochmuth et al. (2012).

where s_{word} is the mean slope of the word-specific functions, and σ_{Lmid} is the standard deviation of word-specific L_{mid} measures. A reduction in the standard deviation of word-specific L_{mid} measures can be observed when word presentation levels are adjusted, which in effect increases the slope of the test-specific function (Hochmuth et al., 2012). Unfortunately, due to time constraints, measured post-normalisation data was unable to be obtained. However, the predicted outcomes for $s50_{test}$ for both constant and babble noise will be evaluated here compared to the initial measured pre-normalisation values. A summary of this data for both constant and babble noise conditions is provided in Table 8.

Table 8

Measured pre-normalisation and predicted post-normalisation word-specific sensitivity (s_{word}) and difficulty (L_{mid}).

Noise	*M. mean s_{word} (%/dB)*	*M. mean L_{mid} (dB SNR)*	*SD (dB)*	*P. Mean L_{mid} (dB SNR)*	*P. SD (dB)*
Constant	14.4	-13.6	2.4	-14.0	0.8
Babble	10.3	-14.0	3.6	-14.9	1.9

'P.' refers to a predicted post-normalisation value, 'M.' Refers to a measured value.

The pre-normalisation (measured) test-specific slope ($s50_{test}$) for constant noise was calculated to be 0.108 dB^{-1} or 10.8%/dB. With the adjustment applied, the standard deviation is predicted to be 0.8 dB, a reduction from the measured 2.4 dB. The test-specific slope ($s50_{test}$) is predicted to increase to 0.139 dB^{-1} or 13.9%/dB, denoting a 3.1%/dB increase in slope with normalisation.

For babble noise, the pre-normalisation test-specific slope was 0.062 dB^{-1} or 6.2%/dB. With the adjustment, and the removal of "shirts", the standard deviation of word-specific L_{mid} was reduced from 5.0 dB to 3.6 dB, resulting in a measured pre-normalisation slope of 7.5%/dB. With the adjustment applied, the predicted standard deviation of L_{mid} was 1.9 dB. The predicted post-normalisation test-specific slope increased to 9.3%/dB; a 3.1%/dB increase in slope from the original measured data. A comparison of the test-specific parameters with published values will be provided in the discussion section (Chapter 4).

5.4 Part II. Generation of sentence lists

Based on the data provided in the previous section, 30 base lists of 20 sentences were generated for each noise condition, which will be used to obtain measured post-normalisation data. This data would be used to 1) evaluate the normalisation process carried out here (i.e. obtain the measured post-normalisation test-specific slope), 2) evaluate the performance of UCAMST in lists (i.e. with sentence scoring), and 3) evaluate the test re-test reliability of these lists.

The lists were constructed in Microsoft Excel using the word-specific slopes (s_{word}) generated from the word normalisation process and the predicted post-normalisation L_{mid} for each word. The average of the five word-specific slopes in each sentence was taken to generate the sentence-specific slope ($s_{entence}$). The predicted list-specific slope ($s50_{list}$) was calculated by probabilistic modelling using the mean slope across words (s_{word}), and predicted standard deviation of L_{mid} (σ_{Lmid}) in each list. The data was used to ensure lists were homogeneous in terms of 1) the distribution of sentence-specific within each list (i.e. similar minimum and maximum $s_{sentence}$ and $\sigma s_{sentence}$), and 2) the predicted list-specific slopes within each noise condition.

In order to maximise the sensitivity and reliability of the lists, the fragments and words which produced abnormal psychometric function (i.e. "wins", or 'nine_good') or required excessive adjustment ("shirts") were

avoided in generating the lists in the affected noise condition. The words "wins" and "shirts" were replaced at random with one of the other nine alternatives in that sentence position. Where possible, word substitutions were made to reduce the distribution of sentence-specific and list-specific slopes. All included sentences were unique in that no sentence was repeated both within, and between, noise conditions. Table 9 shows the proportion of total occurrences (n = 3000) contributed by each matrix word in the final base lists for constant and babble noise.

Table 9

Proportion (%) of matrix words in constant and babble noise lists.

Word	*Constant %*	*Babble %*	*Word*	*Constant %*	*Babble %*
amy	2.00	2.07	old	2.00	2.00
big	2.00	2.13	oscar	2.00	1.90
bikes	2.00	2.23	peter	2.00	2.03
books	2.00	2.27	rachel	2.00	2.10
bought	2.00	2.07	red	2.00	2.00
cheap	2.00	2.00	sees	2.00	2.40
coats	2.00	2.27	ships	2.00	2.20
dark	2.00	1.90	shirts	2.00	0.00
david	2.00	2.03	shoes	2.00	2.17
eight	2.00	2.00	six	2.00	2.07
four	2.00	2.20	small	2.00	1.97
gives	2.00	2.40	sold	2.00	2.73
good	2.00	1.87	some	2.00	2.10
got	2.00	2.17	sophie	2.00	2.07
green	2.00	2.00	spoons	2.00	2.20
hannah	2.00	1.77	ten	2.00	2.03
has	2.00	1.80	thomas	2.00	1.90
hats	2.00	2.23	those	2.00	1.93
kathy	2.00	2.07	three	2.00	2.00
kept	2.00	2.17	toys	2.00	2.23
large	2.00	2.03	twelve	2.00	1.97
likes	2.00	2.23	two	2.00	2.03
mugs	2.00	2.20	wants	2.00	2.03
new	2.00	2.10	william	2.00	2.07
nine	2.00	1.67	wins	2.00	0.00

Word frequencies were homogeneous in constant noise; there was slight variation in babble due to the exclusion of “wins” and “shirts”. Thus, unlike in constant noise, the predicted list-specific slopes ($s50_{list}$) were not equivalent to the test-specific slope ($s50_{test}$) in babble noise, necessitating the calculation of this value for each list. Table 10 and Table 11 display the

resulting descriptive statistics for finalised lists in the constant and babble noise conditions, respectively.

Table 10

Constant noise list descriptive statistics.

List	*Mean* s_{word} *(%/dB)*	$\sigma s_{sentence}$ *(%/dB)*	*Min.* $s_{sentence}$	*Max.* $s_{sentence}$	*P. Mean* L_{mid} *(dB SNR)*	*P.* σ_{Lmid} *(dB)*	*P.* $s50_{list}$ *(%/dB)*
1	14.4	1.3	12.0	17.0	-14.0	0.8	13.9
2	14.4	1.5	11.3	16.8	-14.0	0.8	13.9
3	14.4	1.4	11.3	17.2	-14.0	0.8	13.9
4	14.4	1.4	12.3	16.6	-14.0	0.8	13.9
5	14.4	1.3	12.3	17.2	-14.0	0.8	13.9
6	14.4	1.0	12.5	16.6	-14.0	0.8	13.9
7	14.4	0.9	12.6	16.7	-14.0	0.8	13.9
8	14.4	1.0	12.3	16.4	-14.0	0.8	13.9
9	14.4	1.0	11.9	16.2	-14.0	0.8	13.9
10	14.4	1.0	12.8	16.3	-14.0	0.8	13.9
11	14.4	1.6	11.4	17.3	-14.0	0.8	13.9
12	14.4	0.9	12.9	16.8	-14.0	0.8	13.9
13	14.4	1.7	11.5	17.7	-14.0	0.8	13.9
14	14.4	1.3	11.7	17.4	-14.0	0.8	13.9
15	14.4	1.3	12.0	17.0	-14.0	0.8	13.9
16	14.4	1.4	12.2	16.7	-14.0	0.8	13.9
17	14.4	1.4	12.0	17.0	-14.0	0.8	13.9
18	14.4	1.3	12.3	16.9	-14.0	0.8	13.9
19	14.4	1.0	12.9	16.2	-14.0	0.8	13.9
20	14.4	1.1	12.7	16.2	-14.0	0.8	13.9
21	14.4	1.2	12.0	16.4	-14.0	0.8	13.9
22	14.4	1.4	11.5	17.7	-14.0	0.8	13.9
23	14.4	1.2	12.0	16.9	-14.0	0.8	13.9
24	14.4	0.9	13.0	16.7	-14.0	0.8	13.9
25	14.4	1.2	12.1	17.7	-14.0	0.8	13.9
26	14.4	1.4	12.0	17.3	-14.0	0.8	13.9
27	14.4	1.3	12.5	17.0	-14.0	0.8	13.9
28	14.4	1.0	12.4	16.9	-14.0	0.8	13.9
29	14.4	1.2	12.2	16.9	-14.0	0.8	13.9
30	14.4	1.5	11.4	16.5	-14.0	0.8	13.9
Mean	**14.4**	**1.2**	**12.1**	**16.9**	**-14.0**	**0.8**	**13.9**
SD	**0.0**	**0.2**	**0.5**	**0.4**	**0.0**	**0.0**	**0.0**

P = predicted.

Table 11

Babble noise list descriptive statistics.

List	*Mean* s_{word} *(%/dB)*	$\sigma s_{sentence}$ *(%/dB)*	*Min.* $s_{sentence}$	*Max.* $s_{sentence}$	*P. Mean* L_{mid} *(dB SNR)*	*P.* σ_{Lmid} *(dB)*	*P.* $s50_{list}$ *(%/dB)*
1	10.3	1.2	7.4	12.2	-14.8	1.5	9.6
2	10.2	1.1	8.5	11.9	-14.8	1.8	9.2
3	10.2	1.1	7.9	12.6	-14.9	1.8	9.2
4	10.1	1.1	8.1	12.4	-14.9	1.8	9.2
5	10.2	1.0	8.1	11.9	-14.9	1.8	9.3
6	10.4	0.9	8.8	12.5	-14.8	1.6	9.6
7	10.4	1.2	8.5	12.9	-14.8	1.8	9.3
8	10.3	1.3	8.1	13.3	-14.8	1.5	9.6
9	10.2	0.9	9.0	12.2	-14.9	1.8	9.2
10	10.3	1.0	8.2	11.9	-14.8	1.6	9.5
11	10.3	1.0	8.6	12.9	-14.7	1.6	9.5
12	10.4	1.3	8.2	12.5	-14.9	1.8	9.4
13	10.3	1.1	7.6	12.0	-14.8	1.9	9.3
14	10.2	1.1	7.9	11.8	-14.8	1.6	9.4
15	10.4	1.3	8.2	12.5	-14.9	1.8	9.4
16	10.3	1.3	7.6	12.3	-14.9	1.9	9.2
17	10.2	0.9	8.6	12.3	-14.9	1.8	9.3
18	10.2	1.1	8.3	13.5	-14.9	1.8	9.3
19	10.3	1.4	8.0	13.6	-14.9	1.8	9.3
20	10.2	0.6	9.2	11.4	-14.9	1.7	9.3
21	10.3	0.9	8.1	11.6	-14.9	1.8	9.3
22	10.2	1.3	8.1	12.6	-14.9	1.8	9.2
23	10.4	1.0	8.2	12.8	-14.8	1.5	9.6
24	10.3	1.0	8.3	12.4	-14.8	1.9	9.2
25	10.3	1.1	8.1	11.9	-14.9	1.8	9.3
26	10.3	1.1	8.6	12.6	-14.8	1.5	9.6
27	10.3	0.9	8.7	11.8	-14.9	1.8	9.3
28	10.2	1.4	7.9	13.0	-14.9	1.8	9.3
29	10.5	1.2	8.0	12.3	-14.8	1.6	9.6
30	10.3	0.8	8.8	11.9	-14.9	1.8	9.3
Mean	**10.3**	**1.1**	**8.3**	**12.4**	**-14.9**	**1.7**	**9.4**
SD	**0.1**	**0.2**	**0.4**	**0.5**	**0.1**	**0.1**	**0.1**

P = predicted.

Comparisons between lists within each noise condition will be made first. As word-specific functions all had adequate fits in constant noise, each list contained exactly two occurrences of each matrix word. Thus, as lists are comprised of the same matrix word-specific slopes, the list-specific slope is predicted to be the same across lists (M = 13.9%/dB ± 0.0%/dB). The babble noise (M = 9.4%/dB) is expected to have slightly more variation in list-specific (± 0.1%/dB).

With regards to within-list comparisons, the mean distribution of sentence-specific slopes within each list are comparable (mean $\sigma s_{sentence}$ = 1.2%/dB in constant noise; mean $\sigma s_{sentence}$ = 1.1%/dB in babble noise). This was due to the larger proportion of words and word pairs that were excluded, necessitating the substitution of other words. The complete sentence lists can be found in Appendix B. The following section will compare the values obtained here with those of international MSTs.

5.5 Comparison of parameters with international MSTs

5.5.1 Comparison of test-specific slope ($s50_{test}$)

A comparison shall first be made between the measured pre-normalisation and predicted post-normalisation test-specific slopes of the UCAMST with those of published MSTs, as summarised in Table 12.

Table 12

Test-specific slopes ($s50_{test}$) based on measured pre-normalisation and predicted post-normalisation data across MSTs. Emboldened data is from this project (NZ English).

Language[6]	*Noise*[7]	*M. Pre-norm. $s50_{test}$ (%/dB)*	*P. Post-norm. $s50_{test}$ (%/dB)*	*M. Post-norm. $s50_{test}$ (%/dB)*
Danish	SS	8.7	13.2	13.2
Polish	B	13.9	18.2	17.1
Spanish	SS	10.9	16.0	13.2 (open-set) 14.0 (closed-set)
Italian	SS	9.2	15.2	13.3
NZ English	**SS**	**10.8**	**13.9**	*****
	B	**6.2**	**9.3**	*****

M. = Measured, P. = Predicted.

* To be confirmed by follow-on research.

