

CREATING AND EVALUATING NEW ZEALAND-ACCENTED SYNTHESISED  
VOICES USING MODEL TALKER VOICE BANKING TECHNOLOGY

Michelle Beverley Westley

University of Canterbury

A thesis submitted in partial fulfillment of the requirements for the degree of Master of  
Science in Speech and Language Sciences at the University of Canterbury, 2018

## Acknowledgements

I express great appreciation to my supervisor Dr. Dean Sutherland for his valuable guidance throughout the research. I would like to acknowledge the time and support which Tim Bunnell and the ModelTalker team provided to allow the creation of the synthesised voices. Thank you to Grace Eriksen for the effort put into her honours project and her assistance with our younger voice donors. I also wish to thank various people for their contributions to this project: Dr. Annalise Fletcher for her support with the perceptual experiment methodology and results, Professor Jeanette King for her expertise of the te reo Māori language, the speech-language therapists at the TalkLink Trust, Matthew Edwards and Sidney Wong for proofreading, and the support of my family and friends. I am extremely grateful to the Graduate Women Canterbury branch, whose scholarship enabled me to undertake postgraduate study and present at the American Speech-Language-Hearing Association annual convention in Los Angeles. This project would not be possible without the participants who took part and I am particularly grateful to the voice donors and the parents of our two youngest voice donors for their commitments to the project. Finally, a special acknowledgment goes to a client with whom I had the pleasure of working in my honours year; his humour and outlook on life truly inspired me to continue with research that will benefit people with motor neurone disease.

## Abstract

Communication, in all its modalities, is an important way for individuals to express themselves and connect with others. An individual's own voice portrays many aspects of their personality and identity. Individuals who have conditions which reduce the ability to speak using their natural voice face a lack of personalisation and customisation of the synthetic voices available for speech generating devices. This may lead to a decrease in device acceptance. In New Zealand, there are currently no locally-accented voices for speech generating device users. Voice banking is the process of recording one's voice to create a personalised synthetic voice for use on speech generating device. This study explored the experiences of those who voice bank, and investigated the quality of the resulting voices. Eight healthy adults and two healthy children participated in the ModelTalker voice banking process and completed a questionnaire to gather the voice donors' perceptions of the experience. Fifteen unfamiliar listeners assessed perceptive aspects of the synthetic voices created. The measures used included the Speech Intelligibility Test, intelligibility and naturalness visual analogue scales, and age and gender identification tasks. Personalised synthetic voices were successfully created using the ModelTalker system. The voice donors reported positive experiences and identified multiple strengths and challenges of the ModelTalker voice banking system, which were consistent with previous literature (Creer, Green, & Cunningham, 2009; Hyppa-Martin, Friese, & Barnes, 2017; Jackson et al., 2017). The synthesised voices were found to have intelligibilities similar to those previously reported for synthetic speech, and age and gender estimations followed patterns reported in the literature (Cerrato, Falcone, & Paoloni, 2000; Jreige, Patel, & Bunnell, 2009; Von Berg, Panorska, Uken, & Qeadan, 2009; Waller, Eriksson, & Sorqvist, 2015). Future directions for this area should include perceptions of voice banking experiences for clinical populations such as those with progressive speech loss. Personalisation of the voice banking process for the New Zealand accent should continue, as should the creation of a fully synthetic te reo Māori voice. The voices created by this study are available for New Zealand speech generating device users who want a locally-accented voice on their device. With the availability of these voices, this thesis has addressed the lack of New Zealand-accented synthetic voices available for speech generating device users.

# Table of Contents

<b>Acknowledgements .....</b>	<b>ii</b>
<b>Abstract.....</b>	<b>iii</b>
<b>Table of Contents .....</b>	<b>iv</b>
<b>List of Figures.....</b>	<b>vii</b>
<b>List of Tables .....</b>	<b>viii</b>
<b>Chapter 1. Introduction.....</b>	<b>1</b>
<b>1.1. Populations affected by speech loss and impairment.....</b>	<b>1</b>
1.1.1. Motor Neurone Disease .....	2
1.1.2. Dementia.....	3
1.1.3. Parkinson’s Disease .....	3
1.1.4. Aphasia .....	4
1.1.5. Cerebral Palsy.....	5
1.1.6. Autism Spectrum Disorder .....	6
<b>1.2. Augmentative and alternative communication.....</b>	<b>6</b>
<b>1.3. Speech generating devices .....</b>	<b>7</b>
<b>1.4. Experiences of using speech generating devices .....</b>	<b>8</b>
<b>1.5. Aspects which influence device acceptability.....</b>	<b>9</b>
1.5.1. Intelligibility .....	10
1.5.2. Naturalness .....	12
1.5.3. Representation of the user .....	12
<b>Chapter 2. Personalisation of synthetic voices for New Zealanders.....</b>	<b>15</b>
<b>2.1. New Zealand linguistic culture.....</b>	<b>15</b>
<b>2.2. Current synthesised voice options .....</b>	<b>16</b>
<b>2.3. Voice banking .....</b>	<b>17</b>
2.3.1. CereVoice Me.....	19
2.3.2. My-own-voice .....	19
2.3.3. VocaliD.....	20
2.3.4. ModelTalker .....	20
<b>2.4. Present study.....</b>	<b>23</b>
<b>Chapter 3. Voice banking methodology .....</b>	<b>25</b>
<b>3.1. Study design.....</b>	<b>25</b>
<b>3.2. Ethical considerations .....</b>	<b>25</b>

3.2.1.	Confidentiality of participants .....	25
3.2.2.	Maintenance and intellectual property of the synthesised voices .....	26
3.2.3.	Considerations for child participants .....	26
3.2.4.	Cultural sensitivity .....	27
<b>3.3.</b>	<b>Recruitment .....</b>	<b>27</b>
3.3.1.	Voice donor requirements .....	27
3.3.2.	Screening session .....	28
3.3.3.	Screening process .....	29
3.3.4.	Selection of voice donor participants .....	29
<b>3.4.</b>	<b>ModelTalker voice banking process .....</b>	<b>30</b>
3.4.1.	Materials and set up .....	30
3.4.2.	Initial recording session .....	31
3.4.3.	Recording sessions .....	31
<b>3.5.</b>	<b>Te reo Māori custom inventory .....</b>	<b>32</b>
3.5.1.	Twenty common te reo Māori words .....	32
3.5.2.	Extended te reo Māori custom inventory .....	33
<b>3.6.</b>	<b>Recording the Speech Intelligibility Test sentences .....</b>	<b>33</b>
<b>3.7.</b>	<b>Qualitative methods .....</b>	<b>33</b>
3.7.1.	Participants .....	33
3.7.2.	Materials .....	33
3.7.3.	Procedure .....	33
3.7.4.	Measures .....	34
<b>3.8.</b>	<b>Data analysis .....</b>	<b>36</b>
<b>Chapter 4.</b>	<b>Voice banking results .....</b>	<b>37</b>
<b>4.1.</b>	<b>Voice donor rates of recording .....</b>	<b>37</b>
<b>4.2.</b>	<b>Results from online questionnaire .....</b>	<b>38</b>
4.2.1.	Pre-voice banking knowledge .....	39
4.2.2.	Experiences with voice banking .....	39
4.2.3.	Features of the ModelTalker Voice Recorder .....	41
4.2.4.	New Zealand adaptations .....	48
4.2.5.	Recommendations for other people and other environments .....	51
<b>Chapter 5.</b>	<b>Perceptual experiment methodology .....</b>	<b>57</b>
<b>5.1.</b>	<b>Participants .....</b>	<b>57</b>
<b>5.2.</b>	<b>Materials .....</b>	<b>57</b>
<b>5.3.</b>	<b>Procedure .....</b>	<b>58</b>
<b>5.4.</b>	<b>Measures .....</b>	<b>58</b>

5.4.1.	Speech Intelligibility Test.....	59
5.4.2.	Intelligibility and naturalness visual analogue scales.....	60
5.4.3.	Age and gender identification.....	61
<b>5.5.</b>	<b>Data analysis .....</b>	<b>61</b>
<b>Chapter 6.</b>	<b>Perceptual experiment results .....</b>	<b>63</b>
<b>6.1.</b>	<b>Speech Intelligibility Test .....</b>	<b>63</b>
<b>6.2.</b>	<b>Intelligibility visual analogue scale .....</b>	<b>67</b>
<b>6.3.</b>	<b>Naturalness visual analogue scale .....</b>	<b>69</b>
<b>6.4.</b>	<b>Age estimation .....</b>	<b>71</b>
<b>6.5.</b>	<b>Gender identification .....</b>	<b>73</b>
<b>Chapter 7.</b>	<b>Discussion.....</b>	<b>74</b>
<b>7.1.</b>	<b>Aim One: To explore the effectiveness of the ModelTalker voice banking protocol for New Zealand speakers.....</b>	<b>74</b>
7.1.1.	What is the experience of healthy voice donors during the ModelTalker voice banking process?.....	74
7.1.2.	Are there any alterations that are required to make ModelTalker voice banking suitable for New Zealanders?.....	77
<b>7.2.</b>	<b>Aim Two: To create and evaluate the New Zealand-accented synthetic voices for speech generating devices .....</b>	<b>78</b>
7.2.1.	How do unfamiliar listeners perceive the intelligibility and naturalness of the New Zealand-accented voices created using ModelTalker technology?.....	78
7.2.2.	Can unfamiliar listeners perceive the age and gender of the New Zealand-accented voices created using ModelTalker technology?.....	82
<b>7.3.</b>	<b>Clinical implications.....</b>	<b>83</b>
<b>7.4.</b>	<b>Future directions .....</b>	<b>85</b>
<b>7.5.</b>	<b>Conclusions .....</b>	<b>86</b>
<b>References.....</b>		<b>88</b>
<b>Appendices.....</b>		<b>97</b>

## List of Figures

Figure 1. Set up of the recording session. ....	30
Figure 2. Voice banking participant experiences with voice banking. ....	40
Figure 3. Voice banking participant ratings of preview synthesised voices at the start and end of voice banking.....	45
Figure 4. Voice banking participant ratings of the length of time required to voice bank. ....	46
Figure 5. Voice banking participant recommendations about other people donating their voice vs. voice banking for those who are losing their voice.....	52
Figure 6. Voice banking participant ratings of success if recording at home.....	54
Figure 7. Voice banking participant ratings of success if voice banking independently.....	56
Figure 8. New Zealand-accented voices intelligibility at word level with the Speech Intelligibility Test.....	64
Figure 9. Relationship between length of Speech Intelligibility Test sentence stimulus and percentage words correct. ....	66
Figure 10. Relationship between exposure of synthesised voices and percentage words correct. ....	66
Figure 11. New Zealand synthesised voices average intelligibility visual analogue ratings. ...	67
Figure 12. Correlations between Speech Intelligibility Test percentage of words correct and average intelligibility visual analogue ratings. ....	69
Figure 13: New Zealand synthesised voices average naturalness visual analogue ratings.....	70
Figure 14. Unfamiliar listeners age predictions for the New Zealand synthesised voices. ....	72
Figure 15. Gender identification of the New Zealand synthesised voices. ....	73

## List of Tables

Table 1. Synthesised voice providers and communication systems used by TalkLink clients in New Zealand. ....	16
Table 2. Comparison of English accented voices available by selected synthesised voice providers .....	17
Table 3. Biographical details of voice donor participants. ....	34
Table 4. Voice donor rates of recording. ....	37
Table 5. List of final main themes and associated sub-themes.....	38
Table 6. Paired-sample t-test statistics for voice banking participant ratings of their preview synthesised voice. ....	44
Table 7. Demographical information about unfamiliar listener participants. ....	57
Table 8. Synthesised voice schedule for Group A and B. ....	58
Table 9. Perceptual experiment schedule.....	62
Table 10. T-test statistics for male and female synthesised voices.....	65
Table 11. T-test statistics for sentence A and B in the intelligibility visual analogue measure. ....	68
Table 12. T-test statistics for first heard sentence and second heard sentence in the intelligibility visual analogue measure. ....	68
Table 13. T-test statistics for sentence A and B in the naturalness visual analogue measure. ....	70
Table 14. Names of the New Zealand synthesised voices. ....	84

## Chapter 1. Introduction

Communication is a human right and a way for individuals to express themselves and connect with others (Beukelman & Mirenda, 2013; Nathanson, 2017). Speech plays a significant role in the communication process; many people use speech as their primary modality to communicate. That being said, there are other options for communication if natural speech is unavailable. Throughout their lives, people can lose their ability to speak for a number of reasons. Some individuals are born with conditions which reduce their ability to speak and some people lose their voices as a result of an acquired condition. There are multiple augmentative and alternative communication (AAC) approaches that are used to support communication. The use of high-tech devices with speech generating output capabilities have a wide range of benefits and challenges for individuals and their communication partners (Beukelman & Mirenda, 2013; Lasker & Bedrosian, 2001; Marthy, Yorkston, & Gutmann, 2000; Palmer, Enderby, & Hawley, 2010). Technological advances continue to provide the opportunity for personalisation of speech generating devices (Costello, 2000; Creer, Cunningham, Green, & Yamagishi, 2013). Individuals can use locally-accented voices on their speech generating device (SGD) or can record their own speech for use on their device. Currently in New Zealand there are limited choices for locally-accented voices for SGD users. There is also limited research into the perspectives of those who voice bank: the process of recording one's voice to create a synthetic voice for use on SGDs. This thesis addresses the gap in New Zealand-accented synthetic voice availability and gathers perceptions about the experience of voice banking. The remainder of this chapter will focus on populations affected by speech loss and impairment, an overview of SGDs, and address the benefits and challenges faced by SGD users. Factors that influence acceptability of devices related to voice output options will be discussed. Chapter Two will continue the discussion of personalising synthetic voices within the New Zealand context. Chapter Three describes the voice banking methodology and Chapter Four reports the voice banking results. The perceptual experiment methodology is described in Chapter Five and Chapter Six reports the perceptual experiment results. The findings are discussed in Chapter Seven.

### **1.1. Populations affected by speech loss and impairment**

The diversity of people who use AAC devices as their means of communication is vast. Creer, Enderby, Judge, and John (2016) found that 97.8% of the total number of people who could benefit from AAC have one of nine medical conditions: dementia, Parkinson's disease, autism, learning disability, stroke, cerebral palsy, head injury, multiple sclerosis or motor neurone

disease. It was estimated that in the United Kingdom, approximately 0.5% of the population could benefit from AAC (Creer et al., 2016). Sutherland, Gillon, and Yoder (2005) estimated that 0.15% of people aged under 21 years in New Zealand use AAC.

### **1.1.1. Motor Neurone Disease**

Motor neurone disease (MND) is a progressive neurological disease involving the motor neurons of the brain and spinal cord (Yorkston, Beukelman, Strand, & Hakel, 2010). Individuals with MND experience motor neuron degeneration which affects the ability of the brain and the spinal cord to send messages to the muscles. As the disease progresses, there is increased weakness and wasting of muscles, typically in the limbs. Some people experience weakness and wasting of muscles of the face, throat, and chest which in turn causes difficulty with speech (dysarthria) and difficulty chewing and swallowing (dysphagia) (Yorkston et al., 2010). There are three main subtypes of MND: Amyotrophic Lateral Sclerosis, also known as Lou Gehrig's Disease; Progressive Muscular Atrophy; and Progressive Bulbar Palsy (Murphy, 2004). These differ in the symptoms that individuals may present with initially. Care for individuals with MND focuses on maintaining functional ability and quality of life for as long as possible with the support of multidisciplinary team management (National Institute for Health and Care Excellence, 2016). Multidisciplinary team care may prolong the survival of individuals with bulbar MND by nine months (Simon et al., 2015). Maximizing speech intelligibility and setting up future AAC strategies and education for the individual and family are the primary communication goals for speech therapy (Britton, Cleary, & Miller, 2013). Initial changes in communication include differences in voice quality and the presence of dysarthria. This can include changes in speech related to muscle weakness, presenting as imprecise articulation, decreased speech rate, and fatigued speech which is less intelligible over time. For some individuals, frontotemporal dementia causes changes in memory and cognition which in turn influences receptive and expressive communication (Yorkston et al., 2010). As the disease progresses into the late stages, individuals may have no functional verbal communication and rely exclusively on AAC strategies. Approximately 80% of people with MND will use some form of AAC across the span of the disease (Ball, Beukelman, & Pattee, 2004). People with MND are typically ideal candidates for high-tech AAC devices with speech generating capabilities (Murphy, 2004). Early education about the options available for people to record their own voice before the impact of dysarthria is vital, and it is recommended that people record as soon as possible following diagnosis (Costello, 2016). Recording one's voice can be important for people as a way of leaving legacy messages for their friends and family,

and also to use recordings to create personalised synthetic voices for their SGD (Costello, 2016). People with MND who complete voice banking before they experience moderate dysarthria fare best in terms of the quality of their synthetic voice (Bunnell, Lilley, & McGrath, 2017).

### **1.1.2. Dementia**

Dementia is a syndrome characterised by an acquired cognitive impairment affecting memory, thinking, and social abilities which interfere with daily functioning (Beukelman & Mirenda, 2013). Alzheimer's disease is the most common type of dementia. Currently, 10% of people aged 65-84 years and 47% of people 85 years and older are diagnosed with Alzheimer's disease (Beukelman, Fager, Ball, & Dietz, 2007). The number of people with dementia is projected to grow considerably over the coming years. Dementia can impact an individual's executive function, attention, organisation, visuospatial function, language expression and comprehension (Beukelman & Mirenda, 2013). The primary focus of intervention for individuals with dementia is to enhance their strengths at each stage of progression (Beukelman & Mirenda, 2013). Communication books are common AAC strategies to aid communication and to orientate the person to the current topic. Memory books collate and present information about an individual such as their routines, interests and personality, and can aid memory and provide platforms for conversation (Bourgeois, 2007).