The measured pre-normalisation data will be examined first. The test-specific slope in constant noise (10.8%/dB) was steeper than those of the Danish (8.7%/dB) and Italian (9.2%/dB) MSTs. On the other hand, the test-specific slope for the babble noise (6.2%/dB) was lower than all published values, reflecting the larger distribution of word-specific L_{mid} measures. A small

[6] Author key: Danish (Wagener et al., 2003), Polish (Ozimek et al., 2010), Spanish (Hochmuth et al., 2012), and Italian (Puglisi et al., 2014).

[7] Noise key: SS = Speech-shaped noise, B = Babble noise.

increase to this slope (1.3%/dB) was noted with the removal of "shirts" from the base matrix.

Turning now to predicted values, the test-specific slope in constant noise (13.9%/dB) is comparable with the measured post-normalisation test-specific slopes of the Italian (13.3%/dB), Danish (13.2%/dB), and Spanish (14.0%/dB) MSTs; however, this is lower than the predicted values of these MSTs with the exception of the Danish version (13.2%/dB). The predicted test-specific slope for babble noise (9.3%/dB) is lower than both predicted and measured post-normalisation slopes listed in Table 12. Of particular relevance is a comparison with the Polish MST (18.2%/dB and 17.1%/dB, respectively), which also used babble noise.

5.5.2 Comparison of predicted list values

A comparison will also be drawn between the predicted list values offered in this project, and measured list values of international MSTs. The test-specific slope assumes equal frequencies of included data, and differs from list-specific slope in the current project in that babble lists used slightly heterogeneous word frequencies due to word exclusions. List-specific data also provided inter-list variability, and the mean L_{mid} for each test will also be discussed here. A summary of list-specific and mean L_{mid} data is provided in Table 13.

Table 13

Predicted post-normalisation list-specific parameters ($s50_{list}$) of NZ English (emboldened) in comparison with measured post-normalisation parameters of international MSTs.

Author(s)	Language	Noise	Mean SRT (dB SNR)	Mean $s50_{list}$ (%/dB)
Wagener et al. (2003)	Danish	SS	-8.4 ± 0.2	12.6 ± 0.8
Ozimek et al. (2010)	Polish	B	-9.6 ± 0.2	17.1 ± 1.5
Hochmuth et al. (2012)	Spanish	SS	-6.8 ± 0.2	13.1 ± n.a.
Houben et al. (2014)	Dutch	SS	-8.4 ± 0.2	10.2 ± 0.9
Dietz (2014)	Finnish	SS	-10.1 ± 0.1	16.7 ± 1.2
Puglisi et al. (2014)	Italian	SS	-7.3 ± 0.2	13.3 ±1.2
Current	**NZ English**	**SS**	**-14.0 ± 0.0***	**13.9 ± 0.0****
Current	**NZ English**	**B**	**-14.9 ± 0.1***	**9.4 ± 0.1****

* predicted based on applied adjustments.

** predicted based on probabilistic modelling (see equation 3).

The mean list-specific slope for constant noise is predicted to be steeper than the measured slope equivalents for the Danish, Spanish, Dutch, and Italian MSTs. The measured mean list-specific slope of the Finnish version is 2.3%/dB steeper than the predicted UCAMST function. With regard to babble noise comparisons, the predicted list-specific slope is 6.8%/dB shallower than the Polish version. In terms of difficulty (mean L_{mid}), the NZ English version is predicted to be detectable at lower SNRs than the other MSTs. This is because, as mentioned earlier in this chapter, performance was measured binaurally, as opposed to monaurally.

Chapter 4

Discussion

9.1 Study 1: Noticeability of video judders

9.1.1 Overview

"Judders", or image jerks, were a by-product of the complex editing process in which unedited original sentences were cut and reconstructed to create so-called "synthesised" sentences (Trounson, 2012). As synthesised sentences comprise the majority of possible auditory-visual sentences, it was necessary to establish a subjective index of how "noticeable" the judders were compared to the unedited "original" sentences from which they were constructed. This constituted the main goal of Study 1. Based on this data, sentences with minimal noticeable judder were selected for inclusion in the auditory-visual UCAMST. The following section will provide a commentary relating the results of Study 1 to the initial research questions.

9.1.2 Synthesised sentences vs. no judder sentences

It was initially hypothesised that the mean rating score of synthesised and no judder sentences would significantly differ. This initial prediction was confirmed, as the synthesised sentences were assigned, on average,

significantly higher rating scores than no judder sentences. The no judder sentences were assigned a mean pixel difference value (M = 0.59) that approximated '0' ("no noticeable judder"). This observation suggests that, in general, the requirements of the task were understood, as participants were correctly assigning low rating scores to sentences without judders. Why the mean rating score for this condition did not better approximate 0 is unknown; there are, however, two potential explanations. Firstly, imprecise placement of the cursor on the rating scale may have resulted in a rating score response that did not represent the intended response. Secondly, experimental evidence has shown individual response styles to differ; some participants respond with the extreme values of the scale, whereas some favour more central values. This response style may be related to personality factors (Naemi, Beal, & Payne, 2009). A tendency towards central values may have predominated in Study 1.

A closer analysis of the results in Study 1 reveals how small the relative difference in judder severity is between sentences with no judder and sentences with Tier 2 judder. Tier 2 sentences had pixel difference values between 300,000 to 400,000, which represented a 0.38% (300,000/78,643,200 x 100) to 0.51% (400,000/78,643,200 x 100) absolute change in pixel difference value between transitions, compared to 0.23% to 0.31% for the no judder sentences. A comparison of the absolute change in pixel difference value suggests that small disruptions to the smoothness of video transition translate to a large difference in the subjective measure of noticeability (M =

0.59 vs. M = 3.00). This relationship will be further elucidated with regards to the second aim of Study 1, as detailed in the following section.

9.1.3 Relationship between pixel difference value and rating score

The second aim of Study 1 was to investigate the relationship between the average pixel difference value and rating score. The findings were congruent with the initial prediction: the regression model confirmed that average pixel difference value was a significant predictor of rating score. The relationship constituted a large effect size (Pearson's r = 0.64)[8], and average pixel difference value accounted for approximately 41% of the variation in rating score. This represented a slightly larger proportion of variance accounted for than by judder number (38.1%); thus, judder severity was slightly more influential of rating score than the number of judders per sentence. Combined, these predictors accounted for almost half (48%) of the total variance in rating score. The 52% variation unexplained may be due to a number of factors; one potential source of variation is individual differences in response style, as discussed in the previous section. For example, participant 9 assigned a maximum rating score of 3.01, whereas other participants provided a greater range of rating scores across the 10-point scale. The threshold for what each participant deemed "noticeable" may depend on personality related

[8] Based on Cohen's rule of thumb (see Cohen, 2003).

factors, such as decisiveness (Naemi et al., 2009). Other individual differences, notably the acuity of the individual's visual system, and factors related to attention or fatigue, may also have influenced the rating scores. It is also possible that, because the average pixel difference value accounted for change in head position only, the unexplained variance is due to perceived noticeable judder in other regions of the image (i.e. the actress's mouth and eyes).

The model equation describing the relationship between average pixel difference value, which was the best predictor, and rating score may be utilised as an indication of the level of judder severity that will result in the requisite level of noticeable judder. This model may be of use in the development of future auditory-visual MSTs. However, the accuracy of this method is limited as it is based on averages, and therefore should be used as a guide only.

9.2 Sentences for inclusion in the auditory-visual UCAMST

As a whole, the evidence from Study 1 warranted a compromise between quality (i.e. the "noticeability" of the judders) and quantity (i.e. the number of usable sentences) in the selection of sentences for the auditory-visual UCAMST. This compromise was reached by choosing synthesised sentences with minimal noticeable judder. It was decided that Tier 2 sentences (M= 3.31) had an acceptable level of judder that would be of minimal distraction during UCAMST testing. The Tier 2 limit elucidated a repertoire of

2,494 sentences for inclusion in the auditory-visual UCAMST, including no judder sentences, which were sourced from the pool of sentences used to generate test materials for Study 1. Preferential selection of sentences with no judder (n = 1233) is advised to maintain the least noticeable judder overall. Interestingly, when the "Tier 2" criterion was applied to the 3,000 transitions, fewer transition 2 fragment combinations (n = 521) met this criterion than transition 1 (n = 890) and transition 3 (n = 872). Thus, the pool of candidate sentences was limited by the fewer acceptable transition 2 fragment combinations.

9.3 Study 1: Limitations and future directions

In summary, the main goal of Study 1 was to establish a collection of sentences for use in the auditory-visual UCAMST. Based on these results, sentences with minimal noticeable judder were selected as a means of increasing the number of available sentences. It is important to note that the conditions under which participants performed Study 1 may limit the applicability of the results. Participants were exclusively instructed to rate judders, resulting in active attendance to the judders. By contrast, a naive client performing the UCAMST may not notice the judders—in particular those with low pixel difference values (Tier 2 or 3)—as attention will, at least in part, be focused on interpreting the sentences. This seems a compelling

reason to argue that the rating scores obtained here may be liberal estimates of how noticeable the judders are during performance of the UCAMST.

The video judders were an unfortunate by-product of the complex editing process and were largely attributed to changes in head position across the transitions (Trounson, 2012). This endures as a limitation of the current project, which has revealed on average a significant "noticeability" of these artefacts compared unedited no judder sentences. Although a large a sample of candidate sentences for the auditory-visual UCAMST were provided (n = 2494), approximately one-half of this sample will exhibit one or two slight judders (n = 1261). In the unlikely event that a larger sample size of auditory-visual sentences is desired, one alternative would be to re-record sentences using measures that better support the actress's head and neck (Trounson, 2012). For example, in recording the Malay version of the UCAMST (Jamaluddin & O'Beirne, *in progress*), a plaster cast was used to hold the speaker's head in place underneath her headscarf. Head mounted cameras may also be considered as a means of affixing head position on the lens. Nonetheless, it is expected the sample of sentences produced here is sufficiently large, and low in noticeable judder, to allow for the continued use of these recordings. The next logical step for future research is to compile equivalent lists from these recommended sentences, which will constitute materials for auditory-visual testing.

9.4 Study 2: Normalisation of the auditory-alone UCAMST

9.4.1 Overview

Study 2 was comprised of two complementary phases; each carried out in both constant speech-shaped and six-talker babble noise. The aim of Phase I was to normalise the difficulty of matrix words and fragments by adjusting individual word-specific L_{mid} measures to equal the mean pre-normalisation L_{mid} across fragments. It was predicted that, based on pre-normalisation data, the participants would be more sensitive to constant noise ($s50_{test}$), and would also find the UCAMST more difficult (mean L_{mid}), compared to the babble noise condition. We also hypothesised that the normalisation process would result in a predicted increase to the test-specific slope. Due to the unique way in which the audio component mapped onto fragments, data could be normalised by fragment or by word; both methods were incorporated in the development of sentence lists in Phase II. In this second phase, 30 lists of 20 sentences were generated in each noise condition. The sentences were homogeneous in terms of: 1) the overall proportion of word occurrences within each noise condition, 2) the within-list distribution of sentence-specific slopes ($s_{sentence}$), and 3) the distribution of predicted list-specific slopes ($s50_{list}$). It is important to note that the mean L_{mid} measures observed in this study are lower than published values as materials were presented binaurally. This section will discuss the results of Study 2 in a chronological fashion.

9.4.2 The effect of masking noise on performance

Firstly, the influence of masking noise on performance of the UCAMST in terms of the difficulty (L_{mid}) and sensitivity ($s50_{test}$) will be considered. As predicted, the pre-normalisation data confirmed that participants found it slightly harder to detect words in constant noise (L_{mid} = -14.2 dB SNR) than in babble noise (L_{mid} = -14.9 dB SNR). The direction of this difference is consistent with what is expected with masking release, in that listeners take advantage of momentary dips in the envelope to detect the target signal (Peters et al., 1998; Wagener & Brand, 2005). However, the relative separation of the mean L_{mid} measures in this study is small (0.7 dB). One possible reason for this is the number of talkers in the masker, which has been shown to influence the masking capability of babble. For example, inWagener and Brand (2005), normal-hearing listeners performed significantly better in a one-talker babble (SRT = -21.6 dB SNR) compared to a six-talker babble (SRT = -9.9 dB SNR). Furthermore, a 2.7 dB difference separated the mean listener SRT in six-talker babble from that in constant speech-shaped noise (-6.2 dB SNR). The comparatively smaller separation observed in the UCAMST data (0.7 dB) may be attributed to the use of a closed-set presentation format. Elsewhere, Simpson and Cooke (2005) found that gradually increasing the number of talkers non-monotonically decreased the intelligibility of consonants. The result was ascribed to a number of factors, including the increasingly more non-stationary (i.e. non-fluctuating) nature of

babble maskers with a large number of talkers (see Simpson & Cooke, 2005, for more detail). In summary, listener performance on the UCAMST is not expected to vary greatly between the constant and babble noise conditions.