### **1.1.3. Parkinson's Disease**

Parkinson's disease is a progressive neurological condition that can take many years to develop. The main three motor symptoms include tremor, stiffness, and slowness of movement (Parkinson's New Zealand, 2018). The causes and triggers of Parkinson's disease are unknown and many of the symptoms occur following decreased levels of dopamine in the brain. People with Parkinson's disease experience a different number and combination of symptoms (Parkinson's New Zealand, 2018). Approximately 1 in 500 people have Parkinson's disease, and it becomes more prevalent in older age groups (Parkinson's New Zealand, 2018). It is believed that 1% of people in New Zealand aged over 60 years have Parkinson's disease, and the average age of diagnosis is 59 years (Parkinson's New Zealand, 2018).

Communication changes are almost inevitable for people with Parkinson's disease. Voice changes are experienced by 80–90% of people and 45–50% show alterations in articulation (Miller, Noble, Jones, & Burn, 2006). Many people experience a decrease in speech intelligibility due to imprecise consonants, reduced speech loudness, and voice breathiness. Alphabet boards, a form of low-tech AAC, can aid individuals in reducing speech rate and

provide extra information for communication partners to help in understanding the message. High-tech devices with speech generating output may be used with advanced stages of Parkinson's disease if the individual experiences loss of all functional speech.

#### **1.1.4. Aphasia**

Aphasia is an impairment that results from brain injury, usually due to cerebrovascular accident (stroke). Each day in New Zealand around 24 people will have a stroke, and approximately one third of these people will suffer some form of aphasia (Aphasia New Zealand Charitable Trust, 2010). It is estimated that there are 16,000 New Zealanders currently living with stroke-acquired aphasia (Aphasia New Zealand Charitable Trust, 2010). Aphasia can impair language production, language comprehension, or both. The Wernicke's area and Broca's area are two regions of the brain which are vital for understanding and using language to communicate. Injury to these areas of the brain leads to two main forms of aphasia: receptive aphasia and expressive aphasia. It is rare for pure receptive or expressive aphasia to occur; instead, most forms of acquired brain injury affect several regions of the brain and cause a variety of difficulties with language (Headway: The Brain Injury Association, 2018). Dysarthria and apraxia of speech can also be acquired after a stroke. Both motor speech disorders contribute to reductions in speech intelligibility.

Receptive aphasia is the impairment of language comprehension. In its most severe form, a person may not recognise spoken and written words. Most people retain some understanding, such as recognition of simple words and sentences but not complex sentences. A person with receptive aphasia may have better ability in one area of language than another. For example, they may be able to recognise written words more easily than spoken words or vice versa, or they may retain some non-verbal communication such as gestures or facial expression. Receptive aphasia also affects aspects of speech output. People may speak in long chains of words that have limited meaning or use incorrect words, and they may be unaware of their errors. Reading as a form of receptive language may also be impacted. A person may be unable to recognise individual letters and words or be unable to fully understand simple written sentences.

Expressive aphasia is an impairment in the ability to use and express language. In its most severe form, a person may not be able to produce any meaningful speech. More commonly, there is a decrease in speech fluency and words produced. The speaker may use short and simple sentences. Most people can understand language normally and are aware of their expressive difficulties. Speech output can be difficult, and many people experience word-

finding difficulties which can cause frustration. Writing as a form of expression can also be affected. For example, a person may write words with letters in the incorrect order, write incorrect words, or be unable to write simple sentences or certain letters of the alphabet.

Traditionally, aphasia interventions focus on restoration of functional communication by reducing language impairment (Beukelman et al., 2007). Some people require compensatory support from a variety of AAC strategies, such as drawing, use of gesture, writing, communication books and other low-tech aids, or high-tech devices. These compensatory communication strategies aim to support communication between individuals with aphasia and their communication partners (Beukelman et al., 2007). There has been an increase in high-tech interventions for individuals with severe aphasia. Most of these AAC systems focus on supporting specific communication tasks such as requesting items or help, communicating feelings, and practising everyday scripted conversations (Beukelman & Mirenda, 2013). A recent study by Hux, Knollman-Porter, Brown, and Wallace (2017) reported the use of text-to-speech technology may help people with aphasia by providing simultaneous written and auditory cues to aid comprehension of written materials.

#### **1.1.5. Cerebral Palsy**

Cerebral palsy is a term used to describe a group of conditions which affect movement and posture (Cerebral Palsy Society of New Zealand, 2018). Cerebral palsy is caused by a defect or lesion in one or more specific areas of the brain, usually occurring during foetal development or infancy (Cerebral Palsy Society of New Zealand, 2018). The prevalence of cerebral palsy is 2–2.5 per 1,000 live births (Cerebral Palsy Society of New Zealand, 2018). Approximately 7,000 people in New Zealand has some degree of cerebral palsy and one third are under 21 years of age (Cerebral Palsy Society of New Zealand, 2018). Cerebral palsy is incurable, however training and therapy can help people to improve functions of movement and posture. People with cerebral palsy experience difficulties such as weakness, stiffness, abnormal muscle tone, and difficulty with balance (Cerebral Palsy Society of New Zealand, 2018). Spastic cerebral palsy is the most common type of cerebral palsy, occurring in 70–80% of cases and accompanying the other types in 30% of cases (Cerebral Palsy Society of New Zealand, 2018).

The incidence of dysarthria is estimated to occur in a significant proportion of all people with cerebral palsy (Beukelman & Mirenda, 2013). Individuals may experience poor respiratory control as a result of musculature weakness, laryngeal and velopharyngeal dysfunction, and articulation difficulties from the restricted movement of the muscles of the face and throat (Beukelman & Mirenda, 2013). Children with cerebral palsy may experience a

delay in language development and difficulty learning to read and spell (Pennington, 2008). With the wide variety of motor impairments in this population, the involvement of many professionals is vital for determining the appropriate communication system for individuals across time. A Scottish study reported that 55% of people who used any AAC system had diagnoses of cerebral palsy and that the majority of this group (81%) used high-tech systems (McCall & Moodie, 1998).

#### **1.1.6. Autism Spectrum Disorder**

Autism Spectrum Disorder (ASD) is a life-long developmental disability affecting social and communication skills (Autism New Zealand, 2018). ASD affects 1 in 66 people, which is approximately 65,000 New Zealanders (Autism New Zealand, 2018). People with ASD may experience a wide range of difficulties with understanding and using language. This can include difficulties with receptive communication, expressive communication, joint attention, symbol use, and social pragmatic skills (Ministry of Health and Education, 2008). Early intervention emphasises the development of pragmatics and functional communication skills.

The Picture Exchange Communication System (PECS) is a form of AAC that helps individuals with ASD communicate using pictures. The purpose of PECS is for individuals with ASD to learn how to initiate communication with others through exchanging a picture with a communication partner for their desired object. It supplies the child with an initial symbol-based way to facilitate emerging functional communication which can lead to an increase in speech production over time (Carr & Felce, 2007). SGDs are also used with people with ASD and can support literacy development, communication skills related to requesting, answering questions, and natural speech production (Beukelman & Mirenda, 2013).

#### **1.2. Augmentative and alternative communication**

AAC systems supplement existing speech or replace speech that is not functional to support an individual's ability to communicate. AAC systems are multimodal systems, designed to allow individuals to use every mode possible to communicate (ASHA, 2011). It allows for changes over time in response to the individual's language and physical needs, as well as changes based on context, audience, and communicative intent (ASHA, 2011; Beukelman & Mirenda, 2013). For children who use AAC, many systems can support language acquisition and verbal expression (ASHA, 2011). AAC is typically divided into two categories: unaided and aided. Unaided communication systems rely on the person's body to convey messages such as through the use of gestures, body language or sign language. Aided communication systems require the use of tools or equipment in addition to the user's body. These can be sub-divided

into low-tech and high-tech aids. Low-tech communication aids are defined as those that do not need batteries or electricity to function. They are often simple aids created with words or pictures on a board or book. Depending on physical abilities, the user may indicate the appropriate message with a limb, their head or pointing device. Alternatively they may indicate their choice through a *yes* or *no* as a communication partner scans through the possible options (ASHA, 2011). High-tech aids are electronic devices which allow storage and retrieval of messages. These can range from single message devices through to SGDs.

### **1.3. Speech generating devices**

Also known as voice output communication aids, SGDs are a relatively recent addition to the repertoire of AAC options (Schlosser, 2003). SGDs are used to supplement or replace speech or writing for individuals, enabling them to verbally communicate (ASHA, 2018). SGDs allow the user to input a message, which is then verbalised by the voice output device (Creer, Cunningham, Green, & Yamagishi, 2013). SGDs can either be dedicated communication devices with the sole purpose of communication (e.g., Accent® and Indi™ devices) (Prentke Romich Company, 2018; Tobii Dynavox, 2018) or non-dedicated devices (e.g., tablets) which have been adapted as a communication tool but can also be used for other functions. Each device requires a communication system. There are many applications which form the communication systems for devices. Systems commonly used by TalkLink, the national assistive technology provider, are discussed further in Chapter Two. Most communication systems have dynamic displays where multiple pages of symbols are available, and the communicator navigates the various pages. There are a wide variety of access methods that people use. As with low-tech aids, selection methods are dependent on the abilities of the communicator. Access methods include direct selection on a screen or keyboard with a body part or pointer, and indirect selection using switches or partner assisted scanning (ASHA, 2011).

SGDs use digitised speech output, synthesised speech output, or both. Digitised speech consists of natural speech that has been recorded, stored, and reproduced. This includes phrase banking where a person can record whole words or phrases to store and then play back on the SGD (Creer, Green, & Cunningham, 2009; Shannon, 2016). The natural voice, inflection, and intonation are recorded and reproduced by the SGD (Costello, 2016). There is reduced versatility in creating novel sentences as the SGD has only a limited library of pre-recorded phrases that can be verbalised (Creer et al., 2009). Synthesised speech is produced electronically. Phonemes and allophones of the target language are used to generate digital

speech signals that are transformed into intelligible speech (Beukelman & Mirenda, 2013). This allows SGD users to input any message into their device for the voice output function to verbalise. There are processes that allow the vocal characteristics of the speaker's former or residual speech to be synthesised. Voice banking allows the creation of a synthetic voice that approximates the speaker's natural voice and is discussed further in Chapter Two (Costello, 2016).

#### **1.4. Experiences of using speech generating devices**

SGDs have been shown to have positive influences on quality of life and communication successes (Marthy et al., 2000; Palmer et al., 2010). They allow communication to occur over distance and for people to send messages without first obtaining their communication partner's attention (Beukelman & Mirenda, 2013). SGD users are able to communicate with unfamiliar listeners, people who are non-literate, and those with vision impairments (Beukelman & Mirenda, 2013). Communication can occur in a familiar and non-threatening way for unfamiliar listeners and is functional for use in many day-to-day activities (Beukelman & Mirenda, 2013). SGDs have the potential to allow the user to become an active participant in communication exchanges. AAC systems, including SGDs, are often introduced as an early intervention to support communication, language, literacy, and natural speech development (Blischak, Lombardino, & Dyson, 2003; Schlosser, 2003). SGDs can also assist graphic symbol learning, requesting, social regulation, learner preferences, and challenging behaviours (Schlosser, 2003). Children who use SGDs are able to experience more natural interaction with peers who are typically more accustomed to verbal conversational exchanges (Drager, Reichle, & Pinkoski, 2010). This can support the development of social skills and friendships (Drager et al., 2010). Children who use SGDs also may experience increased independence in communication with a wider range of peers (Drager et al., 2010).

However, there are multiple challenges that SGD users encounter. Firstly, there is a range of perceptions and attitudes that listeners can have towards people who use SGDs. Gorenflo and Gorenflo (1991) found that attitudes of young adult listeners, measured with the *Attitudes Towards Nonspeaking Persons* scale, improved if the AAC system included speech output, compared to an AAC system without speech output. Bedrosian, Hoag, Calculator and Molineux (1992) studied the perceived communicative competence of a person simulating the use of an SGD with digitised speech output and found that the person was perceived as more competent when using phrases compared to using single-word messages. Natural verbal communication is usually rapid, and interactions when using SGDs can be significantly slower

(Palmer et al., 2010). Although a slower speed of communication can result in more successful message delivery, rapid speed is more natural as it approximates typical verbal communication (Lasker & Bedrosian, 2001; Palmer et al., 2010). It is unrealistic to expect most SGD users to engage in natural and fluid conversation while using their device (Stern, 2008). SGD users often speak less than ten words per minute, and text entry creates a high wait time for communication partners (Howey, 2017). Videos which portray SGD users having a conversation with responses appearing in real time often are using pre-programmed utterances. Online text entry for many users is slow, especially for people who have physical conditions which impact on speed and accuracy of message selection (Howey, 2017). Embarrassment and negative attitudes can also create barriers for SGD users (Miller et al., 2006). People who use SGDs can worry that their AAC device may remove the closeness of their personal interactions (Murphy, 2004). Although technologies are constantly evolving and improving, and vital to the existence of SGDs, technology itself can also be a barrier for SGD users. SGDs can be a challenge for new users and their communication partners due to setup processes, and maintenance and programming required of the devices (Baxter, Enderby, Evans, & Judge, 2012; McCall, Marková, Murphy, Moodie, & Collins, 1997). It also takes time and practice to use the technology (Howey, 2017). This technology is reliant on battery power, so keeping the device constantly charged is important (Howey, 2017). The highest rated item by users of SGDs in the communication aid protocol in O'Keefe, Brown, and Schuller (1998) was *battery allowing full day of use*. Howey (2017) recounted stories of SGD users who experienced conversations being cut short mid-sentence because their device had run out of battery.

### **1.5. Aspects which influence device acceptability**

Once an SGD has been introduced and is seen to meet all the individual's communication needs, acceptance of the device is not guaranteed (Lasker & Bedrosian, 2001; Marthy et al., 2000). The acceptability of an SGD is influenced by many factors (Creer et al., 2013; Lasker & Bedrosian, 2001; Marthy et al., 2000). Acceptance often depends on the motivation and attitudes of users and conversation partners, and the features of the device itself (Lasker & Bedrosian, 2001; Marthy et al., 2000). Devices need be reliable, accessible, and flexible to adaptation as the individual's communication needs and physical needs change, or else the likelihood of abandonment is high (Murphy, 2004). SGD acceptance and use has increased considerably over the past decade (Ball et al., 2004; Beukelman et al., 2007). In 1996, acceptance rates of individuals with MND who were recommended SGDs were 73%, while a 2004 study reported a 96% acceptance rate (Ball et al., 2004; Beukelman et al., 2007). Along

with higher acceptance rates, people with MND often continue to use their SGDs throughout the end stages of the disease. Without the use of AAC, 80 – 95% of people with MND would be unable to meet their daily communication needs (Ball et al., 2004; Beukelman et al., 2007). Lasker and Bedrosian (2001) define optimum acceptance of a communication aid as being where the system is used willingly at every opportunity. Lasker and Bedrosian (2001) proposed an AAC acceptance model based on success factors divided into three main categories: milieu (social environment), person, and technology. Voice quality and customisation are two success factors included in the technology category (Lasker & Bedrosian, 2001). These two factors are particularly relevant to SGDs where attempts are made to provide a realistic replication of person to person spoken communication (Creer, 2009). ASHA (2018) recommends that SGDs are set up with a voice that is appropriate to the user’s age, gender, and race or ethnicity, and aligns with the user’s preferences. The following sections detail the acceptability of SGDs in terms of aspects related to the output voice.

### **1.5.1. Intelligibility**

The primary importance of the SGD voice output function is whether the average listener can understand the voice (Creer, 2009; Stern, 2008). Stern (2008) described three components of understandability: intelligibility, comprehension, and naturalness. Intelligibility can be described as the degree to which speech is understood (Anand & Stepp, 2015). Palmer et al. (2010) indicate that AAC devices are most successful and accepted when there is maximum retention of the speed, naturalness, and responsiveness of spoken communication, and when the device adds to intelligibility. Beukelman, Fager, Ball, and Dietz (2007) indicated that speech synthesis technology often interferes with the communication partner successfully understanding messages produced, so the need for natural-sounding intelligible speech output is essential. There is indication such as in Murphy (2004) that sometimes people who use SGDs have been provided with devices which have poor quality voices. Murphy (2004) discuss an example where a partner of an SGD user commented that the voice on her husband’s device was less intelligible than his dysarthric speech. The mean importance ratings by users of SGDs for the two communication aid protocol items *voice is clear* and *aid message is easy to understand* in O’Keefe, Brown, and Schuller (1998) was 4.95 out of a possible 5 points. Voice output that is intelligible offers a number of advantages for communication partners, including ease of understanding messages produced in a familiar modality (Beukelman & Mirenda, 2013). Because there are a range of ages of people who use SGDs and a range of communication partner ages, there is a continuing need for synthesised speech that can be

easily understood by all people (Beukelman, Fager, Ball, & Dietz, 2007). Beukelman et al. (2007) comment that people using SGDs are often in environments that are less than ideal listening conditions such as schools, rest homes, or other places with distractions and high background noise.

Studies which compare synthetic and natural speech reveal that people identify and respond to natural speech more accurately and quickly than to synthetic speech (Koul & Allen, 1993; Koul, 2003; Logan, Greene, & Pisoni, 1989; Miranda & Beukelman, 1987). In the past, most studies of synthesised speech intelligibility investigated listeners' comprehension of DECTalk, a speech synthesiser that was reported superior in intelligibility compared with other commercially available synthetic voices at the time (Duffy & Pisoni, 1992; Von Berg et al., 2009). Percent intelligibility for DECTalk in single word intelligibility tasks have ranged from 81.7% correct with an open response format to 96.7% correct with a closed response format (Greene, Manous, & Pisoni, 1984; Miranda & Beukelman, 1987). In contrast, word intelligibility scores for natural speech have ranged from 97.2% with an open response format to 99% with a closed response format (Logan et al., 1989). Von Berg et al. (2009) investigated the intelligibility of two types of popular adult and child synthesised speech systems in the late 2000s: DECTalk and VeriVox. The experiment used an open response format and the synthesised voice with the highest intelligibility reached a mean of 96.3% accuracy. This indicated that intelligibility of synthesised speech was increasing over time as synthesis technology advanced. The results did reveal differences in intelligibility between synthesised adult and child voices. The mean intelligibility scores for VeriVox and DECTalk child synthesised voices (71.9% and 55.2%) were significantly below those of the adult synthesised voices; the lowest-scoring adult voice was 85.0% intelligible. There were no significant differences found between male and female synthetic voices. Jreige, Patel, and Bunnell (2009) conducted a usability study for a small sample of VocaliD voices. VocaliD voices are created by combining speech sounds from individuals with speech impairments and healthy age and gender matched voice donors to create personalised synthetic voices. This process is discussed further in Chapter Two. Results indicated that intelligibility ranged from 94.0% to 97.6% for the eight synthesised adult voices. This is comparable with Von Berg et al. (2009) results with the commercially available synthetic voices. Jreige et al. (2009) also found no significant differences in intelligibility between male and female voices.

Synthesised speech can be more difficult to process for the elderly, those with intellectual or language impairments, and those who are communicating in a second language (McCord & Soto, 2004; Palmer et al., 2010). McNaughton, Fallon, Tod, Weiner and Neisworth

(1994) suggested three ways to increase the intelligibility of synthesised speech: increase quality and naturalness of the voice, use sentences and phrases to provide contextual information, and train the listener to recognise the patterns of the synthesised speech over time. As new speech synthesis voices are created and used with SGDs, the effectiveness of these voices should be investigated and reported (Beukelman, Fager, Ball, & Dietz, 2007).

### **1.5.2. Naturalness**

One criticism of synthesised voice output is that it is still not as natural-sounding as natural speech itself (Creer, 2009). This can cause negative attitudes of users and communication partners relating to the naturalness of their interactions. The naturalness of synthesised speech may impact the acceptance and use of SGDs (Beukelman, Fager, Ball, & Dietz, 2007). There is a preference for voice output that is perceived as more natural (Ratcliff, Coughlin, & Lehman, 2002). The AAC users in Palmer et al. (2010) indicated that computer-generated speech was acceptable but recorded human speech was preferable due to its natural sound. This is consistent with earlier research by Crabtree, Mirenda and Beukelman (1990), which found that natural voices were preferred over synthetic voices. People who use SGDs often have very little expressive control over their tone of voice (Pullin & Hennig, 2015). Expression and tone of voice are important in communication where the change in intonation or expression can change the meaning of a sentence. Beukelman et al. (2007) found that adults with acquired neurological conditions who use SGDs reported cases of miscommunication due to lack of naturalness, and that some voice output sounded impersonal and robotic. Jreige et al. (2009) included a naturalness rating in their study of VocaliD synthetic voice performance. Unfamiliar listeners rated the naturalness of each recording on a Likert scale from 0 to 5, where 0 was defined as computerised and unnatural, and 5 was defined as natural and human-sounding. The average naturalness rating across all voices was 3.5. Participants thought the voices sounded more human than computerised. Therefore, naturalness continues to be an important factor to develop in synthesised voices.

### **1.5.3. Representation of the user**

British theoretical physicist Professor Stephen Hawking was well known for the robotic, mechanical-sounding voice that he used on his device. Although this voice became part of his identity, most people who use SGDs prefer natural-sounding voices, and this increases the positive attitudes towards the SGD by the users and their conversation partners (Creer, 2009; Nathanson, 2017). In 2014, Professor Hawking trialled a new, more personalised voice with an accent closer to his native accent, however, he chose to retain the mechanised voice he had

used for the past 20 years (Nathanson, 2017). Nathanson (2017) discussed that perhaps Professor Hawking identified more with his mechanised voice than his native voice because it had been entrenched for many years with his sense of self and his public identity. The ability to speak using one's natural voice is often taken for granted until it becomes threatened by illness or disease-related treatment (Nathanson, 2017). While adjusting to the voice output of an SGD, AAC users can feel they are presenting a distorted misrepresentation of themselves, and that can create negative attitudes towards the device and can reduce acceptability (Creer et al., 2013; Nathanson, 2017). An individual needs to identify with and feel positive toward the voice that they are using in order to feel motivated to use the device (Creer et al., 2013). Children and adults want SGDs which are unique and customised to them, as it is important for users to be able to express themselves through their devices (Creer et al., 2013; Light, Page, Curran, & Pitkin, 2007). SGD voices have options for users to select the gender and the age of the voice on their device. Offering users a choice of voices for their SGD including one which matches their vocal identity pre-deterioration could lead to increased acceptance (Creer et al., 2013). Mirenda, Eicher and Beukelman (1989) indicate that although listeners' responsiveness to and attitude about synthetic speech are influenced by intelligibility, the age and gender-appropriateness of the synthesised speech can also be expected to influence listener perceptions. Mirenda, Eicher and Beukelman (1989) found that female listeners across the age ranges indicated that only natural female voices were acceptable alternatives to their own speech. Participants indicated that it was more socially acceptable for a man to use an SGD which produces male-like synthetic speech compared to a female using male-like synthetic speech. In Crabtree et al. (1990) participants preferred natural-sounding voices with the same gender as themselves. Light et al. (2007) had children design the features of AAC systems; the children emphasised that the users should be able to choose the voice on their device. If the user was a young girl, the children recommended they should use a girl's voice close in age. O'Keefe et al. (1998) found that SGD voice outputs which are easy to understand, socially appropriate, and reflective of the user's intelligence, age, and gender were factors that AAC users found were more likely to receive a favourite response from peers.

When inferring the age of a speaker from hearing the voice, a listener may rely on various cues, including the physical attributes of the voice as well as the content of what is being said (Waller et al., 2015). Although age estimates and the chronological age of a speaker are typically correlated, when predicting the age of a speaker by only hearing their voice, listeners tend to overestimate the age of younger speakers and underestimate the age of older speakers (Cerrato et al., 2000; Waller et al., 2015). This suggests a central tendency effect

(Hollingworth, 1910). Cerrato et al. (2000) carried out a study to investigate the extent to which unfamiliar listeners could estimate a speaker's age and gender from a voice recording. This Italian study included 42 male and female voices ranging in age from 18-65 years and 17 unfamiliar listeners. The youngest four age groups of speakers were estimated to be older than the voices' chronological age. The mean deviation between chronological age and age estimates was 10 years for the 18 to 24-year-old voices, 5 years for the 25 to 31-year-old voices and 3 years for the 32 to 38 and 39 to 45-year-old voices. The three older age groups were estimated to be younger than the speakers' chronological ages, by 3 years for the 46 to 52-year-old group, 5 years for the 53 to 59-year-old group and 12 years for the 60 to 65-year-old age group. All gender predictions were accurate for the 42 voices. There are differences in the literature about the extent of the deviations in age estimations. Waller et al. (2015) carried out a study which reported age estimation deviations. For speakers in the younger age group (20–30 years), a deviation of 4.8 years above the chronological age of the speaker was reported. The middle (40–50 years) and older (60–70 years) age groups were reported to have deviations below the chronological age of the speaker by 1.8 years and 12 years respectively. Waller et al. (2015) discussed that people are more accurate at estimating ages from faces than from voices, however it is important for age estimations of the voice and face of the speaker to align.

For SGD users, the restricted selection of voices limits their ability to distinguish themselves from others and to represent their personalities and identities (Creer et al., 2013). Mullennix and Stern (2010) discuss the lack of child voice options for children who use SGDs. The limited number of different voices can cause practical problems in large groups of people who may all be using the same output voice (Mills, Bunnell, & Patel, 2014). For example, in a classroom setting, it might be difficult to identify the person speaking if more than one student uses the same voice on their device. McCord and Soto (2004) discuss accent and language options being a factor which influences device acceptability. The language of the SGD was a barrier for use at home when children who used devices come from bilingual families (McCord & Soto, 2004). The lack of devices having two languages available was identified as a limiting factor, as well as synthesised speech being harder to understand for family members who did not speak English as a first language (McCord & Soto, 2004). Limited vocabulary was identified by Bailey, Parette, Stoner, Angell, and Carroll (2006) as being an obstacle to acceptability. Bailey et al. (2006) reported SGD users feeling frustrated when spelled words were mispronounced by their SGD. Providing choice and customisation for SGD voice output is important to increase positive attitudes and acceptance of the AAC systems.

## Chapter 2. Personalisation of synthetic voices for New Zealanders

A person's voice gives the listener many clues about their identity such as gender, age, ethnicity, and geographical location, and can be seen as a personal identifier of an individual (Creer et al., 2013). Despite the number of people who use SGDs to communicate, there are limited voice output options (Lasker & Bedrosian, 2001; Marthy et al., 2000; Murphy, 2004). Some children use adult voices on their SGD and many adults use voices which are not personalised to them or to the New Zealand linguistic culture. The lack of customisation can facilitate negative attitudes towards the SGD and reduce the acceptability of the device (Creer et al., 2013; Nathanson, 2017). Recent developments in the personalisation of synthesised speech, including the ability to develop a tailor-made voice for the user, hold great promise for encouraging people with speech impairments to adopt this technology (Mullennix & Stern, 2010). Voice personalisation is an issue that is often overlooked, yet may be as important as intelligibility to SGD users (Creer et al., 2013; Mullennix & Stern, 2010). This chapter will outline the New Zealand linguistic culture and current synthesised voice options, review selected methods for voice banking, and end with an outline of the primary aims, research questions and hypotheses of the present study.

### 2.1. New Zealand linguistic culture

The three official languages of New Zealand are English, te reo Māori and New Zealand Sign Language. English is spoken by 3.8 million people, which is 96.1% of the population (Statistics New Zealand, 2013). The New Zealand English dialect is separate to North American and British dialects. Although similar to Australian English, there are some distinct variations with the most significant being the short front vowel shift in New Zealand English (Crystal, 2003). One example is in the phrase *fish and chips* where Australians perceive New Zealanders to say "*fush and chups*" while New Zealanders perceive Australians saying "*feesh and cheeps*" (Crystal, 2003). The vocabulary of New Zealand English has been influenced by the adoption of te reo Māori words and phrases, and by exposure to words and idioms from North America and Australia (Ministry for Culture and Heritage, 2009).

Te reo Māori has ten consonants (p t k m n ng wh h r w), five monophthongs (a e i o u) and eight diphthongs (ae ai ao au ei oi oe ou). The vowels can be short or long, with long vowels in written te reo Māori designated by a macron. Te reo Māori is an open syllable language and consonant clusters are not permitted (Keegan, 2018). All te reo Māori words borrowed from other languages such as English are adapted to conform to the existing phonotactics (Keegan, 2018). Until the mid-19th century, te reo Māori was the predominant

language spoken in New Zealand (Ministry for Culture and Heritage, 2018). The language was increasingly confined to Māori communities as increased numbers of English speakers arrived in the country following the signing of the Treaty of Waitangi in 1840 (Ministry for Culture and Heritage, 2018). In 1987, te reo Māori was recognised as an official language of New Zealand (Ministry for Culture and Heritage, 2018). In 2013, te reo Māori was the second most spoken language with 148,000 speakers: 21% of the Māori population and 3% of all New Zealanders (Statistics New Zealand, 2013). Many proper nouns such as place names and native flora and fauna have te reo Māori origins or translations and are used in everyday conversation. SGD users in New Zealand need voice output options which have New Zealand English accents and systems which allow communication in te reo Māori.

## 2.2. Current synthesised voice options

There are limited voice options available for SGD users to distinguish themselves from others and maintain their personal identities (Creer et al., 2013). Currently most SGDs have voice output that is not personalised to the individual or to New Zealand linguistic culture. Communication systems on SGDs use voices from companies which develop text-to-speech voices. Each communication system uses different synthetic voice providers. The communication systems which TalkLink most commonly provide to clients and the available voices for each system are outlined in Table 1.

Table 1.

*Synthesised voice providers and communication systems used by TalkLink clients in New Zealand.*

<b>Synthesised voice providers</b>	<b>Communication systems</b>
Acapela	Speak For Yourself; Grid 3; Speech Assistant; Communicator 5; Nova Chat; Proloquo2go; TouchChat
CereProc	Speak For Yourself; Speech Assistant; Grid 3
Google Text-to-Speech	Speech Assistant
Ivona	Nova Chat; LAMP
iOS	ClaraCom; ChatAble; Predictable
Nuance	ChatAble; Predictable; Grid 3

*Note.* Information collated from A-Soft Software (2018), Acapela Group (2018), Apple (2018), AssistiveWare (2017), CereProc (2018), Desktop Technology Services Limited (2018), Google (2018), Ivona Software (2018), Saltillo Corporation (2018), Smartbox Assistive Technology (2018), Speak For Yourself (2015), Therapy Box (2018), Tobii Dynavox (2016), and TouchChat (2018).

No synthesised voice provider has a New Zealand English voice option. The English voices that are available are predominantly North American or British-accented with a number of systems also providing Australian options. Current synthetic voices are not able to accurately pronounce te reo Māori words. The three most common synthesised voice providers and their available English accented voices are outlined in Table 2.

Table 2.

*Comparision of English accented voices available by selected synthsised voice providers.*

<b>English voices</b>	<b>Acapela voices</b>	<b>CereProc voices</b>	<b>Nuance voices</b>
<b>Adult male</b>	4x American; 7x novelty 3x British; 2x novelty 1x Australian - - -	4x American; 3x novelty 3x British - 2x Scottish male - - -	2x American 3x British 1x Australian - - 1x Indian male
<b>Adult female</b>	2x American 2x British; 1x novelty 1x Australian 1x Scottish female - 1x Indian female -	4x American female 4x British - 1x Scottish female 1x Irish female - -	5x American 3x British 1x Australian 1x Scottish female 1x Irish female 3x Indian female 1x South African female
<b>Male child</b>	2x American 1x British 1x Australian	- - -	- - -
<b>Female child</b>	2x American 1x British 1x Australian	- - -	2x American - -

*Note.* Novelty voices and expressive voices includes those such as *Will (Bad Guy)* Acapela voice and *Calm and friendly Claire* CereProc voice. Information collated from Acapela Group (2018), CereProc (2018), and Harpo Software (2018).

### 2.3. Voice banking

At the Boston Children’s Hospital (Massachusetts, USA), a unique model of AAC intervention was introduced in 1994 (Costello, 2000). Some children who undergo surgery or medical

procedures have a temporary inability to speak post-operation secondary to procedures such as intubation, tracheostomy, and/or mechanical ventilation (Costello, 2000). Children and their families often experience stress and anxiety in the intensive care environment, in addition to the inability to communicate (Costello, 2000). The process of personalised recordings of an individual's speech emerged to aid the children and their families during recovery in intensive care. At the pre-operative visit, the patient's own voice was digitally recorded into an SGD. This phrase banking process created a digital vocabulary of 30–60 easily accessible messages for the intensive care environment. The children practiced how to access the recordings on the SGD so they became familiar with the communication method pre-operation. Having access to an SGD with a personalised voice helped the children to express their feelings, reduce frustration, and maintain some control in the medical environment (Costello, 2000).

This personalised model of AAC has expanded into use with patients with neurodegenerative conditions such as MND. Progressing from the phrase banking methods, voice banking is a recent type of digitised speech which allows people to bank their own voice (Costello, 2016; Creer et al., 2013). This uses a different approach than phrase banking and creates a fully synthesised voice which can produce any utterance (Creer et al., 2009). Voice banking requires a dataset to be recorded that contains all the sounds in the target language. The dataset is segmented into smaller units which can be recombined to produce new utterances (Creer et al., 2009). This creates an authentic, personalised voice for an individual to use on their SGD. Voice banking was first introduced for use with populations such as those with MND (Costello, 2016; Creer et al., 2013). Voice banking is an important way that SGD users can maintain their own personality and identity when they can no longer speak using their own voice. Chapter One discussed many benefits of SGDs and the need for personalisation. Creer et al. (2009) reviewed the voice banking technologies that were available at the time. Due to the high cost, voice building was reserved mainly for providing voices to companies. The Acapela group (discussed further in section 2.3.2) created synthetic voices from a large database of recordings produced by professional speakers with high quality recording equipment, but were yet to start voice banking options for patients (Acapela Group, 2017; Creer et al., 2009). CereProc (discussed further in section 2.3.1) estimated that a voice built specifically for a company would cost upward of £20,000 (\$38,037.47 NZD March 2018) making it non-viable for individuals (CereProc, 2018; Creer et al., 2009). At the time, CereProc were looking for investments for production of software to allow voice banking for individuals with progressive speech loss; the market became aware of the need for personalised synthesised voices. In 2009, ModelTalker (discussed in section 2.3.4) was the only voice banking software

specifically aimed at people with progressive speech loss (Bunnell et al., 2017; Creer et al., 2009; ModelTalker, 2018). The process required around 1,600 utterances to be recorded with the software prompting the individual to produce an utterance and then screening it for consistency of pitch, loudness, and pronunciation. Today there are multiple online voice banking options, including CereVoice Me by Cereproc, My-own-voice by Acapela, Legacy and BeSpoke voices by VocaliD, and ModelTalker (Acapela Group, 2018; Bunnell et al., 2017; CereProc, 2018; Jreige et al., 2009; ModelTalker, 2018; VocaliD, 2018). The summaries discussed below are based on Eriksen (2018) honours research, which served as a preliminary project for the current study.

### **2.3.1. CereVoice Me**

In 2006 the Scottish company CereProc started the CereVoice Me voice cloning service. Today, CereVoice Me online voice banking service creates a text-to-speech voice compatible with the individual's choice of Windows or iOS systems for £499.99 (\$950.92 NZD March 2018) (CereProc, 2018). The process involves recording 620 sentences, over an average duration of 1.5–2 hours. Languages offered by the software include English, French, Spanish, Italian, German, Portuguese, Japanese, Dutch, and Catalan. There is a custom inventory feature available upon request; however, this is only compatible with the languages listed above. CereProc provides individuals with headset microphones to use during the recording process. The software supports the Firefox, Chrome, and Opera browsers, and the company specifies recordings to be completed in rooms free of background noise, preferably studio-recorded. The staff are involved in the process by providing feedback for ten screening sentences before the individual begins the full inventory, to ensure the quality of the recordings.