With regards to sensitivity, the constant noise had a steeper measured test-specific slope than the babble noise, which is consistent with both the initial hypothesis and the experimental evidence comparing babble and speech-shaped noises (Francart, 2011; Wagener & Brand, 2005). The utility of each noise type in the UCAMST will be detailed in the following section in the context of the predicted post-normalisation data.

9.4.3 Normalisation of the UCAMST

The second aim of Study 2 was to normalise the difficulty of the matrix words in both constant and babble noise. The efficacy of this process was calculated using a probabilistic model (equation 3, Chapter 3) described by Kollmeier (1990). If the level adjustment made to pre-normalisation word-specific L_{mid} measures effectively decreases the distribution of post-normalisation word-specific L_{mid} measures (σ_{Lmid}), a large increase in the test-specific slope ($s50_{test}$) will be observed. Indeed, consistent with the initial hypothesis, an increase in the predicted test-specific slope in both constant (3.1%/dB) and babble noise (3.1%/dB) was observed with probabilistic modelling. This indicates that the adjustments made to pre-normalisation data

resulted in a theoretical decrease in the distribution of L_{mid} measures, homogenising the difficulty of the matrix words.

In comparing the two masking noises, the constant noise was predicted to have a steeper test-specific slope (13.9%/dB) than the babble (9.3%/dB). This has implications for the selection of masking noise to use during UCAMST testing in both clinical and research contexts. Generally a steep slope is desirable as it denotes the reliability of the test, and in effect, the accuracy of the client's SRT (Ozimek et al., 2010). In clinical practice, the audiologist has a limited amount of time to complete a large battery of audiological tests, which are typically repeated at follow-up appointments. Therefore, a test that can quickly and accurately estimate the SRT would be a valuable addition to the audiological test battery, as it could potentially save time. In contrast, babble noise may be considered to have higher face-validity, as it simulates everyday contexts where the masker is largely speech (Killion et al., 2004; Plomp, 1978). This may give the audiologist a better idea of what benefit the rehabilitation plan would have to the individual's life. The UCAMST enables a choice between two types of noise based on the specific goals of rehabilitation or research. However, before firmer conclusions can be drawn on the sensitivity of the UCAMST in either noise type, *measured* post-normalisation values are required to confirm the efficacy of the adjustments

made in this project. Future considerations for the normalisation of the UCAMST are detailed in section 9.6.

9.4.4 Comparison of test-specific slopes with international MSTs

The results section comprised a comparison of UCAMST parameters with those of international MSTs (section 3.5). In general, the constant noise condition produced predicted test-specific and list-specific parameters comparable with the majority of international MSTs, which is consistent with the initial hypothesis. However, contrary to the same hypothesis, the babble noise condition had a much lower predicted slope value (9.3%/dB) when compared to the Polish version (17.1%/dB), which also used babble noise. As discussed previously, a comparatively shallow slope is expected when measuring performance in a fluctuating masker as compared to a speech-shaped noise (Francart, 2011; Wagener & Brand, 2005). The slope estimation in the babble condition may also have been affected by the small sample size used here (n = 8) compared to the Polish version (n = 30) (Ozimek et al., 2010). Further discussion of sample size is provided in section 9.5 ("Study 2: Limitations").

Regarding cross-language comparison, an advantage of the MST format is that it permits comparability across languages, allowing for listener performance to be compared internationally (Zokoll et al., 2013). However, the test-specific slope may be influenced by the unique speech qualities of

each language, such as speed and articulation (Houben et al., 2014). This view is held by Zokoll et al. (2013), who noted a similarity in reference slopes between Polish and Russian MSTs, both of which are Slavic languages. In a similar manner, the unique characteristics of NZ English may affect the parameters of the psychometric function. For example, its rhythm has been found to differ from other forms of English due to vowel duration (Nokes & Hay, 2012). Slight variations between languages may, therefore, result from language-specific factors.

9.4.5 Homogeneity of test lists

With regards to inter-list variation in predicted list-specific slopes, the constant noise was predicted to be homogeneous, as the same word compositions were used in each list (13.9%/dB ± 0.0%/dB). However, word exclusions in the babble condition resulted in slightly different word compositions between lists. In spite of this, there is little predicted inter-list variation in slopes, with a mean list-specific slope of 9.4%/dB and standard deviation of 0.1%/dB. As these are *predicted* values, one should interpret them as such; this will be discussed further in conjunction with plans for evaluation of the UCAMST in section 9.6 (“Study 2: Future research”).

9.5 Study 2: Limitations

9.5.1 Sample size

The current project was limited by the small sample size, which particularly affected data collection in the babble noise condition. The small sample size was attributed to time constraints that arose due to delays in the development of the normalisation software. Follow-on work may consider the recruitment of further participants, particularly in the babble noise condition, to improve the accuracy of slope and L_{mid} estimates.

9.5.2 Data exclusion

Related in part to sample size, one main drawback of the current study was the large proportion of data excluded in Study 2. In Phase I, a small proportion of fragments failed to produce adequate function fits with logistic modelling (see equation 1) due to 1) floor effects (i.e. SNR too low), 2) ceiling effects (i.e. SNR too high), or 3) unreliable or inconsistent performance across presented SNRs. Affected fragments were not included in Phase II to maximise the sensitivity, and therefore reliability, of the sentence lists. The attrition of usable data was expected: poor function fits have been noted in published MSTs using the logistic model. In brief, the Spanish version saw eight out of 500 (1.6%) word realisations owing to poor model fits (Hochmuth et al., 2012), and the Finnish version saw 15 out of 500 (3%) removed due to exceeding the adjustment limit by a further 2 dB (Dietz, 2014). The relatively

large proportion of fragments excluded in this project may be ascribed to a comparatively smaller sample size, as discussed in section 9.5.1. As each participant contributed 4 data points to each file fragment (across the four presented SNRs), each fragment garnered 36 data points in constant noise, and 32 data points in babble noise. The difference in sample size may therefore have contributed, at least in part, to the threefold larger proportion of data removed from the babble condition (12%, n = 8) compared to the constant noise condition (4%, n = 9). Increasing the sample size may improve the fit of the function by minimising the proportion of data that would warrant removal.

The relative ease of the babble noise condition compared to constant noise, as discussed earlier in this chapter in a comparison of L_{mid} measures (see section 9.4.2), may also be associated with the larger proportion of data excluded from this condition. The dominant reason for fragment exclusion from the constant noise condition was ceiling effects (n = 12); thus, as detection of matrix words in the babble condition was, on average, easier than in constant noise, one may expect the proportion of file fragments "hitting the ceiling" of the measurement range to increase. As 24 fragments warranted exclusion from the babble condition due to ceiling effects, this seems a likely rationale. In practice, floor and ceiling effects can be avoided by broadening the range of SNRs, thus essentially extending the "floor" and "ceiling" of the measurement range.

Alternatively, the inability to fit psychometric functions may be the product of inherent word-specific characteristics, and in this manner, impervious to both the type of noise and sample size. The trends observable in excluded fragments propound this view. A large proportion of fragments containing the word "nine" were excluded from both the constant (n = 4) and babble (n = 7) noise conditions due to a ceiling effect. Descriptive analyses in Study 2 found—regardless of noise type—the 'name' and 'number' words were on average the easiest to detect, and therefore, at greater risk of ceiling effects. This suggests that there is a relationship between the grammatical role of the word and the rate of exclusion (due to ceiling and floor effects). Another relevant word-specific characteristic is its phonemic composition. Liu and Eddins (2008) provide confirmatory evidence that certain vowels in American English had lower thresholds in LTSS noise. The phoneme /I/ (i.e. "wins") had a high detection threshold compared to the 11 other vowel sounds included in the study. The authors attributed differences in detection thresholds to the specific vowel spectra. It is quite possible that—particularly at unfavourable SNRs—the participants based word selection on detection of the vowel component. Similarly, word selection may have been influenced by the phonemic composition of *other* words in the matrix. For example, the word "shirts" was removed from the babble condition as it was very difficult for the participants to detect (L_{mid} = 9.7 dB SNR). "Shirts" also proved difficult to detect in constant noise, although not quite to the same extent (L_{mid} = -8.0 dB

SNR). The reason for this apparent difficulty could be due to a confusion of "shirts" with "ships", which both contain the /sh/ sound. In addition to phonemic composition, other contributing factors worthy of consideration are word frequency and consonant sound.

9.6 Study 2: Future research

9.6.1 Evaluation of lists

Study 2 provided predicted post-normalisation parameters based on adjustments made to measured data, and we emphasise that these should be interpreted as predicted estimates until further validation of these parameters can be ascertained. The lists generated in this project should be subjected to evaluation to establish: 1) the efficacy of the normalisation procedure carried out here, by comparing predicted and measured post-normalisation test-specific slopes, and 2) test re-test reliability of the generated lists. Evaluating the lists in these ways would constitute an important step in establishing the UCAMST as a reliable measure of SRT, particularly when repeated testing (and hence, multiple test lists) is warranted.

Proposed here is a method for future evaluation of this normalisation process, which aligns with the methods used in existing international MSTs (i.e., Hochmuth et al., 2012; Ozimek et al., 2010). A large sample of normal-hearing participants (n = 10-20) should assess a small sample of lists (i.e. five)

at two different SNRs that approximate the "pair of compromise" (i.e. $p_1 = .19$ and $p_2 = .81$). This method estimates the list-specific SRT and slope in a quick and efficient manner using an adaptive procedure (see Brand & Kollmeier, 2002, for further details). A comparison of the predicted and measured post-normalisation data, specifically the test-specific specific slope, will be necessary to confirm the efficacy of the adjustments made in this project. A comparison of the inter-list and inter-subject standard deviations of L_{mid} measures would indicate the test re-test reliability of the lists; if the inter-list measure is smaller than the inter-subject measure—that is, the differences between lists is smaller than between subjects—the lists may be used inter-changeably (Dietz, 2014; Houben et al., 2014). On the whole, one can expect post-normalisation parameters to show some departure from these predicted values—in particular, the distribution of list-specific slopes in constant (± 0.0%/dB) and babble (0.1%/dB) noise. These values will naturally be subject to novel sources of variation, such as between-subjects differences. Measured test-specific slopes in this post-normalisation phase will provide reference values for the published version of the auditory-alone UCAMST.

9.6.2 Word normalisation vs. fragment normalisation

An important goal for follow-on work is to validate the "word normalisation" method employed in this project. In brief, word normalisation

involved averaging performance across all realisations of each word (n = 10) to establish word-specific intelligibility functions. However, this approach negated the effect of context on word difficulty. This was not a challenge encountered by published MSTs, as each word realisation of the word mapped onto a single file fragment (i.e., Wagener et al., 2003). If, indeed, context had an effect on word difficulty, the post-normalisation distribution of word-specific L_{mid} measures will not reduce to match the predicted value. The result will be a measured test-specific slope that is lower than the predicted equivalent. If this occurs, "fragment normalisation" should be considered. One drawback of this approach, however, is that it disregards the fact that the audio components that comprise a single word may map onto more than one file fragment. For example, the audio of the word "small" in "those small toys" maps onto two fragments: 'those_s____' and '_mall_toys'. If the former fragment required a level decrease, while the latter fragment required a level increase, an unnatural "level jump" would result at the edited transition ("s-mall"). Fragment normalisation should therefore be piloted prior to its full implementation.

9.6.3 Adjustment limit

The adjustment limit used in this project may also be worthy of future consideration. A ± 3 dB limit was selected as it represented the limit used in the majority of published MSTs, which have confirmed the naturalness of this

limit (i.e., Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek et al., 2012). The Danish version, on the other hand, used a more relaxed limit of ± 4 dB (Wagener et al., 2003). According to Ozimek et al. (2010), the larger adjustment limit used in the Danish version resulted in equal predicted and measured post-normalisation test-specific slopes (both 13.2%/dB). This is because a more conservative adjustment limit will result in a failure to align the word-specific SRTs with the pre-normalisation mean SRT, as evident in published MSTs which used a ± 3 dB limit (i.e., Hochmuth et al., 2012; Ozimek et al., 2010; Puglisi et al., 2014). In Study 2, there were a large proportion of words that exceeded the adjustment limit in both constant (14%) and babble (41%) noise conditions. If a ± 4 dB limit was applied, however, the proportion of words exceeding this limit is greatly reduced in the babble condition (22%), and slightly reduced in the constant noise condition (12%). This would likely increase the test-specific slope, to a greater extent in the babble condition. However, care should be taken when applying the adjustment limit; one that is too liberal may cause an unnatural level jump between words, particularly if consecutive words are adjusted in opposite directions (i.e. one increased in presentation level, one decreased in presentation level). Nonetheless, as a means of increasing the sensitivity of the test, incorporation of a ± 4 dB limit could be fruitfully investigated by future research. Based on the limit used in this project and by published MSTs (i.e., Dietz, 2014; Hochmuth et al., 2012; Houben et al., 2014; Ozimek et al., 2010;

Puglisi et al., 2014), it is tentatively suggested that measured post-normalisation test-specific slopes will not equate to the same level of sensitivity indicated by the predicted values.

9.6.4 Piloting with hearing-impaired individuals

In addition to participants with normal hearing, the auditory-alone UCAMST should also be assessed by individuals with hearing impairment. The rationale for this is to establish reference criteria by which normal hearing and hearing impairment can be categorised. This should be carried out in both noise types independently as, due to the differences in L_{mid} observed here and in previous work (Peters et al., 1998; Wagener & Brand, 2005), babble noise will likely require poorer performance to merit concern.