### **2.3.2. My-own-voice**

My-own-voice is a voice banking service provided by the Acapela Group from 2014. The Acapela Group was established in 2003 after a merger of three companies from Belgium, France and Sweden (Acapela Group, 2018). My-own-voice is free to record and trial on the Acapela website however costs €3,000 (\$5,086.33 NZD March 2018) to download onto a communication device. The recording process takes 5–8 hours to record 1,500 sentences. Many languages apart from English are offered including Dutch, French, German, Italian, Norwegian, Spanish, and Swedish. For the English language, British, American, and Australian accents are supported. Although the user can determine specific pronunciation for words for which the default pronunciation is not accurate, there is currently no option to add the user's own words or words from languages not supported by the software. My-own-voice can be used

on Windows and Android devices as well on any of the communication systems compatible with Acapela voices. Recording takes place via the Firefox web browser and it is recommended to carry out the recordings in a professional studio. If this is not possible, Acapela recommends a quiet room, noise-free without reverberation. There is a screening process where staff will give feedback about recordings before individuals continue with the full recording process.

### **2.3.3. VocaliD**

VocaliD is another online voice banking company (Jreige et al., 2009; VocaliD, 2018). Founded in the USA in 2014, there are two voice banking products available. The first is the Bespoke voice, designed for people with speech impairments who can record three seconds of sound. A personalised digital voice for text-to-speech applications is created using the vocalisations and several hours of recordings from a matched speech donor in *The Human Voicebank*. This is a spoken repository where over 14,000 voice donors from 110 countries have contributed over six million sentences. When Bespoke voices are requested the voice recordings will be age and gender-matched and combined with donor voices from *The Human Voicebank*. This service costs \$1,499 USD (\$2,063.57 NZD March 2018) and voices can be used on most iOS, Windows, and Android applications. Google Chrome is the recommended web browser for use when recording. A room free of background noise and no distractions is indicated as the optimum recording environment.

The second type of VocaliD service is Vocal Legacy. This is focused on voice preservation for individuals are able to record several hours of speech. A digital representation of the individual's speech is created from 3,500 sentences recorded by the individual across a duration of 7–8 hours. A screening process takes place with ten sentences that are assessed by the staff for quality before the individual continues the full recording process. Currently English is the only language offered and all accents are supported. Words from other languages cannot be added to a personalised inventory. There is a visual representation of the English phonetic inventory on screen. Once the user records specific sounds and combinations of sounds, the interface lights up the corresponding phonetic symbols that have been recorded. Downloading the completed voice costs \$1,199 USD (\$1,650.58 NZD March 2018).

### **2.3.4. ModelTalker**

ModelTalker began in 2005 and is an online voice banking service aimed at people with progressive speech loss (Bunnell et al., 2017; ModelTalker, 2018). ModelTalker is currently the only voice banking software which has been used by TalkLink with a handful of New Zealand clients with MND. Developed by the Nemours Speech Research Laboratory in

Delaware, USA, ModelTalker requires an individual to record 1,600 utterances which are synthesised to create the personalised voice. From about 400 sentences a usable voice can be created, however the quality of the voice increases with number of sentences recorded. This provides options for people who are unable to complete all 1,600 sentences, which would typically take 6–8 hours. On-screen pronunciation guides are optimised for American English. The software screens the recordings for consistency of pitch, loudness, and pronunciation in attempt to control the quality of the recordings. ModelTalker recommends recording with a USB headset microphone. Individuals can choose to voice bank or to donate their voice. Individuals who voice bank are those who are creating a voice for use on an SGD. This service was free until mid-2017 when a \$100 USD (\$137.66 NZD March 2018) charge was put into place for users to enable the download of completed voices. This fee is covered by the Ministry of Health for eligible clients in New Zealand. For healthy speakers wanting to donate their voices for other people to use, ModelTalker is free to use to record and download their completed voice. The personalised voice is able to be used on Windows and Android devices; however, only iOS communication systems by Therapy Box (Predictable and ChatAble) are compatible with the ModelTalker voice. Custom words and phrases are able to be added and are recorded via phrase banking, meaning that any languages and pronunciations are able to be recorded in these personal inventories. Creer et al. (2009) trialled the ModelTalker voice banking system and commented that the system was optimised for American English and the pronunciation dial may have to be switched off for speakers of other English accents. The researchers also indicated that the online interface was easy to use and there was minimal wait time between the recordings being submitted for synthesis and the personalised voice being ready for use.

Four speech therapy graduate students at the University of Kansas trialled the ModelTalker process (Jackson, Foutch, Roberts, Duff, & Collins, 2016). The students did not use a headset microphone and instead used the built-in microphones on their Apple MacBook laptops aged 2010–2013. The students each recorded the ten screening sentences at their individual homes. Two of the four students' recordings did not pass the screening on the first attempt and were given feedback to reduce background noise. It took up to four trials for the two remaining students to pass the screening. The students recorded the first 200 sentences of the full ModelTalker inventory. Recording 200 sentences was reported to take 1–1.5 hours. The synthesised voices created from the small number of recordings were reported to not be of high quality, although no perceptual analysis was carried out.

The following year, five different graduate students repeated this study with alterations in methodology to follow the recording guidelines on the ModelTalker website (Jackson et al., 2017). The five students used a Sennheiser PC-36 USB headset microphone and various MacBooks aged 2012–2016. All students recorded in a small carpeted room at the Hearing and Speech department of the University of Kansas. All the students' screens passed on the first attempt. They recorded 400 sentences of the full inventory, which was the minimum number of sentences required for synthetic voices to be created. Recording 400 sentences took approximately two hours for each student. On completion of the 400 sentences and the creation of the synthetic voices, the students reported that their peers who were not involved in the study gave feedback that the synthesised voices sounded like the students who recorded them. The researchers indicated that not all phonemes in the synthesised voices sounded as clear as in natural speech which could have negative effects on speech intelligibility. However, no quantitative analysis was carried out on the synthetic voices. The students compiled a list of advantages and disadvantages for the process of voice banking with ModelTalker. Similar to Creer et al. (2009), the students thought the process was timely and easy and the staff at ModelTalker were helpful when answering queries. The students indicated there were some unclear instructions, such as how to start recording after the screening was passed. They thought emailing the staff seemed redundant, and suggested a chat feature as part of the interface. The students summarised a number of clinical considerations for using ModelTalker with clients. The first consideration was around early education and starting the voice banking process early. Consistent with Costello (2016), the researchers indicated that clinicians should inform clients with progressive speech impairments about the options for voice banking before the quality of the synthetic voice is affected with dysarthria and degeneration of the voice. The computer skills of the client should also be considered. The screening, recording, and installation process requires an individual to be confident with navigating the recording software, downloading files, installing programs, checking sound settings, and using email. The researchers suggested creating a handout with step-by-step instructions and visual aids to assist clients. Time commitments were discussed with the projected time to record 1,600 sentences being eight hours, using the students' rate of recording 400 sentences in two hours. The researchers recommended using the Sennheiser PC-36 headset microphone and suggested that clinics which provide voice banking services should lend headsets out for client use.

Two undergraduate speech therapy students at the University of Minnesota Duluth created and trialled a home-made portable recording booth in order to find out the feasibility of using this booth with clients at home to reduce the demands for the overbooked audiology

sound booth (Hyppa-Martin et al., 2017). The University of Minnesota Duluth voice banking clinic for individuals with MND found that most clients preferred morning appointments when they felt less fatigued and some clients travelled up to two hours to record at the sound booth. A low-cost portable sound booth was proposed, in order to begin a home-based voice banking outreach programme. The booth was made from three foam panels arranged in a C shape around a desk and chair; however, the researchers did not specify the type or cost of the foam. The booth was designed to be used in a quiet room of a house with limited background noise. The foam's purpose was to reduce the resonance and reverberation of sound in the room to increase speech clarity and quality. To assess if it was feasible to use this set up for voice banking, the researchers trialled the ModelTalker voice banking protocol with a healthy adult female in both recording environments: the portable booth set up in a house, and a sound-treated audiology booth. The synthetic voices created from each environment were informally compared to the speaker's natural voice. The researchers indicated that the synthetic voices from both environments sounded similar, however, acoustic analysis would have gained more specific and scientific comparisons of the two voices. The researchers compiled a list of tips for voice banking following their experiences with using ModelTalker. Similar to Jackson et al. (2016), there was discussion of the importance of reducing background noise. Suggestions such as pens that click loudly, noisy windbreaker clothing, and cuffs with buttons were advised to be avoided if possible. The researchers recommended using an inexpensive sponge microphone tip to prevent interference on the quality of recordings from puffs of air on plosive and fricative sounds. They frequently listened to the sentences that were recorded to ensure high quality. Ergonomics such as sitting in a chair and looking at a computer screen for long periods of time were discussed, although there were no specific suggestions included to aid management of these factors. The researchers did suggest providing a drink of water for individuals who are voice banking and that the person should be advised they can take a break from recording at any time.

#### **2.4. Present study**

The present study was designed to address the lack of synthetic voices suitable for the New Zealand context and a gap of literature around the voice banking experiences of people who do not have speech therapy backgrounds. ModelTalker voice banking technology was chosen to reflect the voice banking process that TalkLink currently use with clients in New Zealand. It was also selected because the voice banking process was free of charge for healthy speakers

who consented to donating their voices. A mixed methods design was employed to examine the following primary aims, research questions, and hypotheses:

1. To explore the effectiveness of the ModelTalker voice banking protocol for New Zealand speakers:

a. What is the experience of healthy voice donors during the ModelTalker voice banking process?

Hypothesis I: *Voice banking with ModelTalker technology is a positive experience.*

Hypothesis II: *Children and adults can voice bank using ModelTalker technology with ease.*

b. Are there any alterations that are required to make ModelTalker voice banking suitable for New Zealanders?

Hypothesis III: *ModelTalker can be applied to a New Zealand context.*

2. To create and evaluate the New Zealand-accented synthetic voices for speech generating devices:

c. How do unfamiliar listeners perceive the intelligibility and naturalness of the New Zealand-accented voices created using ModelTalker technology?

Hypothesis IV: *The synthesised ModelTalker voices are intelligible to native New Zealand listeners.*

Hypothesis V: *The synthesised ModelTalker voices are natural-sounding to native New Zealand listeners.*

d. Can unfamiliar listeners perceive the age and gender of the New Zealand-accented voices created using ModelTalker technology?

Hypothesis VI: *The synthesised ModelTalker voices portray the age of the voice donors.*

Hypothesis VII: *The synthesised ModelTalker voices portray the gender of the voice donors.*

## Chapter 3. Voice banking methodology

### 3.1. Study design

The present study was an exploratory trial and evaluation of voice donation via ModelTalker technology. A mixed methods design was chosen for the study, where quantitative and qualitative methods and analyses were employed to evaluate different aspects of the study. After completing their recordings, each voice donor was given a questionnaire on their experiences of voice banking using ModelTalker technology. Perceptual data was collected via an experiment recruiting unfamiliar listeners who evaluated aspects of the synthesised voices (outlined in Chapter 5). In this way, the use of a mixed method design allowed for a broader understanding of the experience of voice banking and evaluation of the synthesised voices created using ModelTalker technology.

### 3.2. Ethical considerations

The present study was reviewed and approved by the University of Canterbury Human Ethics Committee. Information sheets and consent forms were developed for both experiments (see Appendices A–C). Several ethical issues that were considered for the current study are discussed below.

#### 3.2.1. Confidentiality of participants

Prospective voice donors were given alphabetical codes systematically in order of recruitment. Successful voice donors were given codes represented by age group and gender. All data with participant codes only were stored digitally on a password protected computer on the University of Canterbury server in a locked office.

##### 3.2.1.1. Vocal identity of voice donors

The nature of the current project requires the recordings that create the synthesised voices to match the vocal identity of the voice donors. The completed synthetic voices are available for use by New Zealand adults and children who communicate with SGDs, and there is a small chance that someone may identify the speaker. Although the exact number of people who might use these voices is unknown, it is expected to be less than 500 at any one time. The synthesised voices included age group and gender information about the voice donor, but no other identifying information were attached to the voices. Identification of the voice donors through the synthesised voice and this information is unlikely but may be possible in exceptional situations of chance. Voice donors were fully informed of the future use of the synthesised voices and the small chance of identification.

### **3.2.1.2. ModelTalker access to recordings**

ModelTalker staff did not receive any identifying information about the participants. The voice donor age and gender codes were used as the usernames for the ModelTalker protocol. Voice donors were fully informed and gave permission for the engineers at ModelTalker to listen and work with the coded recordings to create the synthesised voices. Personal communication with Bunnell (2017) detailed that all data transmission were encrypted via Secure Sockets Layer (SSL) communication channels. All recorded speech data and user information were stored in restricted-access password-protected databases and file systems on ModelTalker servers. The servers were located in the secure climate controlled restricted access data centres either at a commercial data centre or within Nemours Alfred I. duPont Hospital data centres (Bunnell, personal communication, 2017).

### **3.2.2. Maintenance and intellectual property of the synthesised voices**

ModelTalker will maintain the synthesised voices so that New Zealand SGD users are able to access the voices through TalkLink. ModelTalker holds the intellectual property for the software that created the synthesised voices. Personal communication with Bunnell (2017) detailed that a ModelTalker “voice” consists of two major components: the actual speech recordings and derivative data related to the recordings (called the “voice data”), and the text-to-speech software that can create audio output from the voice data. When a voice donor registers to donate their speech for creating a synthetic voice, they digitally sign an agreement giving ModelTalker unrestricted rights to use the recordings to develop synthetic voices. To receive a donated ModelTalker voice, a user must register as someone who needs the voice for assistive communication. When installing the voice, they must agree to the ModelTalker End User License Agreement. This gives the user rights to use the voice for personal, non-commercial purposes.

### **3.2.3. Considerations for child participants**

A child-friendly information sheet and consent form was given and explained to the child participants (see Appendix B). A full information sheet and consent form were given to the child’s parent or caregiver. The researchers understood that the recording process was a large commitment for a child and their parents or caregivers. The setup of the recording sessions allowed great flexibility for children to take breaks and negotiate the recording session’s duration. The primary researcher was a qualified speech-language therapist and the secondary researcher was in the final year before qualifying as a speech-language therapist. Both

researchers had experience working with children and reading children's cues regarding fatigue and ability to remain on task.

#### **3.2.4. Cultural sensitivity**

Māori consultation was sought and approved through the Ngāi Tahu Consultation and Engagement Group. The following areas were addressed through the consultation process.

##### **3.2.4.1. Māori involvement as participants**

Voice donors of Māori ethnicity were involved in this project. Extra consideration was taken regarding the tapu surrounding the head. The use of a headset microphone was required and participants were always asked permission before the headset was worn or adjusted. Where possible, participants were encouraged to place and adjust the headset themselves. Once created, the New Zealand-accented voices remained anonymous and no specific details about the participant were associated with the voice, with the exception of gender and age group.

##### **3.2.4.2. Significant Māori content and data being sought and analysed**

The voices of Māori voice donors were recorded as well as recordings of core te reo Māori vocabulary. The New Zealand synthesised voices were evaluated by unfamiliar listeners to rate intelligibility, naturalness, and predictions of age and gender.

##### **3.2.4.3. Research that will impact on Māori**

The lack of availability of New Zealand-accented voices with te reo Māori pronunciation capabilities has been a notable gap in the provision of support for Māori clients who experience loss of their natural voice. TalkLink have reported this as an urgent area of need. This project has addressed this need by creating voices recorded by Māori voice donors for people who use SGDs. A small core vocabulary of words in te reo Māori with correct pronunciation was also recorded for use on SGDs.

### **3.3. Recruitment**

Participants were recruited from within Christchurch, New Zealand. To promote the current study, posters were displayed on noticeboards around the University of Canterbury and online via social media such as Facebook. The poster created included brief details about the study and contact details of the researchers.

#### **3.3.1. Voice donor requirements**

The voice donor participants were required to be English speakers of New Zealand European or Māori ethnicity. Voice donors were required to live in Christchurch so they could attend the recordings sessions at the sound-treated room used for the study. The advertised age groups for New Zealand European male and female voice donors were as follows: 10 years, 20 years, 40

years, and 60 years, and the New Zealand Māori voice donor age group was advertised as 30–50 years. Participant age groups were chosen after discussion with TalkLink to attempt to reflect the clinical populations who use SGDs. The maximum number of participants was set at ten voice donors: five males and five females, with two children and eight adults across the age groups.

### **3.3.2. Screening session**

Sixteen prospective participants attended a screening session at the New Zealand Institute of Language, Brain, and Behaviour (NZILBB) sound-treated room at the University of Canterbury. This was to familiarise prospective participants with the recording equipment and environment and to create sample recordings of their voice. The session was recorded using a handheld recorder and included reading the information sheet, signing the consent form, and various speech and language tasks as follows and as detailed in Appendix D. The duration of the screening session was approximately 15 minutes.

#### **3.3.2.1. Conversational speech sample**

Prospective participants were asked open-ended questions about their week. The researchers used follow-up questions related to the information given by the participants to continue conversation. The conversational speech samples were approximately 3–5 minutes in duration.

#### **3.3.2.2. Paragraph reading**

Prospective participants were given a printed copy of *My Grandfather* passage and were instructed to read it out loud in their everyday speaking voice (Darley, Aronson, & Brown, 1975). *My Grandfather* passage was chosen over the shortened version of the *Rainbow Passage* (Fairbanks, 1960). From the Powell (2006) comparison of English reading passages for elicitation of speech samples, *My Grandfather* passage has an increased syllable ( $n = 172$ ), word ( $n = 131$ ) and sentence count ( $n = 10$ ) and a lower Flesch-Kincaid grade level ( $n = 4.98$ ) compared to the shortened *Rainbow Passage* ( $n = 132$ ; 101; 7; 5.50 respectively). Considering prospective child participants, *My Grandfather* passage was selected for the relative ease of vocabulary and increased phonemic opportunities seen as an important factor for screening age-appropriate speech sound inventories.

#### **3.3.2.3. Sentence reading**

Twelve sample sentences were selected from the ModelTalker software. These were the ten “trial” recording sentences that voice donors would encounter when doing the initial recording session and the first two sentences of the full sentence inventory. The first sentence of the full recording was chosen as it had been identified as a potentially challenging first sentence to start

off the recordings, especially for children. The 12 sentences were printed in 24-point font with four sentences per landscape A4 page, and were presented to the participants. The prospective participants were informed that these were samples of sentences that they would encounter in the full recording process and were instructed to read them aloud in their everyday speaking voice.

### **3.3.3. Screening process**

The 16 prospective participant recordings were coded and given to a speech-language therapist of the University of Canterbury Communication Disorders department who was not involved in the project. The independent speech-language therapist was instructed to comment on each prospective participant's suitability in terms of speech, language, voice, and accent being within normal limits for the participant's age and gender.

### **3.3.4. Selection of voice donor participants**

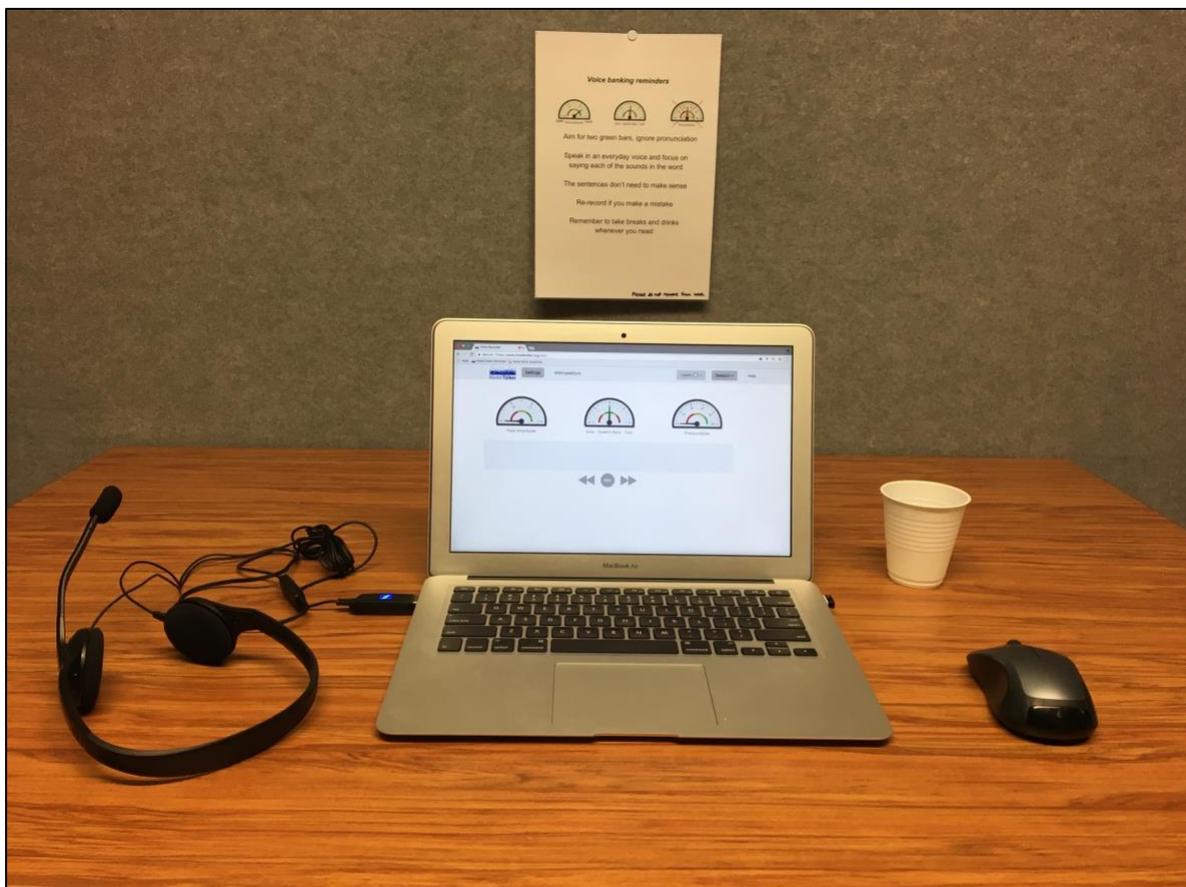
Upon receiving comments regarding suitability of prospective participants from the independent speech-language therapist, ten participants were selected to fill the ten voice donor groups. Two alterations from the initial proposal and recruitment advertising were made. Early in the recording stages a female child participant was unable to continue due to time commitments and travel duration, and so a second female child was sought and recruited. No New Zealand middle-aged males responded to the advertising for participants within the recruitment time frame for the project: however, an Australian male was interested. After consideration by the primary researcher and supervisor it was decided to include this participant as it is likely that there are people from Australia living in New Zealand who use SGDs. All participants who completed the screening session were contacted via their preference of email or text message. Successful participants were invited to continue the voice banking process and given detailed information about the next session. Unsuccessful participants ( $n = 4$ ) were given a short summary of the independent speech-language therapist's comments regarding the presence of vocal fry in their voice and one prospective child was given a short summary of the presence of hypo-nasality in their speech. Once a participant was identified as being a suitable voice donor and confirmed their continuation, prospective participants in the same gender and age group were no longer recruited for the screening session, with exception of the female child re-recruitment as this occurred early in the recruitment phase. Ten participants were selected as voice donor participants for the following categories: male child, female child, young male, young female, middle-aged male, middle-aged female, older male, older female, Māori male, and Māori female.

### 3.4. ModelTalker voice banking process

Recommendations from Creer et al. (2009), Hyppa-Martin et al. (2017), Jackson et al. (2017), and ModelTalker (2018) were followed for the ModelTalker voice banking process.

#### 3.4.1. Materials and set up

The ModelTalker Voice Recorder (MTVR) was displayed through a MacBook Air 13" laptop using ModelTalker's preferred web browser, Google Chrome. Audio input and output was delivered through ModelTalker's preferred USB head-mounted microphone, a Sennheiser PC 36, and a foam tip was used to reduce audio distortion from the airflow of the mouth and nose. All recording sessions were conducted in a sound-treated room with wireless internet access and a table and chairs at the NZILBB facility at the University of Canterbury, as shown in Figure 1.



*Figure 1.* Set up of the recording session.

The primary researcher created a guest account on her personal laptop to ensure that no other applications (e.g., e-mail notifications) would be open or distract recording sessions. The primary researcher created a voice donation login for each voice donor participant using the researchers email address as the contact. The laptop was always unplugged from the charger

for the duration of recording sessions. The researcher fully quit the Google Chrome web browser and plugged in the Sennheiser headset. On the laptop, the researcher opened the *Input* tab of the *Sound Options* in *System Preferences* and ensured that the *Use ambient noise reduction* box was not checked. The *Input* and *Output* tabs were both checked to ensure that the Sennheiser headset was the selected option for microphone and earphones respectively. Google Chrome was launched. The MTVR webpage was opened and the login details were entered. The researcher directly selected the Sennheiser headset from the top right microphone options on the Google Chrome browser. The researcher followed the instructions on the webpage to complete the silence measurement. The silence readings were recommended to be between -80 and -70 dB before beginning recordings. The researcher noted the silence levels and instructed the voice donor to place the headset on their head. The researcher checked the position of the microphone to ensure it was positioned at the corner of the voice donor's mouth. The voice donors were instructed to sit comfortably upright in the chair facing the laptop.

#### **3.4.2. Initial recording session**

The researcher read over the recording procedure reminders, which included information about how to say the sentences and what to aim for with the amplitude and speed dials. A summary poster was displayed on the wall of the sound booth for participants to refer to during the initial session and for following sessions (see Appendix E). The initial session involved recording a short inventory of ten sentences for ModelTalker engineers to review. The participant was prompted to listen to the model voice before reading out the sentence themselves. Each of the ten screening sentence stimuli were displayed one at a time on the computer screen. After recording each sentence, three meters on the top of the screen provided feedback regarding peak amplitude, speech rate, and pronunciation. The participant was instructed to aim to have the peak amplitude and speech rate within the green zone, but to ignore the pronunciation meter as it had been designed for American English and did not apply well to New Zealand English accents. On completion of the ten screening sentence recordings, the researcher listened back to the recordings and instructed the voice donor to re-record any if required (e.g., if outside construction noise was picked up in the recordings or there were audible puffs of air from plosive consonants). The ten screening sentences were then submitted to ModelTalker for review.

#### **3.4.3. Recording sessions**

After receiving confirmation of the quality of the screening sentences, the voice donors returned to begin the recording sessions. For adult voice donors, each recording session was

two hours in duration; the duration of sessions for the child voice donors was 1.5 hours. The voice donors were reminded of the recording procedures from the poster displayed on the wall. The voice donors then began the ModelTalker process as outlined above, recording the 1,600 sentences over multiple recording sessions. Once the adult voice donors were comfortable using the MTVR the researchers left the sound-treated room to allow the voice donors to record independently. The researchers stayed with the children during the recording sessions. The researchers checked in on all voice donors every half hour to remind them to pause and take a drink of water before continuing. Participants were also reminded that they could take a vocal break any time as the MTVR auto-saved each recorded sentence. Voice donors were encouraged to listen periodically to the preview voice that the MTVR created as a sample of their gestating voice. At the end of each recording session the participants' recording progress was auto-saved, allowing the following session to start off from the next sentence.

### **3.5. Te reo Māori custom inventory**

Following discussion with TalkLink regarding the current unavailability of te reo Māori for SGDs, two inventories of te reo Māori vocabulary were produced. Personal communication with Bunnell (2017) guided the process for recording the te reo Māori vocabulary using the custom inventory phrase banking feature on the MTVR. Each te reo Māori word was embedded into three different English phrases to record each word in initial, medial, and final position in the phrase. Some sentences created included multiple te reo Māori words to reduce the number of phrases used. Following recommendations from Bunnell (2017), the word prior to the te reo Māori word ended in a voiceless fricative or plosive and similarly the word immediately following the te reo Māori word began with a voiceless fricative or plosive. This was to make it easier for the automatic segmentation algorithms to cleanly discern the start and end of the te reo Māori words. Macrons were unable to be included as the diacritics presented some challenges for the text processing parts of ModelTalker. Two custom inventories of te reo Māori were created and recorded by the voice donor participants.

#### **3.5.1. Twenty common te reo Māori words**

Under the guidance of Professor Jeanette King, leader of the bilingualism theme at the NZILBB, twenty te reo Māori words that New Zealanders would commonly encounter and use were selected (see Appendix F). All ten voice donors recorded these sentences. The model pronunciations were muted while recording the te reo Māori custom inventories as the text-to-sound rules were unable to pronounce the words accurately. The primary researcher offered a verbal model of the sentence as requested by the participants.

### **3.5.2. Extended te reo Māori custom inventory**

A core board of 65 vocabulary words used with low- and high-tech pictorial communication boards was supplied by TalkLink with English and te reo Māori text. The translations had been carried out by a TalkLink therapist in the North Island, so the te reo Māori words were checked for dialect differences and altered as appropriate with support from the Māori male voice donor whose first language was te reo Māori. The core vocabulary words and the part-of-speech information were sent to ModelTalker to provide a dictionary of the te reo Māori words and to allow each word to be correctly tagged (see Appendix G).

### **3.6. Recording the Speech Intelligibility Test sentences**

The Speech Intelligibility Test (SIT) sentences were generated using a CD copy of the SIT stimuli generator. Trials were administered using the CD to generate ten sets of eleven sentences ranging from five to fifteen words in length. Each of the sentence sets were allocated to one of the voice donors. At the end of the final recording session, the participants were asked to read aloud one of the sentence sets; this was recorded using a hand-held voice recorder. This collected a sample recording of the SIT in the participant's natural voice that could be later compared to the synthesised voice version of the same SIT sentences.

### **3.7. Qualitative methods**

#### **3.7.1. Participants**

A final sample of ten (five female, five male) voice donors who completed the voice banking process were included in the analyses. Table 3 shows the basic biographical details of the participants, including age, gender, and ethnicity. The participants ranged from 9–65 years at the time of completing the screening session.

#### **3.7.2. Materials**

An online questionnaire hosted by the University of Canterbury's Qualtrics survey software was used to present the questionnaire. The questionnaire involved adapted Likert-type scales selected from relevant literature and a small number of open-ended questions.

#### **3.7.3. Procedure**

All voice donors were presented with the questionnaire on the completion of the last recording session and were required to complete this before receiving a \$100 voucher for their participation in the project.

Table 3.

*Biographical details of voice donor participants.*

<b>Voice donor</b>	<b>Age (yy;mm)</b>	<b>Ethnicity</b>
Male child	09;01	NZ European
Female child	13;02	NZ European
Young male	22;05	NZ Māori, NZ European
Young female	21;08	NZ European, Chinese
Middle-aged male	41;10	Australian European
Middle-aged female	48;08	NZ European
Older male	65;02	NZ Māori, NZ European
Older female	58;10	NZ European
Māori male	22;04	NZ Māori
Māori female	38;04	NZ Māori

#### **3.7.4. Measures**

Twenty-one questions were created and placed into the following categories: pre-voice banking knowledge, experiences with voice banking, complexity and time measures, features of MTRV, and recommendations for other people and other environments (see Appendix H).

##### **3.7.4.1. Pre-voice banking knowledge**

“*What did you know about voice banking before participating in this study?*” was asked as an open-ended question to investigate existing knowledge or prior experiences around voice banking services. Voice banking for clinical populations such as people with MND is optimally completed early in diagnosis in order to reduce dysarthria present in the recordings to maximise quality of the synthesised voice (Bunnell et al., 2017). It was of interest what knowledge about voice banking the participants had prior to the study.

##### **3.7.4.2. Experiences with voice banking**

A combination of open-ended questions and adapted Likert-type scales were used to gather participants overall experiences with voice banking. The open-ended question “*Tell us about your experience with the voice banking process*” was chosen deliberately in order to avoid cues that might have encouraged socially desirable response (Schwarz & Oyserman, 2001). A ten-point Likert-type scale with *1=negative* and *10=positive* asked participants to “*Rate your overall experience with the voice banking process.*” Ten-point Likert-type scales were used as

a standard scale throughout the questionnaire with ten symmetrical points between the left most perspective and the right most perspective. Participants were asked open-ended questions such as “*What did you like best about voice banking and why?*” “*What did you like least about voice banking and why?*” and “*How could voice banking be improved?*” and the questions were aimed to not prime participants with any cues for answers.

#### **3.7.4.3. Features of the MTRV**

Seven questions were asked about complexity and time measures using a ten-point Likert-scale and an open-ended “*Please explain your rating.*” Information was gained regarding the ease of the process, the length of time taken to complete the process, and the usefulness of the volume and speed dials. Participants were also asked about the likeability of hearing the preview synthesised voice and how similar the voice sounded to their own voice at the start and end of the process.

#### **3.7.4.4. New Zealand adaptations**

Three questions were asked about aspects of the voice banking process that related to adaptations for New Zealanders. These included Likert-type scales and open-ended explanations of the helpfulness of the model voice, the types of sentences that were recorded, and the process of recording the te reo Māori words.

#### **3.7.4.5. Recommendations for other people and other environments**

Five questions were related to the participants’ recommendations for other people and other environments. Two questions used a 5-point Likert scale with *1=definitely yes*, *2=probably yes*, *3=unsure*, *4=probably not*, and *5=definitely not*, and asked participants “*Would you recommend that other people donate their voice?*” and “*Would you recommend voice banking to someone in the process of losing their ability to talk?*” The purpose of these questions was to see any differences between the recommendations for healthy voice donors compared to people in the process of losing their ability to speak who are going to require use of their synthesised voice as their communication method in the future. Participants were also asked “*What do you see as any barriers that could prevent people from voice banking?*” The final two Likert-scale questions were based around perceived success of voice banking in the home environment and completing voice banking independently. Open-ended explanation questions were used to allow participants to explain their ratings. The questionnaire finished with “*Is there anything else you would like to share?*” to allow participants to record any information that may not have been covered in the previous questions, but that they thought was important for the project.

### **3.8. Data analysis**

The study used multiple data analysis methods due to the range of data collected. Voice donor recording rates and the adapted Likert-scale questions were analysed using descriptive statistics and basic measures of range and central tendency. The qualitative data were analysed using thematic analysis. This was to explore participants' experiences and perspectives. A thematic analysis method as outlined in Braun and Clarke (2006) was used to allow patterns to be identified and grouped to reflect participants' perspectives. Braun and Clarke (2006) describe six phases of thematic analysis that were used as a guide for analysis of the data for the present study. These are: becoming familiar with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. All data collected from the questionnaire were exported from the Qualtrics software programme into a Microsoft Excel sheet. Initial codes were written next to participants' responses and then organised into categories which formed the foundations of generating themes. After identifying and naming the themes, the participants' responses were reviewed once more to ensure that all information was captured by the generated themes.

## Chapter 4. Voice banking results

All ten voice donors completed the voice banking process and the online questionnaire. In this chapter, voice donor recording rates and the data from the online questionnaire are presented. This chapter relates to the effectiveness of the ModelTalker protocol for New Zealand speakers and the two subsequent research questions:

- a. *What is the experience of healthy voice donors during the ModelTalker voice banking process?*
- b. *Are there any alterations that are required to make ModelTalker voice banking suitable for New Zealanders?*

### 4.1. Voice donor rates of recording

The amount of time for each participant to complete the recordings was totalled and recording rates were calculated. Table 4 shows each voice banking participant, the sum of sentences recorded, time taken, and sentences per hour.

Table 4.