9.7 Conclusion

Speech audiometry is an important component of the audiological test battery, providing information essential to both the diagnosis and rehabilitation of hearing loss. The UCAMST presents numerous advantages over the traditionally used method of speech audiometry: it is automatic, easy to administer, uses higher validity sentence stimuli, and has the potential to generate an unlimited repertoire of speech materials. The uniform format of the MST also allows for comparability across languages. Before the UCAMST can be incorporated into clinical and research use, two studies were warranted

to validate and generate materials for auditory-visual and auditory-alone testing.

In Study 1, subjective ratings of judder noticeability were obtained which, combined with an objective measure of judder severity, generated a pool of sentences with minimal noticeable judder from which materials for auditory-visual testing can be drawn. In practice, minimising the "noticeability" of judder is crucial to ensuring the video component of the UCAMST appears natural and without noticeable distractions. The average pixel difference value had a large effect on rating score, thus may be of utility in predicting noticeable judder in future work.

Study 2 involved the normalisation of the auditory-alone UCAMST to ensure equal difficulty of test materials in both constant speech-shaped noise and six-talker babble. The predicted post-normalisation test-specific slopes are comparable with the measured equivalents of published international MSTs, with the exception of the babble noise condition; therefore, further development of this condition, in the form of a larger sample size, should be considered. Additionally, follow-on work should validate the adjustments made here to ensure predicted post-normalisation slope values align with the measured post-normalisations values.

In sum, the current project constituted the second instalment in a series of projects that will ultimately result in the inclusion of the UCAMST into the

UCAST platform, which comprises a battery of audiological speech tests for use in NZ clinical and research contexts (O'Beirne et al., 2012).

References

Acar, B., Yurekli, M. F., Babademez, M. A., Karabulut, H., & Karasen, R. M. (2011). Effects of hearing aids on cognitive functions and depressive signs in elderly people. *Archives of Gerontology and Geriatrics, 52*(3), 250-252. doi: http://dx.doi.org/10.1016/j.archger.2010.04.013

Arlinger, S. (2003). Negative consequences of uncorrected hearing loss-a review. *International journal of audiology, 42*, 2S17-12S20.

Ashmore, J. (2008). Cochlear outer hair cell motility. *Physiological Reviews, 88*(1), 173-210. doi: 10.1152/physrev.00044.2006

Boothroyd, A. (2008). The performance/intensity function: an underused resource. *Ear and hearing, 29*(4), 479-491. doi: 10.1097/AUD.0b013e318174f067

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America, 84*(1), 101-114.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America, 111*(6), 2801-2810.

Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, N.J: L. Erlbaum Associates.

Dalton, D. S., Cruickshanks, K. J., Klein, B. E., Klein, R., Wiley, T. L., & Nondahl, D. M. (2003). The impact of hearing loss on quality of life in older adults. *The Gerontologist, 43*(5), 661-668.

Darwin, C. J., & Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America, 107*(2), 970. doi: 10.1121/1.428278

Dietz, A. (2014). The development and evaluation of the Finnish Matrix Sentence Test for speech intelligibility assessment. *Acta Oto-Laryngologica, 134*(7), 728-737. doi: 10.3109/00016489.2014.898185

Donkelaar, H. J., & Kaga, K. (2011). The Auditory SystemClinical Neuroanatomy: Brain Circuitry and Its Disorders, Springer.

Francart, T. (2011). Comparison of fluctuating maskers for speech recognition tests. *International Journal of Audiology, 50*(1), 2-13. doi: 10.3109/14992027.2010.505582

Garstecki, D., & Erler, S. (2009). Management of Adults with Hearing Loss. In J. Katz, L. Medwetsky, R. Burkard & L. Hood (Eds.), *Handbook of Clinical Audiology* (6th ed.). Baltimore: Lippincott Williams & Wilkins

Gates, G. A., & Mills, J. H. (2005). Presbycusis. *The Lancet, 366*(9491), 1111-1120. doi: 10.1016/S0140-6736(05)67423-5

Gilchrist, J. M., Jerwood, D., & Ismaiel, H. S. (2005). Comparing and unifying slope estimates across psychometric function models. *Perception & psychophysics, 67*(7), 1289-1303. doi: 10.3758/BF03193560

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons, reprinted in 1988 by Peninsula Publishing, Los Altos, CA.

Greville, K. (2005). *Hearing impaired and deaf people in New Zealand: an update on population numbers and characteristics*: Oticon Foundation in New Zealand.

Gulya, A. J., Glasscock, M. E., Minor, L. B., & Poe, D. (2010). *Glasscock-Shambaugh's Surgery of the Ear* (6th ed.): People's Medical Publishing House-USA.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian audiology, 11*(2), 79-87.

Hällgren, M., Larsby, B., & Arlinger, S. (2006). A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition: Una versión sueca de la Prueba de Audición en Ruido (HINT) para evaluar el reconocimiento del lenguaje. *International Journal of Audiology, 45*(4), 227-237.

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., & Kollmeier, B. (2012). A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International journal of audiology*, *51*(7), 536-544.

Hope, R. V. (2010). *Towards the Development of the New Zealand Hearing in Noise Test (NZHINT).* (MAud), The University of Canterbury.

Hopkins, K., & Moore, B. C. J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *The Journal of the Acoustical Society of America, 125*(1), 442-446. doi: 10.1121/1.3037233

Houben, R., Koopman, J., Luts, H., Wagener, K. C., van Wieringen, A., Verschuure, H., & Dreschler, W. A. (2014). Development of a Dutch matrix sentence test to assess speech intelligibility in noise. *International Journal of Audiology*, *53*(10), 760-763. doi: doi:10.3109/14992027.2014.920111

Katz, J. (2009). *Handbook of Clinical Audiology* (6th ed.). Baltimore: Lippincott Williams & Wilkins.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 116*(4), 2395-2405.

King, S. M. (2010). *Development and evaluation of a New Zealand Digit Triplet Test for auditory screening: a thesis submitted in partial fulfilment of the requirements for the degree of Master of Audiology in the University of Canterbury.* (Dissertation/Thesis). Retrieved from http://hdl.handle.net/10092/5679

Ko, J. (2010). Presbycusis and its management. *British Journal of Nursing*, *19*(3), 160-165.

Kochkin, S. (2007). MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids. *The Hearing Journal, 60*(4), 24-51.

Kollmeier, B. (1990). *Messmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache (in German). (Methodology, modeling, and improvement of speech intelligibility measurements).* (Habilitation), University of Göttingen, Göttingen.

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America, 102*(4), 2412-2421.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America, 49*(2B), 467-477.

Li, C.-M., Zhang, X., Hoffman, H. J., Cotch, M. F., Themann, C. L., & Wilson, M. R. (2014). Hearing Impairment Associated With Depression in US Adults, National Health and Nutrition Examination Survey 2005-2010. *JAMA otolaryngology--head & neck surgery, 140*(4), 293.

Liu, C., & Eddins, D. A. (2008). Categorical dependence of vowel detection in long-term speech-shaped noise. *The Journal of the Acoustical Society of America, 123*(6), 4539-4546. doi: 10.1121/1.2903867

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, 103*(49), 18866-18869.

Mathers, C., Smith, A., & Concha, M. (2000). Global burden of hearing loss in the year 2000. *Global burden of Disease, 18*, 1-30.

Mattheyses, W., Latacz, L., & Verhelst, W. (2009). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing, 2009*, 1.

McArdle, R. A., Wilson, R. H., & Burks, C. A. (2005). Speech recognition in multitalker babble using digits, words, and sentences. *Journal of the American Academy of Audiology, 16*(9), 726-739.

Mendel, L. L. (2008). Current considerations in pediatric speech audiometry. *International journal of audiology, 47*(9), 546-553.

Moncur, J. P., & Dirks, D. (1967). Binaural and Monaural Speech Intelligibility in Reverberation. *Journal of speech and hearing research, 10*(2), 186-195.

Moore, B. C. J. (2008). The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People. *JARO: Journal of the Association for Research in Otolaryngology, 9*(4), 399-406. doi: 10.1007/s10162-008-0143-x

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of personality, 77*(1), 261-286. doi: 10.1111/j.1467-6494.2008.00545.x

Napierala, M. A. (2012). What is the Bonferroni correction? *AAOS Now*, 40.

Newman, C. W., & Sandridge, S. A. (2004). Hearing loss is often undiscovered, but screening is easy. *Cleveland clinic journal of Medicine, 71*(3), 225-232.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085-1099.

Nokes, J., & Hay, J. (2012). Acoustic Correlates of Rhythm in New Zealand English: A Diachronic Study. *Language Variation and Change, 24*(1), 1. doi: 10.1017/S0954394512000051

O'Beirne, G. A., McGaffin, A. J., & Rickard, N. A. (2012). Development of an adaptive low-pass filtered speech test for the identification of auditory processing disorders. *International journal of pediatric otorhinolaryngology, 76*(6), 777-782.

Ozimek, E., Kutzner, D., & Libiszewski, P. (2012). Speech intelligibility tested by the Pediatric Matrix Sentence test in 3–6year old children. *Speech Communication, 54*(10), 1121-1131.

Ozimek, E., Warzybok, A., & Kutzner, D. (2010). Polish sentence matrix test for speech intelligibility measurement in noise. *International journal of audiology, 49*(6), 444-454.

Patuzzi, R. (2009). Cochlear Mechanics *Encyclopedia of Neuroscience* (pp. 1041-1049). Oxford: Academic Press

Peters, R. W., Moore, B. C., & Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America, 103*(1), 577-587. doi: 10.1121/1.421128

Pickles, J. O. (2012). *An Introduction to the Physiology of Hearing* (4th ed.). United Kingdom: Emerald Group Publishing Limited.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical Society of America, 63*(2), 533. doi: 10.1121/1.381753

Plomp, R., & Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology, 18*(1), 43-52.

Porter, H. L., Grantham, D. W., Ashmead, D. H., & Tharpe, A. M. (2014). Binaural Masking Release in Children With Down Syndrome. *Ear and hearing, 35*(4), e134-e142. doi: 10.1097/AUD.0000000000000026

Puglisi, G. E., Warzybok, A., Hochmuth, S., Astolfi, A., Prodi, N., Visentin, C., & Kollmeier, B. (2014). *Construction and first evaluation of the Italian Matrix Sentence Test for the assessment of speech intelligibility in noise.* Paper presented at the PROCEEDINGS OF FORUM ACUSTICUM.

Puria, S., Fay, R. R., & Popper, A. N. (2013). *The Middle Ear : Science, Otosurgery, and Technology* Retrieved from http://canterbury.eblib.com.au/patron/FullRecord.aspx?p=1205335

Scarinci, N., Worrall, L., & Hickson, L. (2008). The effect of hearing impairment in older people on the spouse. *International Journal of audiology, 47*(3), 141-151.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *Journal of the Acoustical Society of America, 118*(5), 2775-2778.

Spencer, G. A. (2011). *Effects of speaker age on speech understanding and listening effort in older adults: a thesis submitted in partial fulfilment of the requirements for the degree of Master of Audiology in the University of Canterbury.* (Dissertation/Thesis), http://hdl.handle.net/10092/6343. Retrieved from http://hdl.handle.net/10092/6343

Starr, A., Picton, T. W., Sininger, Y., Hood, L. J., & Berlin, C. I. (1996). Auditory neuropathy. *Brain : a journal of neurology, 119 (Pt 3)*(3), 741-753. doi: 10.1093/brain/119.3.741

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215. doi: http://dx.doi.org/10.1121/1.1907309

Trounson, R. H. (2012). *Development of the UC Auditoryvisual Matrix Sentence Test*. MAud Thesis. Department of Communication Disorders. The University of Canterbury.

Twisk, J., & Rijmen, F. (2009). Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. *Journal of clinical epidemiology, 62*(9), 953-958.

Tye-Murray, N. (2014). *Foundations of aural rehabilitation: Children, adults, and their family members* (4th ed.): Cengage learning.

Wagener, K., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology, 44*(3), 144-156. doi: 10.1080/14992020500057517

Wagener, K., Josvassen, J. L., & Ardenkjær, R. (2003). Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido. *International journal of audiology, 42*(1), 10-17.

Walsh, T. E. (1953). Speech Audiometry. *The Journal of Laryngology and Otology, 67*(3), 119-127. doi: 10.1017/S0022215100048416

Wilson, R. H., McArdle, R. A., & Smith, S. L. (2007). An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners With Normal Hearing and Listeners With Hearing Loss. *Journal of Speech, Language, and Hearing Research, 50*(4), 844-856.

Wong, L. L., & Soli, S. D. (2005). Development of the Cantonese hearing in noise test (CHINT). *Ear and hearing, 26*(3), 276-289.

Zeng, F.-G., & Liu, S. (2006). Speech Perception in Individuals With Auditory Neuropathy. *Journal of Speech, Language, and Hearing Research, 49*(2), 367-380. doi: 10.1044/1092-4388(2006/029)

Zokoll, M. A., Hochmuth, S., Warzybok, A., Wagener, K. C., Buschermöhle, M., & Kollmeier, B. (2013). Speech-in-Noise Tests for Multilingual Hearing Screening and Diagnostics. *American journal of audiology, 22*(1), 175-178.

Appendix A

- Human Ethics Committee approval letter
- Support letter from Maori Research Advisory Group (MRAG)
- Participant information sheet
- Participant consent form

HUMAN ETHICS COMMITTEE

Secretary, Lynda Griffioen
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2014/49

27 June 2014

Amber McClelland
Department of Communication Disorders
UNIVERSITY OF CANTERBURY

Dear Amber

The Human Ethics Committee advises that your research proposal "Naturalisation and normalisation of the UC auditory-visual matrix sentence test" has been considered and approved.

Please note that this approval is subject to the incorporation of the amendments you have provided in your email of 24 June 2014.

Best wishes for your project.

Yours sincerely

Lindsey MacDonald
Chair
University of Canterbury Human Ethics Committee

Māori Research Advisory Group

Tel: +64 3 364 3050 Fax: + 64 364 2950
Email: john.pirker@canterbury.ac.nz

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Monday 16 July 2014

Tēnā koe Amber,

Re: Naturalness and Normalisation of the UC Auditory-Visual Matrix Sentence Test.

I write on behalf of the Maori Research Advisory Group (MRAG). Thank you for your Maori consultation form that you submitted as part of your MAud. The Maori Research Advisory Group (MRAG) is happy to support your research. It would be appreciated if a summary of your findings could be presented to the MRAG upon its completion. We wish you all the best with your research and please feel free to contact me if you have any further questions.

Nāku noa (Yours sincerely)

John Pirker
School of Biological Sciences/College of Science
Research Consultant-Maori (Acting)
University of Canterbury
Private Bag 4800
Christchurch 8020
New Zealand

http://www.biol.canterbury.ac.nz/
Ph: +64 +3 364 3050
Fax: +64 3 364 2590
Skype: john.pirker

Information Sheet

Full Project Title: Naturalness and Normalisation of the UC Auditory-Visual Matrix Sentence Test.

Principal Researcher: Amber McClelland, MAud student (2nd year)
Department of Communication Disorders

Research Supervisor: Associate Professor Greg O'Beirne
Department of Communication Disorders

Associate Supervisor: Dr. Donal Sinex, Senior Lecturer
Department of Communication Disorders

This study is part of a project to produce an auditory-visual speech test in NZ English to supplement the information gathered from other tests typically used in audiology. The study contains three parts. Part one of this project aims to assess how natural the sentences used in this test are. Parts two and three will assess the difficulty of the sentences used in this test.

The test will take place at the University of Canterbury (either in the Audiology clinics of the Department of Communication Disorders, or the Audiology laboratory in Rutherford 801).

To be eligible to participate, you must:
-be 18 years of age or older
-be a native NZ English speaker
-have normal hearing
-have no chronic dexterity issues

A quick hearing check will be undertaken first to determine whether you are able to participate. You will be asked for a history of your ear health and hearing, which ethnic group you belong to, and your ears will be examined. You will then have a hearing check (if you have not provided an audiologist-completed audiogram dated within six months), and I will inform you of the results. If you would like me to, I can write a letter summarising the results if

you would like to follow up on this with your GP or an audiologist. In the event of an unexpected diagnosis of a hearing loss, a full audiological assessment will be offered at the University of Canterbury Speech and Hearing Clinic free of charge. If you choose to follow up with your GP, this will be at your own expense. If a conductive hearing loss were to be identified during the hearing check, you will receive a $10 fuel voucher for your time.

In part one you will watch video of short sentences being read in quiet. In some of the clips, there are noticeable edits that will cause the image to “judder”, while others will appear smooth with no noticeable edits. At the end of each sentence you are to select how much judder you perceived that sentence to have, from "no noticeable judder" to "highly noticeable judder", on the sliding scale provided. Part one should take no more than 1 hour, and will be completed after the hearing check.

Parts two and three will be completed in two independent sessions. In both of these parts, you will hear short sentences being read in noise. The words will change in loudness and may at times be difficult for you to hear. After each sentence has been read, you are to choose the sentence you heard by selecting the words on a screen. Part two should take no longer than 2 hours. Part three will have fewer sentences to appraise than part two and should take no longer than 1 hour.

This study is being carried out as part of a Masters of Audiology. The information I obtain from you will be used in further development of this test so that it may be used as a diagnostic tool.

I am happy to answer any queries you may have. My phone and email details are provided in case you have any questions at a later date. In recognition of the time and effort involved on your behalf, you will receive an honorarium of $30, as well as a free hearing check.

I have provided a consent form for you to sign prior to participating in this study.

Signing this indicates your understanding that the data collected in this study will not be anonymous, but it will be confidential, and only viewed by people directly involved in this study (those listed at the top of the first page). Participation is voluntary and you have the right to withdraw at any stage

without penalty. If you withdraw, I will remove all of the information relating to you.

The project has been reviewed and approved by the University of Canterbury Human Ethics Committee.

For your own reference, please take this form away with you.

With thanks,

Amber McClelland
2nd year MAud Student
Department of Communication Disorders
University of Canterbury
Email: amber.mcclelland@pg.canterbury.ac.nz
Phone: 021 0677 364

Greg O'Beirne, PhD
Primary research supervisor & Associate Professor in Audiology
Department of Communication Disorders
University of Canterbury
Private Bag 4800, Christchurch 8140, New Zealand
Email: gregory.obeirne@canterbury.ac.nz
Phone: +64 3 364 2987 ext. 7085

Donal Sinex, PhD
Secondary research supervisor & Senior Lecturer in Audiology
Department of Communication Disorders
University of Canterbury
Private Bag 4800, Christchurch 8140, New Zealand
Email: donal.sinex@canterbury.ac.nz
Phone: +64 3 364 2987 ext. 7851

Alternatively, if you have any complaints, please contact the Chair of the University of Canterbury Human ethics committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz), phone: +64 3 364 2987.

Consent Form for Persons Participating in Research Studies

Full Project Title: ***Naturalness and Normalisation of the UC Auditory-Visual Matrix Sentence Test.***

I have read and understand the Information Sheet.

I, _____ agree to participate in this project according to the conditions in the Information Sheet. I will be given a copy of Information Sheet and Consent Form to keep.

The researcher has agreed not to reveal the participant's identity and personal details if information about this project is published or presented in any public form.

I agree that research data gathered in this study may be published and used in future studies. I provide consent for this publication and the re-use of the data with the understanding that my name or other identifying information will not be used.

I understand that participation is voluntary and I may withdraw at any time without penalty. Withdrawal of participation will also include the withdrawal of any information I have provided should this remain practically achievable.

I understand that all data collected for the study will be kept in locked and secure facilities and/or in password protected electronic form and will be destroyed after five years.

I understand the risks associated with taking part and how they will be managed.

I understand that I can contact the researcher or supervisor for further information. If I have any complaints, I can contact the Chair of the University of Canterbury Human Ethics Committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz)

I would like to receive a report on the findings of the study at the conclusion of the study (please tick one):

Yes □ **No** □

If yes, please provide a contact email and/or postal address below:

……

By signing below, I agree to participate in this research project.

Signature **Date**

…… **……**

***Note: All parties signing the Consent Form must date their own signature*. Please return the consent form to the researcher before you actively participate in this research.**

Appendix B

- Test lists for constant noise
- Test lists for babble noise

Constant noise sentence lists

List 1					
	David	kept	those	cheap	shirts
	Thomas	sees	twelve	old	ships
	Amy	wins	some	red	bikes
	Sophie	likes	twelve	big	hats
	David	got	three	large	ships
	Rachel	wants	nine	good	shirts
	Peter	has	six	red	coats
	Oscar	sold	some	old	toys
	Hannah	has	three	new	mugs
	Hannah	sold	four	good	books
	Amy	sees	nine	small	books
	William	bought	ten	dark	spoons
	Kathy	gives	six	dark	shoes
	Peter	likes	eight	green	toys
	Oscar	got	ten	new	hats
	Sophie	bought	four	small	bikes
	Rachel	wins	those	large	spoons
	William	wants	two	green	coats
	Thomas	kept	eight	cheap	shoes
	Kathy	gives	two	big	mugs

List 2					
	Peter	kept	eight	dark	coats
	William	likes	three	red	shoes
	Amy	gives	two	cheap	shirts
	Kathy	sees	eight	small	toys
	Oscar	wants	six	green	books
	Sophie	wins	four	new	spoons
	David	has	nine	big	ships
	David	bought	ten	new	hats
	Peter	wins	some	green	toys
	Hannah	got	twelve	large	mugs
	Sophie	sold	nine	dark	coats
	Kathy	likes	two	good	ships
	Oscar	gives	four	red	shirts
	Rachel	sees	ten	good	mugs
	Thomas	wants	three	cheap	bikes
	Amy	sold	twelve	old	shoes
	Rachel	kept	some	large	bikes
	William	got	those	old	spoons
	Thomas	has	six	big	hats
	Hannah	bought	those	small	books

List 3					
	Oscar	got	nine	new	shoes
	Kathy	has	ten	red	ships
	Thomas	gives	four	new	bikes
	Amy	got	three	cheap	ships
	Peter	bought	nine	large	toys
	William	sees	ten	big	shoes
	Kathy	sold	eight	large	shirts
	David	sold	those	good	mugs
	Hannah	bought	two	dark	hats
	Oscar	wins	two	red	toys
	Rachel	likes	twelve	dark	coats
	Peter	has	some	small	books
	Rachel	wants	six	green	spoons
	Hannah	wants	six	big	shirts
	David	wins	four	small	bikes
	Sophie	gives	twelve	old	hats
	William	kept	those	old	books
	Amy	sees	eight	cheap	mugs
	Sophie	kept	three	green	spoons
	Thomas	likes	some	good	coats

List 4					
	Oscar	wants	nine	green	ships
	Peter	has	three	red	mugs
	Sophie	got	nine	old	toys
	Sophie	likes	six	cheap	books
	David	wins	twelve	red	books
	David	bought	two	small	coats
	Oscar	kept	four	large	coats
	William	gives	eight	new	shirts
	Hannah	kept	some	dark	hats
	Rachel	likes	two	good	hats
	Kathy	has	four	big	mugs
	Rachel	wants	those	green	spoons
	Hannah	got	eight	big	spoons
	Peter	sees	those	new	shoes
	Thomas	bought	six	good	bikes
	William	wins	ten	old	shoes
	Amy	sees	three	dark	ships
	Kathy	sold	ten	cheap	toys
	Thomas	sold	twelve	large	bikes
	Amy	gives	some	small	shirts

List 5

Hannah	wants	four	large	coats
William	wants	four	big	ships
William	has	twelve	green	ships
Sophie	likes	those	old	books
David	sees	two	large	spoons
Rachel	gives	ten	small	toys
Oscar	wins	eight	dark	shoes
Oscar	bought	some	good	spoons
Thomas	sold	three	green	shirts
Rachel	gives	twelve	red	shirts
Kathy	sees	some	cheap	mugs
Peter	sold	those	big	bikes
Amy	wins	nine	red	mugs
Thomas	likes	nine	cheap	bikes
Peter	kept	eight	good	shoes
Sophie	got	six	small	toys
David	got	three	new	coats
Amy	bought	two	dark	hats
Hannah	kept	six	new	hats
Kathy	has	ten	old	books

List 6

Peter	kept	two	cheap	shirts
Sophie	wants	twelve	good	hats
Thomas	sold	those	green	spoons
Peter	has	some	large	shoes
Hannah	bought	nine	green	mugs
Rachel	sold	six	small	mugs
Thomas	has	two	red	coats
David	got	nine	old	toys
Sophie	gives	three	dark	bikes
Rachel	wins	eight	cheap	ships
Amy	kept	six	small	coats
William	wants	four	new	shoes
Kathy	sees	three	new	books
William	likes	four	big	ships
Hannah	bought	ten	dark	shirts
Oscar	sees	twelve	red	books
Amy	gives	eight	big	hats
Oscar	got	those	good	toys
David	wins	ten	old	spoons
Kathy	likes	some	large	bikes

List 7

Amy	sees	ten	red	spoons
Kathy	has	some	new	mugs
William	bought	ten	large	ships
Rachel	gives	four	small	coats
Sophie	wins	six	green	ships
Kathy	has	nine	big	coats
Hannah	kept	those	new	hats
Oscar	wins	two	big	shoes
Rachel	wants	three	small	shirts
Oscar	got	those	dark	spoons
David	likes	six	large	mugs
Peter	bought	some	red	toys
David	got	twelve	good	shoes
Peter	sold	two	old	books
Thomas	sees	eight	green	toys
Thomas	kept	eight	good	bikes
Hannah	wants	three	cheap	bikes
William	likes	four	old	shirts
Sophie	gives	nine	dark	hats
Amy	sold	twelve	cheap	books

List 8

William	has	eight	red	bikes
Amy	bought	four	old	spoons
Peter	gives	nine	cheap	mugs
Amy	bought	six	good	hats
Thomas	sold	some	new	mugs
Rachel	wants	ten	red	ships
Sophie	sees	those	green	shoes
Oscar	wants	three	dark	coats
Rachel	sold	twelve	green	books
Thomas	kept	two	dark	hats
William	sees	eight	old	shirts
Sophie	wins	four	large	spoons
Kathy	has	nine	big	books
Hannah	got	six	good	toys
Hannah	wins	some	large	bikes
Oscar	likes	ten	cheap	coats
David	gives	those	new	toys
David	likes	three	small	shirts
Kathy	got	twelve	big	ships
Peter	kept	two	small	shoes