*Voice donor rates of recording.*

<b>Voice Donor</b>	<b>Sum of sentences</b>	<b>Sum of time (hours)</b>	<b>Sentences per hour</b>
Male child	1640	11.00	149.09
Female child	1640	5.50	298.18
Young male	1640	4.17	393.60
Young female	1640	3.17	517.89
Middle-aged male	1640	5.25	312.38
Middle-aged female	1640	6.17	265.95
Older male	1640	7.00	234.29
Older female	1640	5.50	298.18
Māori male	1835	5.67	323.82
Māori female	1835	4.00	458.75

*Note.* The ModelTalker protocol is made up of 1,600 sentences. All ten participants recorded 40 custom inventory sentences for the common te reo Māori words. Additionally, the two Māori speakers recorded an extended inventory of 65 te reo Māori core vocabulary words.

The median recording duration was 5 hours and 30 minutes. The minimum recording time was 3 hours and 10 minutes (young female) and the maximum was 11 hours (male child). The median recording time was chosen as the measure so that the maximum value did not disproportionately influence the average calculations. With exception of the youngest participant taking the most time, there did not appear to be any significant relationships between time taken to record the sentences and age or gender. Detailed calculations were not performed because each voice donor was a case study for each age and gender group.

#### 4.2. Results from online questionnaire

This remaining of the chapter will report on the findings obtained from the online questionnaire completed by all ten voice donor participants. The chapter is organised into sections from the questionnaire as outlined in the methodology. Each section will explore voice donor experiences and perspectives using Likert-scale descriptive statistics, thematic analysis summaries, and participant quotes. Table 5 shows the list of final main themes and associated subthemes which will be discussed.

Table 5.

*List of final main themes and associated sub-themes.*

<b>Theme</b>	<b>Sub-themes</b>
Gained awareness	Increased knowledge about voice banking Appreciated and recommended opportunity to help others Reflected how patients could benefit by voice banking
Positive features of ModelTalker	Ease of the process Usefulness of the speed and volume dials Progression of the synthesised voice previews
Challenges of ModelTalker	Time constraints Number of sentences recorded Energy and fatigue levels Ergonomics
Kiwi adaptations	Mute the American model voice Include more te reo Māori and New Zealand content

#### **4.2.1. Pre-voice banking knowledge**

Most of the participants indicated no knowledge about voice banking prior to the study:

*“I had no experience with voice banking and knew nothing about it. So it was great to hear knowledge about what voice banking is.”*

*(Young female)*

Two participants described some knowledge: one from university communication disorders classes and one from attending a presentation on the topic given by the primary researcher earlier in the year. One participant had detailed knowledge about voice banking as they were employed in work and research in a similar field. Participants gained awareness about voice banking throughout the study and communicated this through the questionnaire:

*“This has been a fascinating experience and something I didn’t know existed. I didn’t know what I was signing up for but it has truly inspired me.”*

*(Māori female)*

#### **4.2.2. Experiences with voice banking**

Figure 2 shows the majority of participants had positive experiences. This matched with the participants’ comments regarding the voice banking experience, where participants used positive descriptors in their responses:

*“I very much enjoyed taking part in the voice recording, it’s such a positive thing to do. Thank you for letting me be part of a hugely exciting project. I feel extremely lucky to have taken part.”*

*(Middle-aged female)*

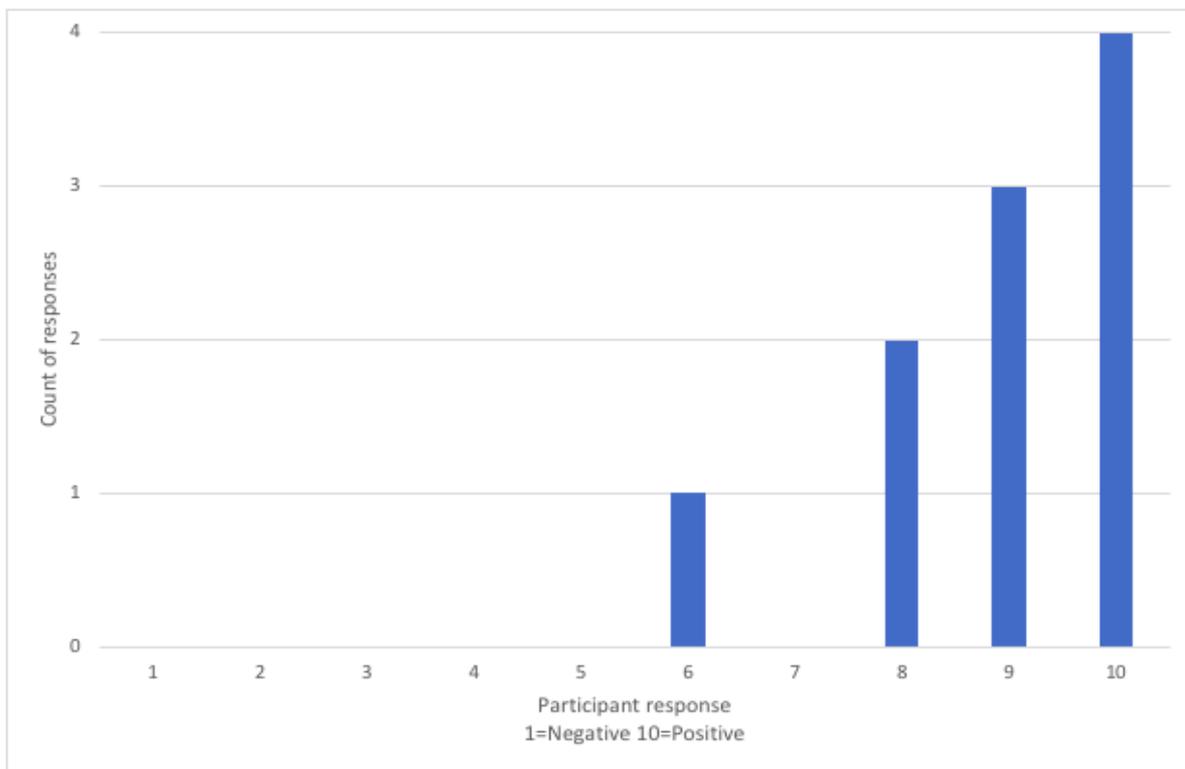


Figure 2. Voice banking participant experiences with voice banking.

Six of the participants answered the open-ended question of “*What did you like best about the voice banking process and why?*” with comments related to the opportunity the project gave them to help others and one participant included a reflection of the importance of her own voice to her:

*“I loved the idea that my voice was going to possibly help people using AAC devices.”*  
*(Young female)*

*“The best part was the fact it could benefit other kiwis in the long term.”*  
*(Māori female)*

*“I was enjoying this one time opportunity and thinking of the people I would help.”*  
*(Female child)*

*“For me it was the chance that I may be able to help someone. My voice is very important to me and I couldn’t imagine life without it.”*

*(Middle-aged female)*

Three of the participants indicated that listening to their own voice being synthesised was their favourite part:

*“The best part was listening back to my synthesised voice.”*

*(Young male)*

There were a range of themes for the participants’ least liked aspects of voice banking. The main responses were focused on the repetitiveness of the process of recording, the duration of time required, and the number of sentences recorded:

*“The only part I kind of disliked was the amount of sentences but it wasn’t too bad.”*

*(Female child)*

*“It is hard concentrating on something so mind-numbing.”*

*(Young female)*

*“Not that it can be helped or avoided but the duration of the recording sessions that’s required is quite long.”*

*(Māori male)*

### **4.2.3. Features of the ModelTalker Voice Recorder**

This section reports the findings from the questions related to the features of the ModelTalker Voice Recorder (MTVR). There were a number of common positive features discussed as well as challenges that voice banking participants faced.

#### **4.2.3.1. Ease of the process**

The median and mode ratings of the ease of voice banking was 9 and 10 respectively on the scale of *1=difficult* and *10=easy*. Many of the participants referred to ModelTalker’s simplicity and functionality:

*“The software design was easy to understand. It was clear and simple on the computer screen with the green bars. The functionality of the process was easy and straightforward. It was easy to go ahead and record, and re-record if you made a mistake.”*

*(Young female)*

*“You could follow the model if unsure of words. You could monitor your pace and amplitude yourself. You could re-record easily. The technology was easy to use. Instructions were clearly given.”*

*(Older female)*

The participant who had the lowest rating of the ease of voice banking commented the following:

*“Speaking wasn’t the difficult part really, but being clear with speech and keeping a consistent tone is not easy.”*

*(Māori male)*

#### **4.2.3.2. Usefulness of the speed and volume dials**

Eight of the participants rated the volume and speed dials as 9 or 10 on the Likert scale with *1=useless* and *10=useful*. Participants used the speed and volume dials as feedback about the recordings that they were doing and whether they needed to re-record:

*“They kept me consistent I think, and allowed me to make sure I was on the right track”*

*(Māori male)*

*“They helped me slow down when needed or raise my voice. They were also motivational.”*

*(Māori female)*

*“I was able to watch them as I spoke a sentences and re-do if I felt necessary and I did that mostly because of the dials”*

*(Female child)*

*“After a little practice I was able to keep within the green range.”*

*(Older male)*

A couple of participants commented about some difficulties with the speed dial:

*“[The dials were] useful but you could get hung up on them but they were also motivational”*

*(Older female)*

*“I paid more attention to the speed dial. The volume dial was typically in an acceptable region, but the speed one was tricky: I felt it was inconsistent”*

*(Middle-aged male)*

#### **4.2.3.3. Progression of the synthesised voice previews**

Nine of the participants rated the feature of being able to listen to the preview of their synthesised voice as 10 on the Likert scale with 1=*didn't like to hear the preview voice* and 10=*liked to hear the preview voice*. Participants provided many responses that indicated their likability of hearing the synthesised voice:

*“It has been great watching the Model Talker learn my voice and then play it back to me. At first it was a little funny, but then it sounded like me.”*

*(Māori female)*

*“The preview at the start of the recordings was quite broken up sounding. The end product was far smoother and very like my voice”*

*(Middle-aged female)*

The youngest participant was the only one who did not rate the preview voice option as a 10. He rated this a 6:

*“[The thing I liked the least was] the first time I heard my voice. I hated it at the start but now I'm sort of getting used to it.”*

*(Male child)*

Other participants commented on the experience of hearing their own voice:

*“Always strange hearing your own voice”*

*(Middle-aged female)*

*“I found listening back to my voice very fascinating, being able to type anything in and have my voice repeat it back to me.”*

*(Young male)*

Two of the participants reflected on how the experience gave them insights about how their own voice is perceived by others:

*“I like hearing my voice synthesised. I’m not sure why, but I guess this allowed me to hear what others might hear when I speak to them.”*

*(Māori male)*

*“I liked to hear [my synthesised voice] to give me an idea of how I sound from a digital device.”*

*(Female child)*

Paired-sample *t*-tests were conducted to examine the change in participants’ perceptions of the preview synthesised voice at the start compared to the end of recording. The means, standard deviations, *t*-values, and *p*-values are shown in Table 6.

Table 6.

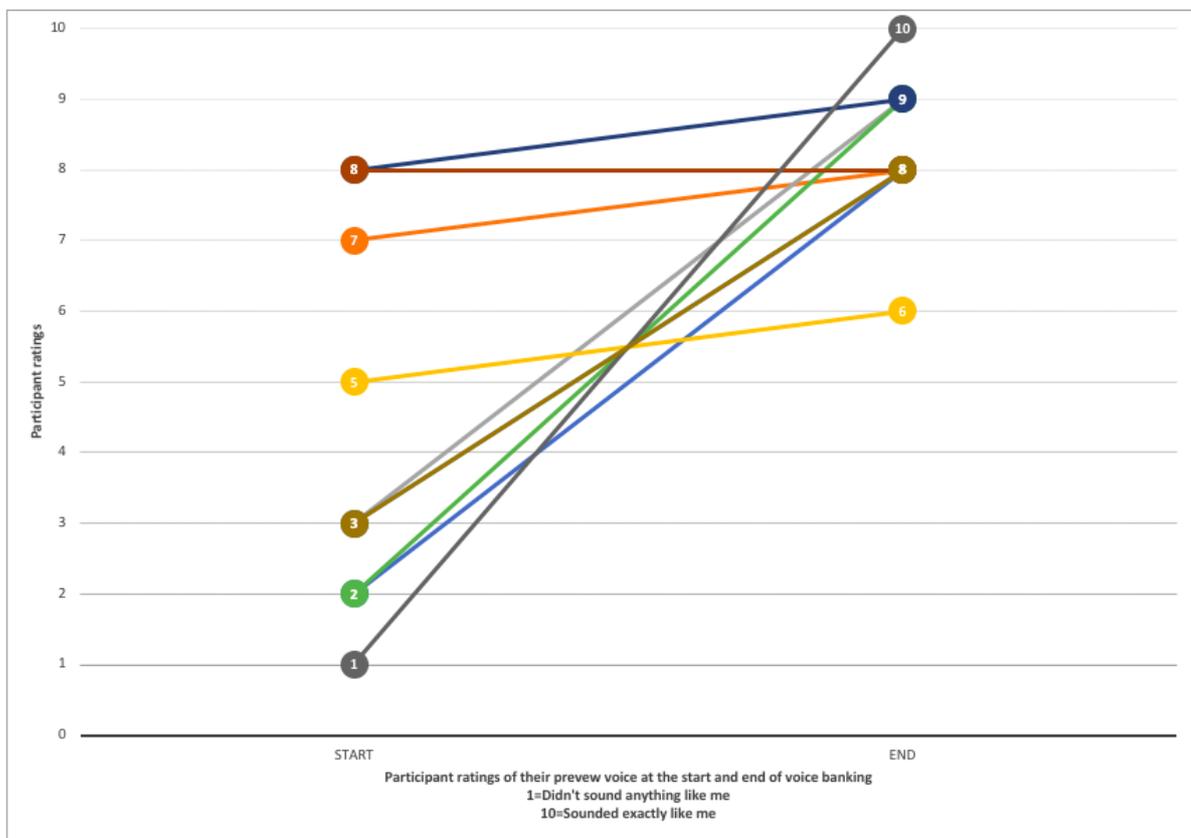
*Paired-sample t-test statistics for voice banking participant ratings of their preview synthesised voice.*

	Rating at start		Rating at end		<i>t</i> -value	<i>p</i> -value
	Mean	Standard deviation	Mean	Standard deviation		
Rating of synthesised voice	3.9	2.7	8.4	1.1	6.4	0.00010

The results show there is a significant difference at  $p \leq 0.05$  that participants rated the preview voices as sounding more like themselves at the end compared to the start preview voice. All participants reported that the preview synthesised voice improved over time and began to sound more like themselves. Figure 3 shows the positive trend for each participants ratings of their preview voice at the start and the end of voice banking. One participant who rated his voice as an 8 at both the start and end could hear his voice throughout, but commented on the characteristics of the voice across time:

*“It always sounded like me, it just sounded either clumsy and robotic or smooth and natural.”*

*(Middle-aged male)*



*Figure 3.* Voice banking participant ratings of preview synthesised voices at the start and end of voice banking.

*Note.* Each coloured line represents one voice banking participant.

#### 4.2.3.4. Voice banking time constraints

Figure 4 shows the range of responses for the question “Please rate the overall time taken to complete the voice banking process.” Participants who thought it took little time reported:

*“It was a lot less time than expected.”*

*(Young male)*

*“It took a lot less time than I initially thought. But overall it is a long time to be starting at a screen.”*

*(Young female)*

Three participants commented on the length of time being unavoidable such as:

*“It was understandable that it was long.”*

*(Older female)*

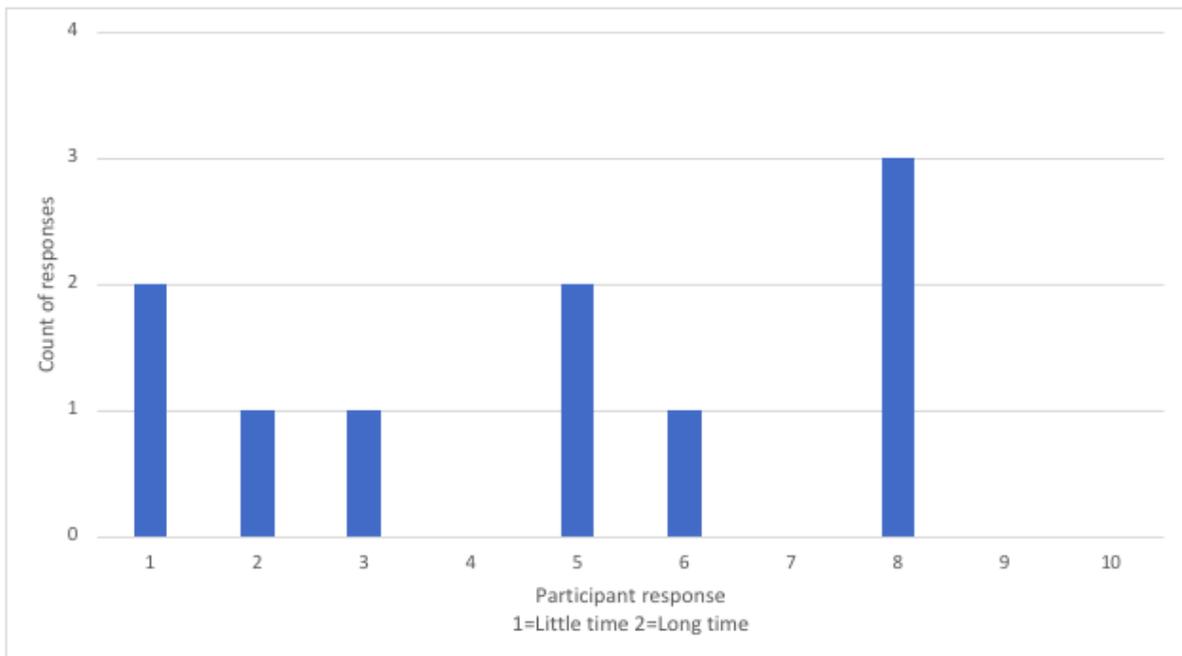


Figure 4. Voice banking participant ratings of the length of time required to voice bank.