List 9

William	has	eight	big	shoes
Amy	bought	four	cheap	spoons
Peter	gives	nine	dark	hats
Amy	bought	six	good	coats
Thomas	sold	some	green	bikes
Rachel	wants	ten	large	shirts
Sophie	sees	those	new	hats
Oscar	wants	three	old	toys
Rachel	sold	twelve	red	ships
Thomas	kept	two	small	toys
William	sees	eight	big	ships
Sophie	wins	four	cheap	coats
Kathy	has	nine	dark	books
Hannah	got	six	good	bikes
Hannah	wins	some	green	shoes
Oscar	likes	ten	large	mugs
David	gives	those	new	mugs
David	likes	three	old	shirts
Kathy	got	twelve	red	spoons
Peter	kept	two	small	books

List 10

Amy	gives	eight	dark	coats
Oscar	sees	four	dark	hats
Hannah	has	nine	big	mugs
David	sold	six	good	ships
Hannah	likes	some	small	shirts
Kathy	sees	ten	big	shoes
Sophie	kept	those	green	spoons
Oscar	wins	three	new	toys
Peter	kept	twelve	large	bikes
Rachel	sold	two	cheap	books
Sophie	bought	eight	large	coats
Rachel	wants	four	good	hats
Thomas	has	nine	old	mugs
David	got	six	cheap	ships
Amy	likes	some	small	shirts
Peter	got	ten	old	shoes
William	wins	those	red	spoons
William	bought	three	red	toys
Kathy	wants	twelve	green	bikes
Thomas	gives	two	new	books

List 11

Sophie	bought	ten	big	coats
Hannah	gives	four	cheap	hats
Oscar	got	eight	dark	mugs
Amy	has	four	good	ships
Thomas	kept	six	green	shirts
Oscar	likes	nine	large	shoes
David	sees	those	new	spoons
Rachel	sold	three	old	toys
Kathy	wants	twelve	red	bikes
William	wins	some	small	books
William	bought	eight	big	coats
Thomas	gives	those	cheap	hats
Hannah	got	three	dark	mugs
Amy	has	six	good	ships
Kathy	kept	twelve	green	shirts
Peter	likes	ten	large	shoes
Peter	sees	two	new	spoons
Sophie	sold	nine	old	toys
David	wants	two	red	bikes
Rachel	wins	some	small	books

List 12

Kathy	wins	eight	green	coats
Sophie	got	four	old	hats
Oscar	gives	nine	large	mugs
Peter	sold	six	cheap	ships
Thomas	bought	some	new	shirts
Rachel	sold	ten	good	shoes
David	sees	those	green	spoons
David	sees	three	cheap	toys
Thomas	has	twelve	big	bikes
William	wants	two	small	books
Amy	wants	eight	dark	coats
Hannah	bought	four	large	hats
Peter	kept	nine	dark	mugs
Sophie	likes	six	small	ships
Rachel	gives	some	big	shirts
Kathy	likes	ten	new	shoes
Oscar	got	those	red	spoons
Amy	wins	three	good	toys
William	kept	twelve	old	bikes
Hannah	has	two	red	books

List 13

Rachel	gives	eight	cheap	hats
Peter	sold	two	small	ships
Kathy	gives	six	new	toys
Sophie	wins	those	big	coats
Amy	wants	ten	dark	shirts
David	has	some	large	spoons
Oscar	likes	nine	old	toys
Thomas	has	two	dark	shirts
William	bought	three	green	shoes
Kathy	sees	those	green	shoes
William	sold	twelve	small	books
Hannah	got	four	red	bikes
Peter	bought	eight	old	coats
Oscar	kept	six	good	mugs
Thomas	wins	ten	good	hats
Rachel	sees	twelve	cheap	mugs
Sophie	kept	some	large	books
Amy	got	three	red	ships
Hannah	likes	four	new	spoons
David	wants	nine	big	bikes

List 14

William	bought	three	good	shoes
Kathy	gives	some	big	books
Rachel	wins	those	dark	shoes
Sophie	sees	some	red	books
Hannah	has	those	old	hats
David	wants	twelve	big	ships
David	got	twelve	old	coats
Peter	got	ten	green	spoons
Rachel	sees	nine	new	bikes
Oscar	wins	four	small	spoons
William	wants	eight	good	coats
Peter	sold	two	cheap	hats
Amy	sold	six	dark	toys
Thomas	kept	two	large	shirts
Kathy	likes	nine	cheap	bikes
Oscar	bought	four	small	mugs
Amy	kept	six	large	toys
Hannah	gives	three	new	mugs
Thomas	has	eight	green	shirts
Sophie	likes	ten	red	ships

List 15

David	wins	some	big	spoons
Oscar	sees	six	good	toys
Peter	wins	nine	large	shirts
Hannah	has	ten	good	spoons
Amy	bought	ten	large	coats
Kathy	likes	four	new	hats
Rachel	wants	two	cheap	toys
David	sold	those	small	ships
Peter	kept	nine	cheap	books
Sophie	got	six	big	books
William	kept	three	green	bikes
Oscar	gives	four	old	hats
Amy	wants	eight	red	ships
Rachel	sees	twelve	dark	shoes
William	likes	three	dark	mugs
Kathy	sold	two	red	mugs
Hannah	gives	some	old	bikes
Sophie	got	eight	new	shoes
Thomas	bought	twelve	small	shirts
Thomas	has	those	green	coats

List 16

David	wins	some	old	spoons
Oscar	sees	six	green	ships
Peter	wins	nine	new	hats
Hannah	has	ten	big	toys
Amy	bought	ten	large	shirts
Kathy	likes	four	dark	bikes
Rachel	wants	two	good	books
David	sold	those	dark	shoes
Peter	kept	nine	green	spoons
Sophie	got	six	red	shirts
William	kept	three	old	bikes
Oscar	gives	four	small	books
Amy	wants	eight	red	mugs
Rachel	sees	twelve	cheap	coats
William	likes	three	small	toys
Kathy	sold	two	large	hats
Hannah	gives	some	big	shoes
Sophie	got	eight	cheap	mugs
Thomas	bought	twelve	new	coats
Thomas	has	those	good	ships

List 17

Sophie	kept	two	old	toys
William	wins	three	green	ships
Peter	sold	nine	dark	books
Hannah	likes	four	red	books
Amy	bought	three	red	spoons
Peter	kept	those	large	coats
Oscar	got	ten	small	hats
William	likes	six	old	shoes
Thomas	bought	eight	new	shirts
Oscar	sold	those	big	bikes
Kathy	sees	some	new	bikes
David	got	ten	good	shoes
Amy	has	twelve	cheap	mugs
Rachel	wins	some	large	toys
David	gives	nine	small	ships
Thomas	wants	four	good	coats
Rachel	sees	six	big	spoons
Hannah	gives	two	cheap	shirts
Kathy	has	eight	dark	mugs
Sophie	wants	twelve	green	hats

List 18

Oscar	bought	those	big	ships
William	got	some	cheap	bikes
Peter	sold	four	dark	shoes
Sophie	kept	some	good	hats
Kathy	wants	ten	green	spoons
Hannah	kept	six	large	books
Hannah	likes	eight	new	toys
Amy	wins	two	old	coats
Kathy	got	two	red	books
Rachel	likes	those	small	shirts
David	gives	six	big	toys
Thomas	has	nine	cheap	mugs
Oscar	gives	twelve	dark	spoons
Thomas	wants	ten	good	hats
Peter	sees	three	green	coats
Rachel	wins	nine	large	mugs
William	sold	three	new	ships
Amy	has	twelve	old	shirts
Sophie	bought	four	red	shoes
David	sees	eight	small	bikes

List 19

Amy	bought	eight	cheap	shoes
David	gives	four	new	books
Hannah	got	nine	old	bikes
Oscar	kept	six	small	mugs
Kathy	has	some	red	hats
Peter	likes	ten	old	bikes
Rachel	sees	those	small	coats
Sophie	sold	three	large	shirts
Thomas	wants	twelve	green	toys
William	wins	two	dark	spoons
Amy	bought	eight	cheap	ships
David	gives	four	green	ships
Hannah	got	nine	dark	mugs
William	has	six	red	toys
Kathy	kept	some	large	shoes
Peter	sees	ten	good	spoons
Rachel	likes	those	good	books
Sophie	sold	three	big	hats
Thomas	wants	twelve	new	coats
Oscar	wins	two	big	shirts

List 20

Hannah	kept	three	red	hats
David	wins	eight	good	hats
Kathy	bought	nine	good	shoes
Hannah	got	some	cheap	bikes
Thomas	has	those	old	toys
Kathy	sees	eight	dark	toys
Peter	kept	nine	small	shirts
David	wins	those	dark	coats
Rachel	sees	ten	large	books
Peter	gives	ten	big	spoons
Thomas	got	two	new	mugs
Oscar	gives	six	new	shoes
Sophie	wants	twelve	old	bikes
Sophie	sold	two	big	shirts
Amy	bought	six	green	spoons
Rachel	sold	four	red	ships
William	likes	twelve	small	mugs
William	wants	three	cheap	ships
Oscar	likes	some	green	coats
Amy	has	four	large	books

List 21					
	Thomas	bought	eight	big	mugs
	Peter	gives	four	cheap	ships
	William	got	nine	dark	shirts
	David	has	six	good	shoes
	Amy	kept	some	green	spoons
	Hannah	likes	ten	large	toys
	Peter	sees	those	new	bikes
	Sophie	sold	three	old	books
	Sophie	wants	twelve	red	coats
	Kathy	wins	two	small	hats
	Kathy	bought	eight	big	mugs
	Oscar	gives	four	cheap	ships
	David	got	nine	dark	shirts
	Hannah	has	six	good	shoes
	William	kept	some	green	spoons
	Thomas	likes	ten	large	toys
	Oscar	sees	those	new	bikes
	Rachel	sold	three	old	books
	Rachel	wants	twelve	red	coats
	Amy	wins	two	small	hats

List 22					
	Hannah	wins	those	dark	shirts
	Thomas	has	two	small	bikes
	Sophie	likes	nine	cheap	shoes
	Oscar	sees	six	big	toys
	David	likes	three	red	books
	Thomas	sold	three	green	shoes
	Peter	wins	four	new	mugs
	Rachel	kept	some	red	toys
	Hannah	bought	twelve	new	ships
	William	gives	nine	old	spoons
	Amy	sees	those	large	hats
	David	bought	two	large	hats
	Rachel	wants	ten	old	books
	William	has	eight	green	spoons
	Peter	gives	eight	small	coats
	Kathy	wants	six	big	bikes
	Oscar	got	four	cheap	coats
	Amy	kept	twelve	good	mugs
	Kathy	got	ten	dark	shirts
	Sophie	sold	some	good	ships

List 23					
	William	wants	six	big	shoes
	Peter	wants	three	big	bikes
	Oscar	kept	nine	cheap	spoons
	David	gives	three	small	coats
	Amy	got	two	good	hats
	Thomas	sold	nine	dark	bikes
	Amy	kept	four	old	spoons
	Thomas	sees	some	green	shirts
	Kathy	has	twelve	old	toys
	Rachel	gives	ten	small	mugs
	Sophie	bought	two	dark	books
	Rachel	sold	six	good	coats
	William	got	some	new	hats
	Peter	has	those	large	ships
	Kathy	wins	those	red	books
	Hannah	likes	four	large	shoes
	Sophie	sees	eight	cheap	toys
	Oscar	wins	ten	new	ships
	Hannah	bought	eight	red	shirts
	David	likes	twelve	green	mugs

List 24					
	Peter	likes	twelve	small	shoes
	David	has	four	small	mugs
	Rachel	wins	twelve	old	shirts
	Hannah	has	some	large	hats
	Oscar	got	two	good	toys
	Sophie	sees	six	new	books
	Sophie	wants	nine	dark	books
	Thomas	bought	four	green	ships
	Kathy	got	some	red	bikes
	Peter	sold	nine	cheap	hats
	William	bought	six	cheap	ships
	Hannah	kept	ten	good	mugs
	Amy	gives	three	big	spoons
	Rachel	wants	two	red	toys
	Oscar	wins	eight	large	shoes
	David	kept	those	old	spoons
	William	sold	three	dark	bikes
	Amy	gives	eight	green	coats
	Kathy	likes	those	big	shirts
	Thomas	sees	ten	new	coats

List 25

Rachel	kept	ten	dark	hats
Oscar	sees	some	good	spoons
Hannah	wants	two	new	toys
Amy	has	ten	small	ships
Amy	bought	two	new	books
Sophie	got	those	dark	mugs
William	bought	three	red	coats
Kathy	got	eight	small	coats
Sophie	gives	six	red	shirts
William	sees	eight	old	mugs
Peter	has	twelve	large	ships
Rachel	wants	nine	good	spoons
Hannah	kept	six	big	toys
Kathy	wins	some	old	shirts
Peter	likes	twelve	green	hats
Oscar	sold	nine	green	bikes
Thomas	likes	three	big	books
David	gives	four	cheap	shoes
David	wins	those	large	bikes
Thomas	sold	four	cheap	shoes

List 26

Kathy	kept	ten	new	ships
Hannah	bought	six	cheap	coats
Rachel	gives	four	green	shirts
Amy	has	ten	red	coats
Sophie	sees	those	red	bikes
David	sees	three	new	spoons
Thomas	got	three	big	ships
Amy	sold	two	old	shoes
Sophie	gives	nine	good	shirts
William	kept	some	small	shoes
David	has	six	big	spoons
William	wins	twelve	green	hats
Oscar	bought	some	dark	books
Hannah	wants	eight	good	books
Peter	wants	those	large	bikes
Rachel	likes	two	small	mugs
Oscar	wins	nine	large	hats
Thomas	got	eight	old	toys
Kathy	likes	four	cheap	toys
Peter	sold	twelve	dark	mugs

List 27

William	likes	four	small	books
Kathy	sees	those	green	coats
Thomas	sees	nine	good	ships
Sophie	got	twelve	red	shoes
Hannah	kept	eight	dark	spoons
Sophie	wins	ten	dark	bikes
Rachel	gives	two	big	books
Amy	got	three	large	bikes
David	has	some	cheap	toys
Rachel	sold	three	new	toys
William	wants	nine	big	shirts
Amy	has	twelve	red	hats
David	bought	those	cheap	mugs
Oscar	wins	four	old	hats
Hannah	likes	some	old	coats
Oscar	wants	six	new	shoes
Kathy	sold	ten	good	spoons
Peter	kept	two	small	shirts
Peter	bought	eight	large	mugs
Thomas	gives	six	green	ships

List 28

Amy	sees	those	big	hats
Thomas	wants	eight	old	shirts
Hannah	bought	three	red	hats
Sophie	sees	eight	dark	toys
Hannah	has	four	green	bikes
William	wins	some	small	shirts
Peter	wants	two	dark	shoes
Amy	gives	three	big	books
Sophie	wins	ten	old	spoons
Peter	kept	six	cheap	spoons
David	likes	four	large	coats
Oscar	got	six	green	bikes
Thomas	has	ten	new	toys
Rachel	kept	some	large	ships
Rachel	sold	twelve	good	shoes
William	got	those	new	mugs
Oscar	gives	two	good	coats
Kathy	sold	nine	cheap	books
Kathy	likes	twelve	small	mugs
David	bought	nine	red	ships

List 29					
	Sophie	bought	two	dark	ships
	Rachel	kept	nine	good	bikes
	David	bought	eight	small	mugs
	Rachel	gives	twelve	big	shirts
	Oscar	got	three	new	books
	William	has	ten	red	shirts
	William	wants	some	cheap	bikes
	Thomas	likes	those	green	mugs
	Peter	sold	four	new	spoons
	Oscar	sees	three	large	hats
	Amy	has	some	good	toys
	Amy	gives	two	green	shoes
	Sophie	wants	those	dark	coats
	David	kept	ten	big	books
	Kathy	sees	eight	old	shoes
	Kathy	wins	six	old	ships
	Hannah	likes	four	small	hats
	Peter	sold	six	large	spoons
	Hannah	wins	twelve	red	toys
	Thomas	got	nine	cheap	coats

List 30					
	Rachel	wins	four	green	spoons
	Kathy	got	those	large	shirts
	Sophie	kept	ten	cheap	coats
	Peter	kept	three	good	ships
	Amy	bought	six	large	hats
	Sophie	likes	two	good	spoons
	Oscar	likes	six	red	shoes
	Kathy	wants	ten	old	ships
	Peter	sees	nine	big	toys
	Hannah	gives	those	red	toys
	Thomas	wants	three	small	bikes
	Rachel	gives	eight	green	hats
	Thomas	got	four	dark	shoes
	Hannah	sees	twelve	small	mugs
	Oscar	wins	some	old	bikes
	Amy	sold	two	new	shirts
	David	bought	nine	cheap	books
	William	has	eight	big	mugs
	David	has	some	dark	books
	William	sold	twelve	new	coats

Babble noise sentence lists

List 1					
	Hannah	sold	twelve	dark	bikes
	Amy	likes	two	good	ships
	Hannah	sees	eight	green	bikes
	Sophie	sees	those	cheap	hats
	Kathy	has	those	large	ships
	Thomas	got	six	red	books
	Peter	wants	nine	new	coats
	William	bought	ten	big	toys
	William	likes	two	cheap	mugs
	Rachel	bought	some	large	books
	Sophie	sold	ten	big	books
	Peter	got	some	small	spoons
	Kathy	has	those	good	shoes
	Oscar	gives	three	new	toys
	Oscar	kept	six	red	hats
	Amy	kept	twelve	old	bikes
	Thomas	gives	four	small	spoons
	David	sold	three	green	coats
	David	gives	eight	old	shoes
	Rachel	wants	four	dark	mugs

List 2					
	Kathy	bought	eight	dark	bikes
	Oscar	sold	three	red	coats
	Amy	got	two	cheap	shoes
	William	sold	eight	small	books
	Sophie	likes	six	green	ships
	Rachel	sold	four	new	shoes
	Oscar	likes	nine	big	coats
	Peter	kept	some	new	bikes
	David	got	some	green	coats
	Sophie	sees	twelve	dark	hats
	Thomas	gives	nine	big	ships
	Hannah	sold	two	good	mugs
	David	kept	four	red	toys
	Kathy	has	ten	good	mugs
	Thomas	wants	three	cheap	spoons
	Peter	bought	twelve	old	spoons
	Hannah	sees	ten	large	books
	Rachel	gives	those	old	coats
	Amy	has	six	large	toys
	William	gives	those	small	hats

List 3					
	Rachel	got	some	red	shoes
	William	has	those	large	ships
	Oscar	gives	six	big	bikes
	Peter	got	ten	old	ships
	Oscar	bought	twelve	red	toys
	Sophie	sees	two	cheap	shoes
	Amy	sold	some	small	hats
	Sophie	sold	nine	new	mugs
	William	bought	eight	dark	hats
	Rachel	gives	two	old	toys
	Kathy	has	four	cheap	coats
	David	likes	four	large	books
	Thomas	wants	three	good	spoons
	Hannah	sees	ten	good	mugs
	David	got	those	green	bikes
	Thomas	gives	eight	big	hats
	Amy	kept	nine	new	books
	Kathy	sees	three	small	mugs
	Hannah	kept	twelve	dark	spoons
	Peter	likes	six	green	coats

List 4					
	Rachel	wants	two	old	books
	Thomas	got	eight	big	coats
	Kathy	has	nine	red	hats
	David	likes	four	cheap	books
	William	likes	two	large	spoons
	Oscar	sold	four	good	mugs
	Amy	kept	some	old	spoons
	Rachel	bought	six	new	hats
	Sophie	kept	twelve	big	ships
	Hannah	sees	four	good	bikes
	Kathy	has	three	small	toys
	Peter	wants	ten	dark	ships
	Amy	got	those	small	shoes
	Peter	sees	eight	dark	shoes
	Thomas	sold	nine	new	bikes
	Sophie	likes	some	green	mugs
	Oscar	sees	three	red	coats
	Hannah	sold	ten	large	toys
	William	bought	those	green	ships
	David	gives	six	cheap	shoes

List 5

Hannah	kept	four	large	coats
William	has	eight	dark	spoons
William	likes	twelve	big	bikes
Sophie	bought	those	cheap	shoes
David	gives	two	green	mugs
Rachel	sold	ten	green	books
Oscar	kept	four	new	toys
Oscar	sees	some	large	spoons
Thomas	sees	three	big	coats
Rachel	likes	twelve	small	ships
Kathy	sold	some	old	books
Peter	wants	those	good	bikes
Amy	bought	nine	red	mugs
Thomas	got	nine	red	hats
Peter	gives	eight	old	spoons
Sophie	wants	six	dark	shoes
David	got	three	small	toys
Amy	got	two	cheap	hats
Hannah	sold	six	new	ships
Kathy	has	ten	good	toys

List 6

Peter	sees	four	dark	books
Sophie	likes	some	large	hats
Thomas	wants	twelve	cheap	bikes
Peter	sold	some	new	toys
Hannah	gives	ten	large	shoes
Rachel	likes	three	good	shoes
Thomas	sees	eight	old	mugs
David	gives	six	cheap	hats
Sophie	sold	eight	big	ships
Rachel	gives	those	small	toys
Amy	has	three	old	books
William	got	some	green	spoons
Kathy	has	two	red	coats
William	kept	nine	new	spoons
Hannah	sold	ten	good	books
Oscar	bought	ten	small	bikes
Amy	got	twelve	dark	mugs
Oscar	bought	two	red	ships
David	wants	four	green	coats
Kathy	kept	six	big	bikes

List 7

Rachel	wants	ten	dark	spoons
Kathy	sees	ten	good	mugs
Oscar	kept	four	big	ships
William	has	ten	new	coats
Thomas	sold	six	small	ships
Hannah	sees	nine	red	coats
Peter	likes	those	large	hats
Amy	has	two	green	shoes
Thomas	got	three	cheap	coats
Oscar	gives	those	new	spoons
David	got	six	dark	mugs
Kathy	sold	twelve	old	toys
Sophie	kept	some	small	shoes
Hannah	gives	twelve	old	books
Sophie	bought	eight	good	toys
Peter	wants	two	large	bikes
Rachel	likes	three	cheap	bikes
William	gives	four	green	books
David	sold	nine	big	hats
Amy	sees	twelve	red	books

List 8

Peter	bought	some	red	bikes
Amy	bought	some	old	coats
Thomas	likes	three	cheap	spoons
Thomas	gives	ten	good	hats
David	wants	six	new	shoes
William	wants	twelve	red	toys
Amy	sees	those	green	books
Rachel	wants	ten	dark	hats
Oscar	sold	eight	green	mugs
Oscar	kept	six	dark	coats
Hannah	sees	eight	old	mugs
Hannah	sold	six	large	mugs
Rachel	has	two	big	bikes
Kathy	got	those	good	hats
Peter	has	twelve	large	spoons
William	likes	four	cheap	toys
Sophie	gives	two	new	ships
Sophie	likes	four	small	shoes
Kathy	got	nine	big	books
David	kept	six	small	ships

List 9

Oscar	bought	two	big	shoes
Amy	has	eight	big	spoons
Kathy	gives	ten	red	hats
Sophie	bought	ten	good	coats
David	sold	two	new	bikes
Hannah	sold	six	old	ships
Kathy	sees	eight	large	hats
Rachel	wants	nine	new	toys
Amy	wants	three	cheap	ships
Oscar	kept	some	green	toys
Rachel	sees	four	old	ships
Peter	gives	those	large	coats
William	has	twelve	small	books
Thomas	got	those	green	bikes
William	likes	some	small	shoes
David	likes	three	good	mugs
Hannah	gives	six	dark	mugs
Peter	likes	twelve	dark	shoes
Thomas	got	nine	red	spoons
Sophie	kept	four	cheap	books

List 10

Amy	has	eight	old	coats
Oscar	sold	four	cheap	hats
Hannah	gives	four	big	mugs
David	likes	six	big	ships
Hannah	sees	some	small	spoons
Kathy	kept	nine	red	shoes
Sophie	bought	those	old	spoons
Oscar	sold	twelve	large	toys
Peter	has	three	new	bikes
Rachel	sold	two	good	books
Sophie	likes	eight	big	coats
Rachel	wants	four	green	hats
Thomas	got	three	good	mugs
David	sees	some	red	ships
Amy	bought	six	green	toys
Peter	gives	ten	dark	shoes
William	kept	some	new	spoons
William	sold	three	dark	toys
Kathy	wants	twelve	cheap	books
Thomas	sold	two	large	bikes

List 11

Amy	has	four	cheap	toys
Peter	kept	twelve	small	bikes
Hannah	kept	twelve	big	spoons
David	kept	four	old	shoes
Oscar	gives	ten	good	mugs
Rachel	likes	six	green	books
Amy	got	three	good	spoons
Kathy	wants	those	red	hats
Oscar	sold	those	large	ships
William	got	ten	red	shoes
David	sold	three	dark	ships
Kathy	sees	nine	new	coats
Rachel	sold	those	small	bikes
Sophie	gives	eight	new	mugs
Hannah	sees	some	old	books
Peter	likes	some	large	hats
Thomas	got	two	cheap	bikes
William	bought	eight	new	toys
Thomas	sees	two	big	coats
Sophie	wants	six	green	books

List 12

William	kept	six	green	shoes
David	sold	three	old	books
Rachel	gives	six	large	shoes
Oscar	bought	twelve	cheap	bikes
Thomas	kept	twelve	new	mugs
Sophie	got	two	good	toys
Oscar	sold	six	green	ships
Peter	got	ten	big	toys
Amy	wants	those	cheap	mugs
Rachel	likes	two	small	ships
Thomas	sees	eight	dark	coats
David	likes	four	large	spoons
Amy	has	three	dark	spoons
William	sold	four	small	coats
Hannah	kept	four	big	coats
Kathy	wants	some	new	books
Kathy	sees	some	red	bikes
Peter	bought	ten	good	hats
Hannah	gives	eight	old	hats
Sophie	sold	nine	red	hats

List 13

Amy	has	ten	big	mugs
Hannah	sold	twelve	old	toys
Thomas	gives	those	cheap	coats
Kathy	sold	those	good	coats
Amy	kept	some	large	ships
Peter	got	some	new	books
Sophie	gives	six	small	bikes
David	gives	eight	dark	ships
Hannah	sees	four	new	mugs
Sophie	wants	ten	large	bikes
William	kept	nine	red	mugs
William	wants	six	green	shoes
Oscar	sold	two	red	spoons
Thomas	sees	two	small	ships
David	gives	twelve	dark	books
Oscar	likes	three	cheap	hats
Rachel	bought	nine	big	toys
Peter	likes	four	green	hats
Kathy	got	eight	good	shoes
Rachel	bought	three	old	spoons