Participants employed a number of intrinsic motivation strategies to manage the length of time:

*“Luckily I managed to find a technique that let me do 80-100 sentences every 10mins, so it was like a little goal to hit 600 each hour, depending on the length of each sentence”*

*(Māori male)*

*“I found that once I got on a roll the recoding was quick and easy to do.”*

*(Middle-aged female)*

The youngest participant rated the time as an 8 and commented:

*“It took me 8 sessions. It got in the way of me having playdates.”*

*(Male child)*

The primary researcher set up small goals and competitions related to seeing how many sentences the youngest participant could record in a set period of time, to assist the participant’s motivation.

#### **4.2.3.5. Number of sentences recorded**

The median and mode ratings of complexity of the sentences recorded was 8 on the scale of *1=complex* and *10=simple*. Many participants thought that the sentences spoken were simple to record however had unexpected combinations of words or sounds:

*“Leaning more towards simple, but there were some very weird sentences that were tongue twisters at times.”*

*(Māori male)*

*“The sentences were quite unusual in context but easy to read/say.”*

*(Middle-aged female)*

Participants thought that the process of recording the sentences was repetitive and required energy and endurance:

*“It requires a lot more energy than I thought, and certainly is an endurance test, but worth it in the end.”*

*(Māori male)*

The number of sentences recorded was a common response for the question asking about least favourite aspects of the voice banking process:

*“It would be better if there were less sentences and shorter sentences.”*

*(Male child)*

*“Maybe if there was the option to read paragraphs rather than sentences individually, it would help with overall duration. I’m not sure if the software would do that though.”*

*(Māori male)*

#### **4.2.4. New Zealand adaptations**

Participants’ responses included adaptations that the primary researcher had implemented in the methodology such as the inclusion of te reo Māori recordings and also those which the participants independently adapted. These included muting the model voice, requests for more te reo Māori and New Zealand content in the form of displaying macrons, and having New Zealand-culturally relevant sentences to record.

##### **4.2.4.1. Muting the American model voice**

The majority of the participants thought that the American accent of the model voice influenced their own pronunciation when recording sentences and chose to mute the model voice to record without an auditory prompt:

*“I found I was mimicking the accent, so had to shut it off.”*

*(Middle-aged male)*

*“If I listened to the voice then I often pronounced words in an American voice rather than kiwi.”*

*(Māori female)*

There were two aspects of hearing the model voice that participants identified as being helpful, using the voice as a guide for pronunciation of unfamiliar words and speech rate:

*“If there were words I was unsure of, it was good to hear a model. However I found the American accent influenced my accent, and so decided not to listen to the model voice for the majority of the sentences”*

*(Young female)*

*“Didn’t find it useful except for the one time I didn’t know how to pronounce a word. It slows the process down a lot if you listen to each one before reading it. I muted it after the first hour”*

*(Māori male)*

*“The only useful thing was to figure out which speed to speak at, apart from that it subliminally makes you speak in an American accent”*

*(Young male)*

The youngest participant thought that voice banking could be improved by:

*“Not putting the American voice in it. Because Michelle or myself could have done it.”*

*(Male child)*

The youngest participant opted to mute the model voice from his second recording session onwards; however, he required support from the primary researcher with pronunciation of unfamiliar words. He also required guidance to ensure he was speaking the sentence with the correct word order and with appropriate volume and rate. The primary researcher ended up modelling the majority of the sentences by reading them aloud for the participant before he recorded the sentence himself.

#### **4.2.4.2. Include more te reo Māori and New Zealand content**

Eight of the participants rated the complexity of the te reo Māori sentences recorded as an 8, 9 or 10 on the scale of 1=complex and 10=simple:

*“I am familiar with the Māori language so the words used were simple.”*

*(Young female)*

*“Commonly used and heard in the media so if you have not learned Māori you have probably used them yourself and have an idea about how they should be said.”*

*(Older female)*

The two te reo Māori speakers who recorded the extended inventory of te reo Māori words said:

*“Very simple, but the combinations of sounds were surprisingly difficult at times. Changing between languages can be hard sometimes too because of the shape that the mouth requires for certain sounds.”*

*(Māori male)*

*“I would have liked to have the macrons present, or at least an A4 piece of paper with them on so I could make sure I pronounced the words correctly. They’re quite important and make a big difference.”*

*(Māori female)*

Participants also wanted to have increased relevance to the New Zealand culture in relation to the types of sentences they were recording for the 1,600 English sentences:

*“I think the sentences that I read had a distinctly American tone – not that that’s a bad thing – I would have just liked to have a more New Zealand theme to it for example instead of wolves, dogs and rabbits – kiwis, wild boar, native birds and Hector’s dolphins.”*

*(Middle-aged female)*

*“It would be great if there could be more of a te reo Māori focus for our fluent speakers who may need this in the future.”*

*(Māori female)*

#### **4.2.5. Recommendations for other people and other environments**

This section includes recommendations for voice donating, voice banking, recording in a home environment and recording independently.

##### **4.2.5.1. Recommendations for voice donations vs voice banking**

Figure 5 shows the results from participants' recommendations about other people donating their voice compared to people voice banking because they are losing their voice. The majority of participants supported the recommendation for other people to donate their voice:

*“It is such a rewarding thing to do! Such a helpful thing and going to a great cause. And for little effort.”*

*(Young female)*

*“Absolutely, this is such a valuable gift. To give voice to someone is a privilege.”*

*(Middle-aged female)*

*“The voice you communicate with is incredibly important. I can imagine the frustration of listening to several candidates and not being happy with any of them. Giving more choice could make all the difference to someone and how well they feel liked themselves while communicating.”*

*(Middle-aged male)*

Two of the participants mentioned about the time commitments as barriers for people to donate their voice:

*“If I knew they could handle the hours, and would enjoy having their own synthesised voice at the end that could help less able people, I would definitely recommended it.”*

*(Māori male)*

*“If they are prepared to commit the time then the more options the better and who knows when people might need them even for themselves in the future.”*

*(Older female)*

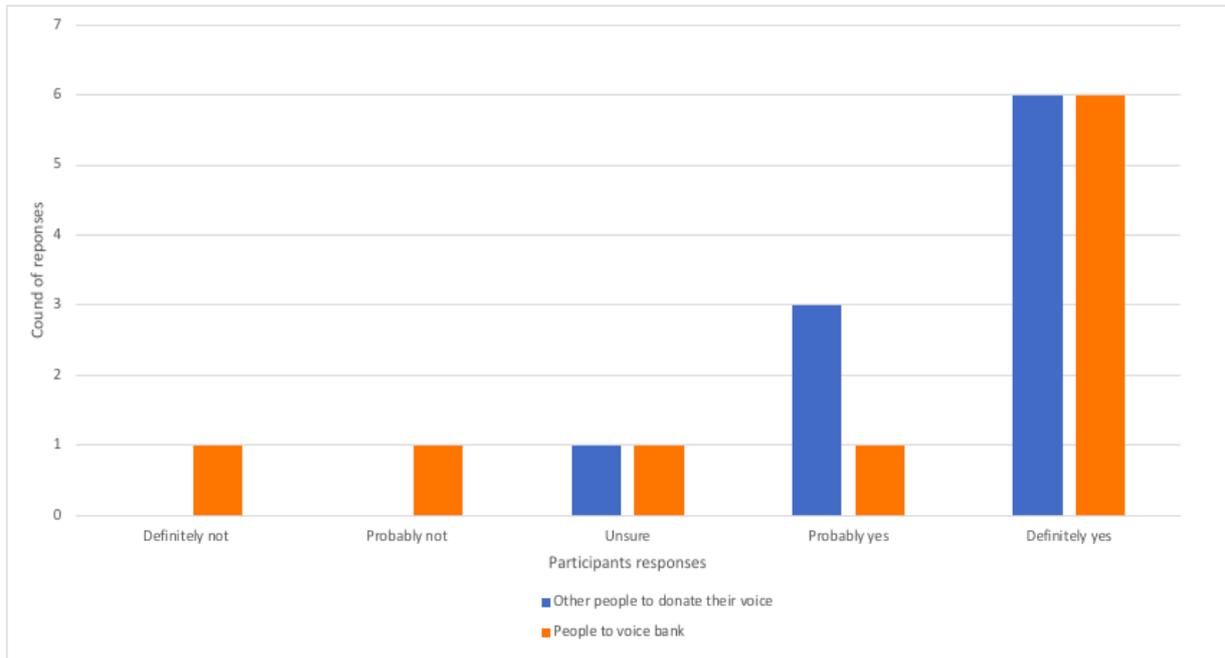


Figure 5. Voice banking participant recommendations about other people donating their voice vs. voice banking for those who are losing their voice.

Figure 5 shows there was a greater spread of responses when participants were asked about their recommendations for people to voice banking when they were in the process of losing their voice. Five participants responded *definitely yes* to both questions. Themes of responses which supported the use of voice banking for people who are losing their voices included keeping the person’s identity, having a natural voice, and being able to use the voice on AAC devices:

*“Absolutely. The opportunity to have your own voice means you keep a part of your identity.”*

*(Middle-aged female)*

*“So they can have it as an option to make the AAC device more relatable to them and listeners.”*

*(Older female)*

*“They would retain their some of their identity and not sound like a robot.”*

*(Māori female)*

The three participants who responded *definitely not*, *probably not* and *unsure* about voice banking had all responded with a degree of yes for the previous question about voice donation. The two reasons for this change in recommendations were fatigue and the idea of preservation of the original voice compared to dysarthric speech:

*“It is tiring on the voice.”*

*(Young female)*

*“It can definitely take a toll on your voice for two hours straight.”*

*(Young male)*

*“I don't know really. It would depend on how far lost their voice was. If it was still 85%+ like their original then I would probably suggest it. However, anything below that and they might not want their voice remembered that way. A memory of how it used to be might be preferred.”*

*(Māori male)*

One of the participants who responded with *definitely yes* also referred to the preservation of the original voice through early voice banking:

*“Yes I would recommend it but at early stages. It would be great for people who can't speak to use their voice once more.”*

*(Female child)*

Other barriers that participants discussed within the questionnaire included limited attention span, poor eyesight, being literate, time constraints, awareness of services, and access to recording areas.

#### **4.2.5.2. Success of voice banking at home**

Figure 6 shows participants' ratings of success if voice banking at home. The main themes reported for participants who thought recording would be successful included finding a quiet space at home and being able to record at their own pace:

*“If you had a quiet room that was being unused, or could ensure no distractions, I think it would work well, as you would feel a lot more comfortable.”*

*(Young female)*

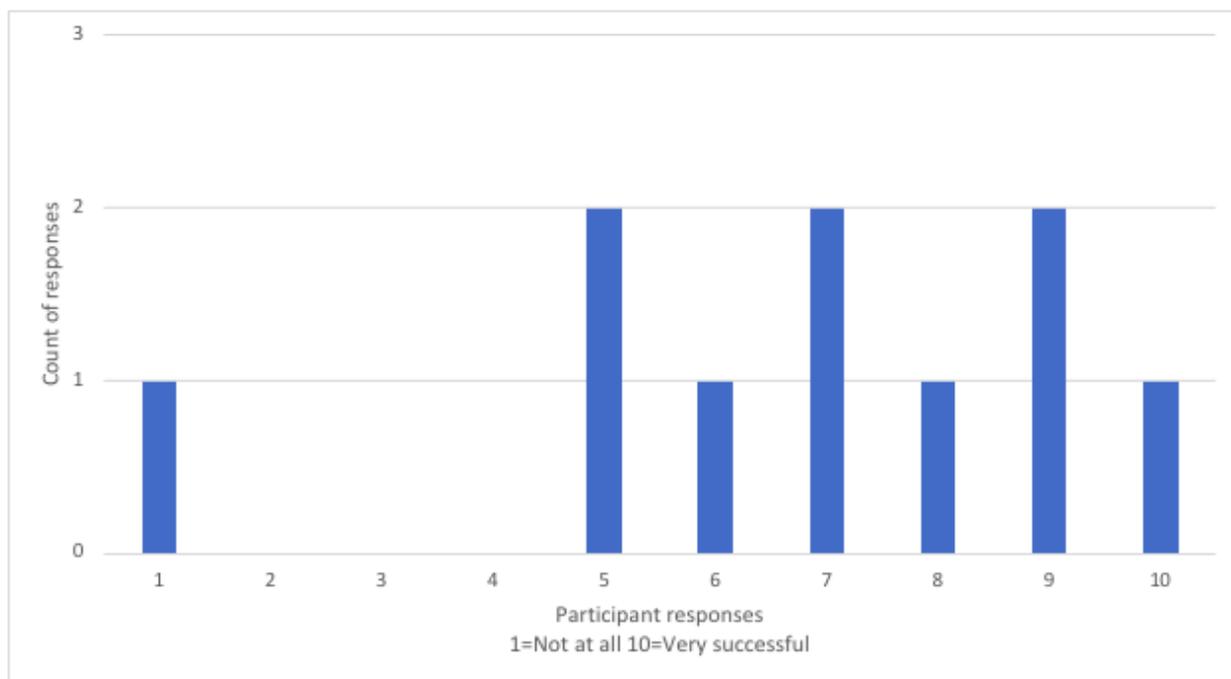


Figure 6. Voice banking participant ratings of success if recording at home.

Another theme across the questionnaire related to ergonomics which could be more flexible at home. Two participants commented on ergonomic aspects of voice banking that they liked the least:

*“The chair as it was uncomfortable. This could be improved by having chair and table heights altered to suit individual.”*

*(Older female)*

*“It’s a long time to be staring at a screen.”*

*(Young female)*

One participant described how he would arrange the recording sessions:

*“I could split the hours up rather than do 2 hours at a time. I might be able to do 30 minutes in the morning, 30 minutes after uni and an hour after dinner for example.”*

*(Māori male)*

Participants who rated a 5 or 6 commented about noise levels and distractions:

*“It would be a bit loud.”*

*(Male child)*

*“Distractions and noise interference would impact.”*

*(Older female)*

The participant who rated a 1 said:

*“It is very loud at my home, it would be a disaster.”*

*(Māori female)*

#### **4.2.5.3. Success of voice banking independently**

Figure 7 shows participants’ ratings of success if recording independently. The majority of the participants thought they would have success with recording on their own. Themes from participants who rated a 9 or 10 included that the process was straightforward however they would like someone to explain and familiarise themselves with the process before starting:

*“Once the process has been explained, I think doing it on your own would be successful because it is such a straightforward process and easy to do.”*

*(Young female)*

*“It was a user friendly system. It’s always good to have someone around to assist though.”*

*(Middle-aged female)*

The participant who worked and researched in a related field indicated he would like further technical information about the process:

*“I think a really technical and in-depth tutorial video that explained the purpose of the different sentence types would be really useful in guiding how we say things. For example, how important is it that we exclaim exclamations, or use rising intonation in questions?”*

*(Middle-aged male)*

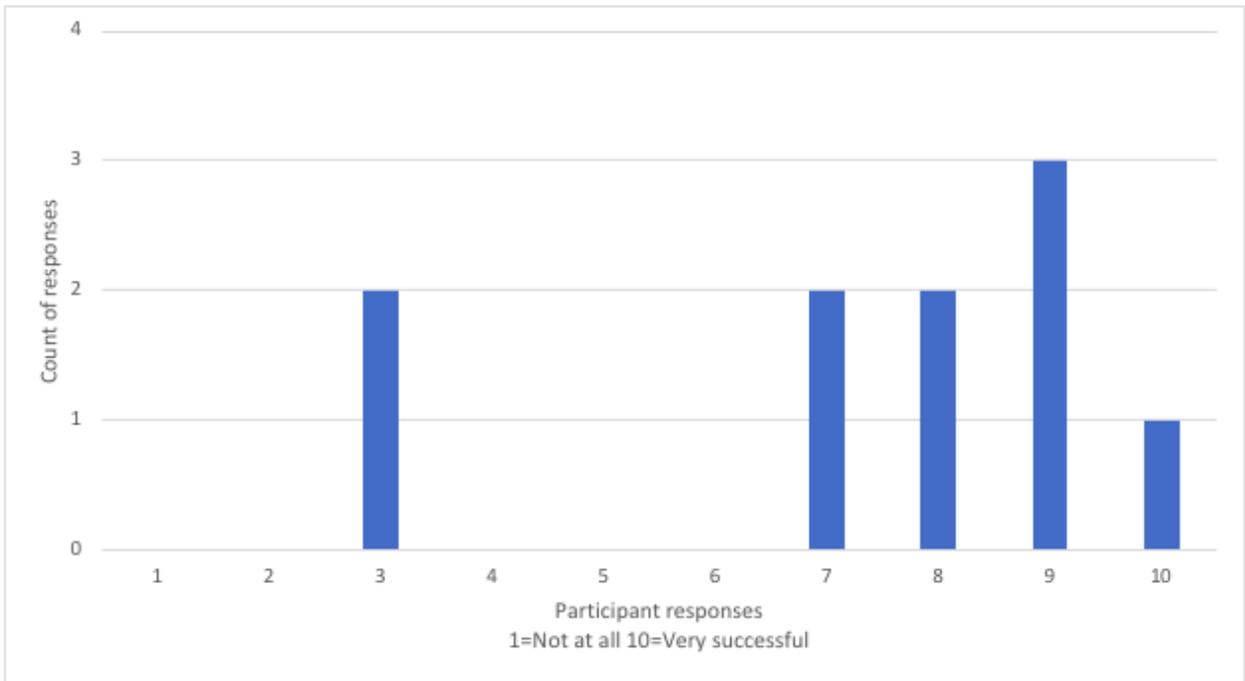


Figure 7. Voice banking participant ratings of success if voice banking independently.