List 14

Oscar	bought	eight	good	shoes
William	gives	eight	big	books
Sophie	gives	twelve	dark	shoes
Rachel	sees	ten	red	books
Kathy	has	six	old	hats
Peter	wants	three	big	ships
Peter	got	six	old	coats
Thomas	got	ten	green	spoons
Amy	sees	twelve	new	bikes
Hannah	kept	four	small	spoons
Oscar	likes	eight	good	coats
Hannah	sold	some	large	hats
Rachel	sold	nine	big	toys
Sophie	kept	four	cheap	shoes
David	likes	two	cheap	bikes
Amy	bought	three	small	mugs
Kathy	kept	four	large	toys
David	gives	two	new	mugs
William	has	some	green	spoons
Thomas	likes	those	red	ships

List 15

William	kept	six	cheap	coats
Sophie	sees	some	large	toys
Sophie	gives	nine	big	shoes
William	has	ten	good	toys
Kathy	likes	ten	large	coats
Kathy	bought	four	new	hats
Oscar	sees	two	cheap	spoons
Thomas	sees	those	small	bikes
Amy	has	nine	big	ships
Hannah	sold	six	cheap	toys
Amy	got	three	green	spoons
Oscar	likes	four	old	shoes
Rachel	sold	eight	red	ships
David	got	twelve	dark	bikes
Rachel	wants	three	dark	books
Thomas	gives	two	red	bikes
Peter	kept	some	old	books
Rachel	bought	eight	new	hats
David	bought	twelve	small	mugs
Peter	wants	those	green	mugs

List 16

Sophie	sees	three	red	books
Amy	sees	those	cheap	mugs
Hannah	kept	six	red	toys
Kathy	has	four	good	toys
William	bought	those	old	bikes
Amy	likes	nine	big	mugs
Rachel	wants	six	dark	spoons
William	sold	two	big	hats
Hannah	kept	twelve	small	ships
David	got	three	good	spoons
Sophie	kept	nine	new	shoes
Oscar	gives	twelve	dark	bikes
Kathy	wants	some	green	hats
Thomas	sees	some	small	books
David	likes	two	old	coats
Peter	sold	ten	large	ships
Oscar	gives	eight	new	coats
Rachel	got	some	green	books
Peter	bought	ten	large	coats
Peter	has	four	cheap	shoes

List 17

David	kept	four	old	spoons
Kathy	bought	ten	green	hats
Oscar	sold	two	dark	bikes
Rachel	likes	four	red	mugs
David	bought	eight	red	coats
Hannah	kept	six	large	spoons
Sophie	got	eight	small	books
Sophie	likes	four	old	hats
Amy	bought	some	new	coats
Rachel	sold	twelve	big	mugs
Peter	sees	those	new	toys
Kathy	got	three	good	ships
Oscar	likes	three	cheap	toys
Thomas	sold	twelve	large	books
Hannah	gives	two	small	hats
Thomas	wants	those	good	shoes
Peter	sees	nine	big	ships
William	gives	nine	big	mugs
William	has	those	green	bikes
Amy	wants	some	large	shoes

List 18

William	bought	six	cheap	hats
David	got	nine	big	toys
Kathy	sold	twelve	dark	mugs
Hannah	kept	four	good	ships
Kathy	wants	some	green	shoes
Rachel	kept	twelve	large	books
Thomas	likes	two	new	mugs
Amy	bought	eight	old	bikes
Rachel	got	nine	red	toys
Peter	likes	those	small	coats
Sophie	gives	eight	big	shoes
William	has	three	cheap	coats
Thomas	gives	three	good	shoes
Sophie	wants	six	dark	hats
Amy	sees	two	green	books
Kathy	sees	ten	large	spoons
Oscar	sold	some	new	spoons
Amy	has	ten	old	ships
Peter	bought	those	red	hats
David	sees	four	small	bikes

List 19

Amy	bought	eight	old	ships
David	got	four	cheap	spoons
Hannah	gives	nine	red	toys
Oscar	kept	six	small	coats
Kathy	has	some	large	mugs
Peter	likes	ten	green	coats
Rachel	sees	those	small	ships
Sophie	sold	three	good	shoes
Thomas	wants	twelve	large	toys
William	wants	two	dark	shoes
Amy	bought	eight	good	bikes
David	gives	four	green	mugs
Sophie	got	nine	big	books
William	has	six	old	bikes
Kathy	kept	some	new	books
Peter	sees	ten	dark	hats
Rachel	likes	those	cheap	spoons
Sophie	sold	three	big	spoons
Thomas	wants	twelve	red	toys
Oscar	sold	two	new	hats

List 20

William	sees	ten	new	shoes
Kathy	bought	two	old	bikes
Hannah	sold	four	old	ships
Rachel	got	three	good	books
William	has	two	dark	ships
Amy	sees	nine	red	books
Oscar	kept	twelve	red	spoons
Sophie	sold	twelve	small	coats
Thomas	sees	eight	green	hats
David	gives	those	new	spoons
David	got	eight	green	mugs
Sophie	gives	four	dark	bikes
Rachel	wants	nine	big	hats
Amy	sold	six	large	shoes
David	bought	those	cheap	books
Hannah	sold	six	cheap	coats
Oscar	likes	some	large	bikes
Kathy	wants	ten	good	toys
Peter	likes	three	big	mugs
Peter	has	some	small	toys

List 21

Thomas	sees	those	cheap	hats
Peter	likes	some	green	shoes
William	has	eight	good	spoons
David	kept	some	small	shoes
Amy	gives	two	dark	bikes
Hannah	sold	six	large	coats
Peter	bought	four	old	books
Sophie	got	ten	good	toys
Sophie	wants	twelve	old	coats
Kathy	sees	twelve	small	mugs
Kathy	bought	two	large	hats
Oscar	kept	nine	new	spoons
David	wants	ten	green	coats
Hannah	sees	three	big	ships
William	bought	those	new	books
Thomas	likes	three	big	mugs
Kathy	got	six	cheap	bikes
Rachel	has	nine	red	toys
Rachel	sold	four	dark	hats
Amy	gives	eight	red	ships

List 22

Sophie	likes	three	dark	mugs
Peter	has	some	small	bikes
Sophie	likes	two	cheap	shoes
Oscar	sees	ten	big	toys
David	wants	six	red	books
Thomas	sold	those	green	shoes
Amy	bought	two	new	toys
Rachel	kept	some	red	mugs
Thomas	got	twelve	old	ships
David	gives	nine	new	spoons
William	sees	four	large	hats
Hannah	sees	eight	large	hats
Kathy	wants	twelve	old	books
Amy	has	those	green	spoons
William	gives	three	small	coats
Oscar	likes	eight	good	bikes
Peter	got	nine	big	coats
Oscar	kept	six	cheap	mugs
Kathy	got	four	dark	ships
Rachel	sold	ten	good	ships

List 23

Kathy	wants	nine	new	shoes
Rachel	wants	those	cheap	bikes
Oscar	kept	some	green	spoons
Rachel	gives	eight	good	coats
Amy	got	three	new	hats
Hannah	sold	four	large	bikes
David	kept	some	small	spoons
Peter	sees	four	old	coats
Amy	has	ten	red	toys
Sophie	gives	ten	dark	mugs
Sophie	bought	three	dark	books
Amy	sold	some	small	coats
William	got	twelve	new	hats
Peter	has	six	old	ships
Thomas	sees	two	large	books
Rachel	likes	eight	big	shoes
William	sees	twelve	big	toys
David	gives	two	red	ships
Kathy	bought	six	cheap	shoes
Thomas	likes	those	good	mugs

List 24

David	sees	four	good	toys
William	sold	twelve	cheap	coats
Rachel	wants	twelve	small	shoes
William	has	some	large	shoes
Hannah	gives	two	green	books
William	gives	six	cheap	hats
Amy	bought	nine	new	spoons
David	kept	four	old	bikes
Peter	sold	eight	large	mugs
Oscar	bought	nine	new	toys
Thomas	kept	six	old	spoons
Amy	got	ten	small	coats
Sophie	gives	three	red	spoons
Rachel	sold	two	dark	ships
Sophie	kept	some	red	hats
Oscar	sees	those	big	mugs
Sophie	likes	three	green	books
Peter	got	eight	good	ships
Kathy	wants	those	big	bikes
Kathy	likes	ten	dark	toys

List 25					
	Oscar	kept	some	large	toys
	Thomas	sees	ten	good	bikes
	William	wants	two	new	ships
	Kathy	has	ten	new	mugs
	David	bought	two	dark	books
	Peter	got	those	big	books
	Oscar	bought	three	small	books
	Thomas	got	eight	red	hats
	Sophie	gives	six	green	toys
	Peter	sees	eight	cheap	spoons
	William	has	twelve	dark	bikes
	David	wants	six	old	coats
	Hannah	kept	six	large	hats
	Amy	sold	some	green	ships
	Sophie	likes	nine	red	bikes
	Hannah	sold	nine	big	shoes
	Amy	likes	three	small	shoes
	Rachel	gives	four	old	mugs
	Rachel	wants	those	big	spoons
	Kathy	sold	four	cheap	coats

List 26					
	Thomas	wants	six	new	mugs
	Amy	got	twelve	red	toys
	Peter	kept	four	large	coats
	William	has	some	old	bikes
	Amy	has	eight	small	shoes
	Oscar	sold	eight	dark	coats
	Kathy	sees	two	good	shoes
	Sophie	kept	some	new	toys
	Sophie	got	two	green	books
	David	likes	ten	green	spoons
	David	bought	twelve	dark	hats
	Kathy	sees	those	big	hats
	Peter	got	three	cheap	bikes
	Oscar	gives	nine	big	spoons
	Thomas	likes	two	green	ships
	Hannah	sold	ten	large	ships
	Hannah	gives	three	cheap	hats
	Rachel	sees	four	small	books
	William	bought	eight	red	mugs
	Rachel	wants	six	old	coats

List 27					
	William	has	two	small	books
	Hannah	gives	those	green	coats
	Sophie	likes	eight	good	toys
	William	bought	eight	red	toys
	William	has	nine	big	bikes
	Oscar	gives	twelve	dark	hats
	Peter	gives	three	dark	hats
	Amy	kept	some	large	spoons
	David	wants	twelve	cheap	spoons
	Sophie	sees	three	new	shoes
	Peter	sold	four	big	mugs
	Thomas	sees	ten	red	books
	Thomas	kept	some	old	shoes
	Kathy	bought	ten	new	ships
	Amy	sold	those	old	bikes
	Rachel	got	nine	new	mugs
	Rachel	got	two	good	coats
	Oscar	likes	six	small	mugs
	Kathy	got	six	large	ships
	David	wants	four	green	ships

List 28					
	Sophie	sees	three	big	bikes
	Rachel	wants	two	dark	toys
	David	bought	ten	small	shoes
	William	has	two	new	hats
	Hannah	sees	those	cheap	ships
	Peter	bought	three	red	spoons
	William	wants	twelve	old	coats
	Peter	gives	twelve	dark	bikes
	Thomas	likes	eight	old	spoons
	Sophie	kept	nine	red	mugs
	Oscar	likes	ten	large	books
	Thomas	got	those	small	shoes
	Hannah	kept	four	cheap	books
	Kathy	has	four	cheap	spoons
	Amy	sold	six	green	ships
	David	got	eight	good	shoes
	Kathy	gives	six	new	mugs
	Oscar	sold	nine	big	hats
	Amy	likes	some	large	coats
	Rachel	bought	some	green	toys

List 29						List 30					
	Rachel	has	two	large	ships		Hannah	kept	four	large	spoons
	Amy	bought	ten	small	bikes		David	bought	those	green	bikes
	William	bought	eight	dark	mugs		Sophie	wants	ten	big	coats
	David	sold	twelve	cheap	toys		Thomas	sold	three	good	ships
	Thomas	wants	three	small	books		Sophie	sees	two	large	hats
	Hannah	sold	ten	big	toys		Peter	has	six	dark	spoons
	Hannah	kept	some	old	bikes		Thomas	sees	four	old	shoes
	Kathy	likes	those	green	mugs		Amy	kept	six	cheap	ships
	Oscar	gives	four	good	spoons		Oscar	likes	nine	big	toys
	Kathy	sees	nine	red	hats		Oscar	gives	those	new	toys
	Rachel	wants	some	new	toys		Amy	got	three	green	books
	Rachel	wants	two	green	shoes		Rachel	sold	eight	dark	hats
	Sophie	likes	those	new	coats		David	wants	ten	red	shoes
	Sophie	sees	ten	old	books		William	bought	twelve	new	mugs
	William	got	eight	large	shoes		Peter	gives	some	old	bikes
	David	got	four	good	ships		Rachel	likes	two	good	books
	Peter	gives	four	big	hats		Kathy	has	nine	red	bikes
	Peter	kept	six	red	spoons		Hannah	sold	eight	small	mugs
	Amy	sold	three	dark	toys		William	got	some	small	books
	Thomas	kept	twelve	cheap	coats		Kathy	gives	twelve	cheap	coats