Participants who rated a 7 or 8 indicated motivation and the social aspect of voice banking with others:

*“Depends on motivation, I would be okay but I could see others getting distracted and forgetting to do it.”*

*(Māori female)*

*“It would be okay I guess although there is a social aspect of doing it with other people.”*

*(Female child)*

The two participants who rated a 3 said:

*“I found the explanation and support from the researcher most helpful.”*

*(Older male)*

*“I might struggle on a hard word.”*

*(Male child)*

## Chapter 5. Perceptual experiment methodology

### 5.1. Participants

Fifteen participants were recruited as unfamiliar listeners. Participants lived in Christchurch, were native speakers of New Zealand English with normal hearing and had no significant previous experience with SGDs. Table 7 shows the basic demographical details of the participants, including age, gender, and ethnicity.

Table 7.

*Demographical information about unfamiliar listener participants.*

<b>Participant demographics</b>	<b>Count</b>
Ethnicity	
New Zealand European	13
Chinese	1
Middle Eastern	1
Age (years)	
20–24	12
25–29	2
30–34	0
35–39	1
Gender	
Male	6
Female	9

### 5.2. Materials

An existing NZILBB MATLAB program was utilised to randomly present the stimuli for the perceptual experiment. The Sennheiser PC-36 headset presented the audio recordings. The experiment involved the SIT (Yorkston, Beukelman, & Trice, 1996), adapted visual analogue scales selected from relevant literature, and multiple choice options.

### 5.3. Procedure

Due to delayed recording and completion of the final two synthesised voices, participants attended two 30-minute sessions to allow recruitment of participants and initial data collection without requiring all the synthesised voices to be complete. This meant that each participant listened to the five Group A voices at the first session and the five Group B voices at the second session, as outlined in Table 8.

Table 8.

*Synthesised voice schedule for Group A and B.*

<b>Voice name</b>	<b>Age (nearest 5 years)</b>	<b>Listener group</b>
Male child	10	A
Young male	20	A
Young female	20	A
Middle-aged female	50	A
Older male	65	A
Female child	10	B
Māori male	20	B
Middle-aged male	40	B
Māori female	40	B
Older female	60	B

Participants attended the two 30-minute sessions in a clinic room at the University of Canterbury Child Language Centre. At the initial session, information sheets were read and consent forms were signed by the participant. Each participant was assigned a numerical code. The primary researcher gave each participant verbal instructions and visual sentence stimuli where appropriate as outlined in section 5.4. On completion of the second session participants received a \$10 petrol voucher for participation in the project.

### 5.4. Measures

As discussed in Chapter One it is important for synthetic voices to be intelligible, natural sounding, and to match the user's age and gender. Measures of intelligibility, naturalness, and age and gender identification were included in the current study and are discussed in the following sections.

### **5.4.1. Speech Intelligibility Test**

Sentence imitation tasks have been used to assess the intelligibility of synthesised speech. Von Berg, Panorska, Uken, and Qeadan (2009) used 30 sentence pairs randomly selected from *Sentences for Phonetic Inventory*. Participants were instructed to repeat verbatim each sentence while researchers judged words correctly repeated (Fairbanks, 1960). Jreige, Patel, and Bunnell (2009) conducted a usability study for a small sample of VocaliD voices where *Harvard Sentences* and *Semantically Unpredictable Sentences* were used as the stimuli for listeners to transcribe.

For the present study, the SIT was utilised for an adapted sentence imitation task. Instead of verbally repeating words such as in Mills et al. (2014), participants were asked to type what they heard in a text box. The SIT is a test of intelligibility through orthographic transcription of standard speech samples by unfamiliar listeners (Hustad, 2011). Listeners hear a sentence-length speech sample and then write down what they thought the speaker said. This method is widely used in clinical applications particularly for people with MND (Beukelman, Ball, & Fager, 2008). The primary researcher used the SIT CD to generate ten sets of 11 different sentences, each 5–15 words in length. The SIT sentence sets were randomly assigned to each of the ten synthesised voices to become the SIT sentence stimuli for each voice (see Appendix I). The SIT sentence stimuli were recorded with an internal playback and recording method through the MacBook Air terminal. This recorded individual audio files for each of the 11 SIT sentence stimuli as spoken by the synthesised voice. This procedure was completed for each of the ten synthesised voices. Using Praat, the primary researcher ensured that each SIT sentence stimulus was a similar intensity, so that intensity would not be an influencing factor for the experiment. The range of intensities was found to be 69–72 dB and deemed an acceptable range.

#### **5.4.1.1. Speech Intelligibility Test trial sentences**

Intelligibility was judged first to avoid familiarisation effects (Anand & Stepp, 2015). Participants were informed by the primary researcher that they would hear sentences ranging from 5 to 15 words in length and they were to type verbatim each word they heard into a textbox displayed on the MATLAB program. Participants were informed that they could start typing as soon as they heard speech and they did not need to wait for the sentence to be completely spoken before they began to type. This was to reduce the short-term memory load on the participants, especially for the longer sentences. Participants heard two trial sentences to ensure they understood the process before beginning the full experiment (see Table 9).

#### **5.4.1.2. Speech Intelligibility Test experiment**

The order of the 55 sentences in each session were presented to the participant at random. Participants were only able to listen each sentence once. The MATLAB program displayed a three second countdown before the presentation of the next sentence. There was no time limit for participants to type in their responses; they pressed enter once they had finished typing, and then the countdown for the next sentence began. The MATLAB program also displayed the number of remaining sentences left in the experiment. Table 9 details visual stimuli for this task, participant input, and MATLAB output.

#### **5.4.2. Intelligibility and naturalness visual analogue scales**

Two modified visual analogue scales were adapted from the following literature to explore the participants' perceptions of intelligibility and naturalness of the synthesised voices. Anand and Stepp (2015) investigated listener perceptions of mono-pitch, naturalness, and intelligibility for speakers with Parkinson's Disease. Participants provided ratings using a visual sort and rank method through a custom-designed user interface developed in MATLAB. Listeners heard each sentence stimulus and sorted the stimuli into ratings on a scale between 0 and 100 that represented two ends of the continuum, for example, *0=least natural* and *100=highly natural*. Hux, Knollman-Porter, Brown, and Wallace (2017) investigated the perceptions and auditory comprehension accuracy of 20 people with aphasia when listening to sentences generated with natural digitised speech and synthesised speech. A 5-point Likert scale was used for participant subjective preference rankings and ratings about ease of understanding, clarity of speech, and voice naturalness. The researchers asked the participants the following three questions in sequence: "*How easy was this voice to understand?*" "*How clear was this voice?*" and "*How natural sounding was this voice?*" The 5-point Likert-type scale where a score of one represented the lowest rating and a score of five represented the highest rating served as a means of quantifying participant preferences. One or two-word descriptors such as *Easy–Difficult* and *Natural–Unnatural* marked the high and low endpoints of the scale respectively.

The current study employed two visual analogue scales to investigate intelligibility and naturalness using an existing MATLAB program that was set up for visual analogue-style data collection. Four sentences (two for each scale so each voice was evaluated twice) were randomly selected from the previous SIT sentence stimuli (see Appendix I). Each synthesised voice was recorded with these four sentences using the internal playback process as outlined in Section 5.4.1. Visual stimuli with the two sentences for each scale were printed and laminated for the participants to refer to in this part of the experiment. This was to decrease the

chance that the initial voices heard speaking the sentences were rated lower due to unfamiliarity with the sentence stimuli. Table 9 shows the experiment schedule, including the number of sentences presented, visual stimuli, participant input mode, and output by MATLAB for the two visual analogue scales.

### **5.4.3. Age and gender identification**

The final experimental task required the participants to identify the age and gender of each of the synthesised voices. Two SIT sentence stimuli were randomly selected and recorded for each synthesised voice using the internal playback and recording method as discussed in Section 5.4.1. The MATLAB program was utilised to enable random presentation of the stimuli. Participants were instructed to type into the textbox the age (rounded to the nearest five years) and gender of the voice they heard. Table 9 shows further details for the age and gender identification task including number of sentences presented, visual stimuli, participant input mode, and output by MATLAB.

### **5.5. Data analysis**

The ten synthesised voices were analysed as individual case studies. The study was not interested in the performance of individual listeners or listener groups. All data collected from the measures were exported as Microsoft Excel spreadsheets. The SIT data consisted of transcribed sentences, where the primary researcher scored each word as correct or incorrect based on whether they matched the intended words of the speaker. Percent intelligibility scores were calculated using:

$$\text{intelligibility (\%)} = \frac{\text{correct words}}{\text{total words}} \times 100$$

Descriptive analyses were carried out to calculate basic measures of range and central tendency. Independent sample *t*-tests were employed as a way of assessing the difference between male and female synthesised voice intelligibilities. The relationship between sentence length and percentage of words correct was investigated using the *R*-squared statistical measure of how close the data are to the fitted regression line. The relationship between the number of times participants heard the synthesised voice and the percentage of words correct were also investigated using this method. The two visual analogue scale data consisted of numerical values where descriptive statistics were employed to calculate means and standard deviations. Independent sample *t*-tests were carried out to assess the difference between the two sentence stimuli used in each visual analogue scale. Correlations between the two measures of

intelligibility and the two visual analogue scales were calculated using *R*-squared statistics. The age estimation data was converted into positive or negative numbers to represent the difference between participant estimations and the speakers' chronological ages. Descriptive statistics including means and standard deviations were calculated. Gender identification data was analysed by calculating percentage correctly identified.

Table 9.

*Perceptual experiment schedule.*

	<b>Sentences per session</b>	<b>Sentences per voice</b>	<b>Printed visual stimuli</b>	<b>Visual stimuli on screen</b>	<b>Input mode</b>	<b>Output by software</b>
<b>SIT trial</b>	2 (Primary researcher's synthesised voice)	2	No	"Input what you hear"	Participant typed words heard	Excel sheet with typed responses
<b>SIT</b>	55	11	No	"Input what you hear"	Participant typed words heard	Excel sheet with typed responses
<b>Intelligibility visual analogue scale</b>	10	2	2 stimuli sentences	"How easy is this voice to understand?"	Participant moved line on visual analogue scale with the left side being <i>difficult</i> and the right side being <i>easy</i>	Excel sheet with participant selection converted into numerical responses 0–100 with 2 decimal points
<b>Naturalness visual analogue scale</b>	10	2	2 stimuli sentences	"How natural does this voice sound?"	Participant moved line on visual analogue scale with the left side being <i>unnatural</i> and the right side being <i>natural</i>	Excel sheet with participant selection converted into numerical responses 0–100 with 2 decimal points
<b>Age and gender predictions</b>	10	2	2 stimuli sentences, list of ages in 5 year increments from 5–70, and <i>male</i> or <i>female</i> option	"Input what you hear"	Participant typed selected age and gender	Excel sheet with typed responses

## Chapter 6. Perceptual experiment results

Fifteen unfamiliar listeners completed the perceptual experiment. The following chapter will report listeners' perceptions of the intelligibility and naturalness of the synthesised voices, and the estimations of the synthesised voices' age and gender. This chapter relates to the evaluation of the New Zealand-accented synthetic voices for SGDs and the two subsequent research questions:

- c. *How do unfamiliar listeners perceive the intelligibility and naturalness of the New Zealand-accented voices created using ModelTalker technology?*
- d. *Can unfamiliar listeners perceive the age and gender of the New Zealand-accented voices created using ModelTalker technology?*

### 6.1. Speech Intelligibility Test

Figure 8 shows the synthesised voices' intelligibility at word level with the SIT. The male child's voice had the lowest average intelligibility at word level with minimum, median, and maximum scores of 60.17%, 73.53%, and 82.90% words correct. The older male's voice had the second lowest percentage of words correct with minimum, median, and maximum scores of 67.37%, 86.00%, and 89.72%. Both voices' maximum scores were below the majority of the other eight voices' median scores, indicating that most unfamiliar listeners found the male child and older male's voices to be less intelligible. The young female and the Māori female's voices had similar median scores to the older male with 86.12% and 88.27%. However, these two voices' minimum and maximum scores were higher at 73.36% and 93.00% for the young female and 75.64% and 93.07% for the Māori female voice. The young male's voice had a minimum, median, and maximum score of 78.63%, 91.45%, and 97.27%. The remaining five voices all had maximum scores of 100% words correct. The female child's voice had a wide range of scores with minimum and median scores being 75.64% and 91.43%. The middle-aged male's voice had a minimum and median score of 82.07% and 94.49%. Although the Māori male's voice had a lower minimum score of 85.18% compared to the middle-aged female and older female's voices (88.45% and 88.67%), the Māori male's voice had the highest median score with 98.47%. The median score for the middle-aged female and older female's voices were 95.02% and 96.96%.

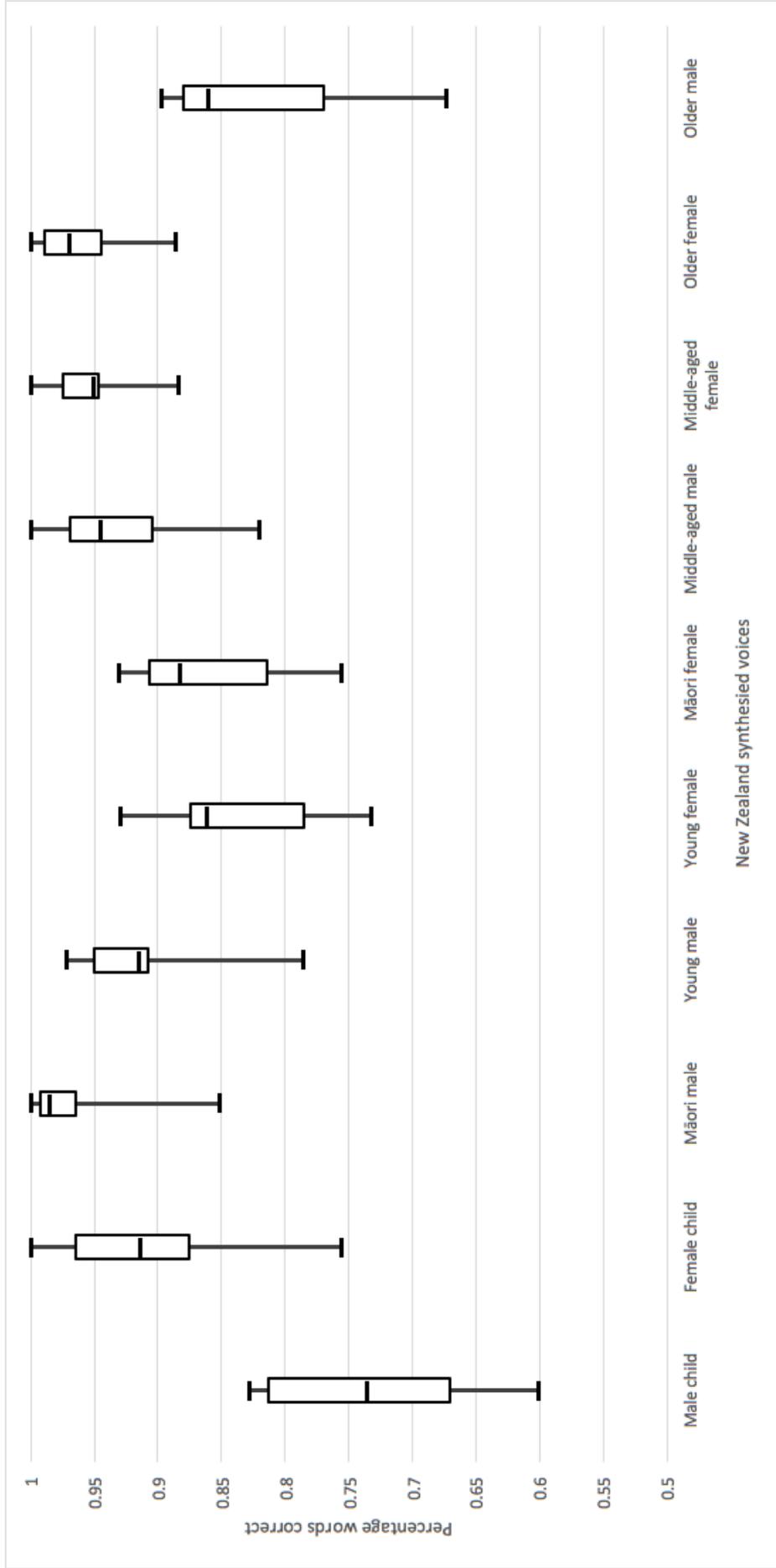


Figure 8. New Zealand-accented voices intelligibility at word level with the Speech Intelligibility Test

Independent sample *t*-tests were conducted to compare the mean intelligibility of the male voices compared to the female voices. The means, standard deviations, *t*-values, and *p*-values are shown in Table 10. The results show there is no significant difference at  $p \leq 0.05$  of percentage words correct between the male and female synthesised voices.

Table 10.

*T-test statistics for male and female synthesised voices.*

	Male voices		Female voices		<i>t</i> -value	<i>p</i> -value
	Mean	Standard deviation	Mean	Standard deviation		
Percentage words correct	87.48	9.44	90.52	5.61	-0.618	0.554

Sentence length was investigated to see if the number of words in the SIT sentence influenced percentage of words correct. Figure 9 shows the relationship between sentence length and SIT accuracy. As sentence length increases, percentage of words correct tends to decrease. The *R*-squared value of the linear trend line was 0.47 indicating that sentence length explains 47% of the variation in percentage of words correct. The male child's SIT scores were further analysed to investigate if increasing sentence length and decreasing intelligibility were linked for the less intelligible voice, however the *R*-squared value for the trend line was 0.098 indicating that sentence length did not explain a large proportion of the variability in the male child's intelligibility scores.

Exposure to the synthesised voices were investigated to see if number of times hearing the voice in the SIT influenced percentage of words correct. Figure 10 shows the relationship between exposure of the voice and SIT accuracy. The *R*-squared value for the trend line was 0.0073 indicating that number of exposures to the synthesised voice did not explain a large proportion of the variability of number of words correct. The male child SIT scores were further analysed to investigate if exposure increased intelligibility for this voice, however the *R*-squared values for the trend line was 0.029 indicating limited significance.

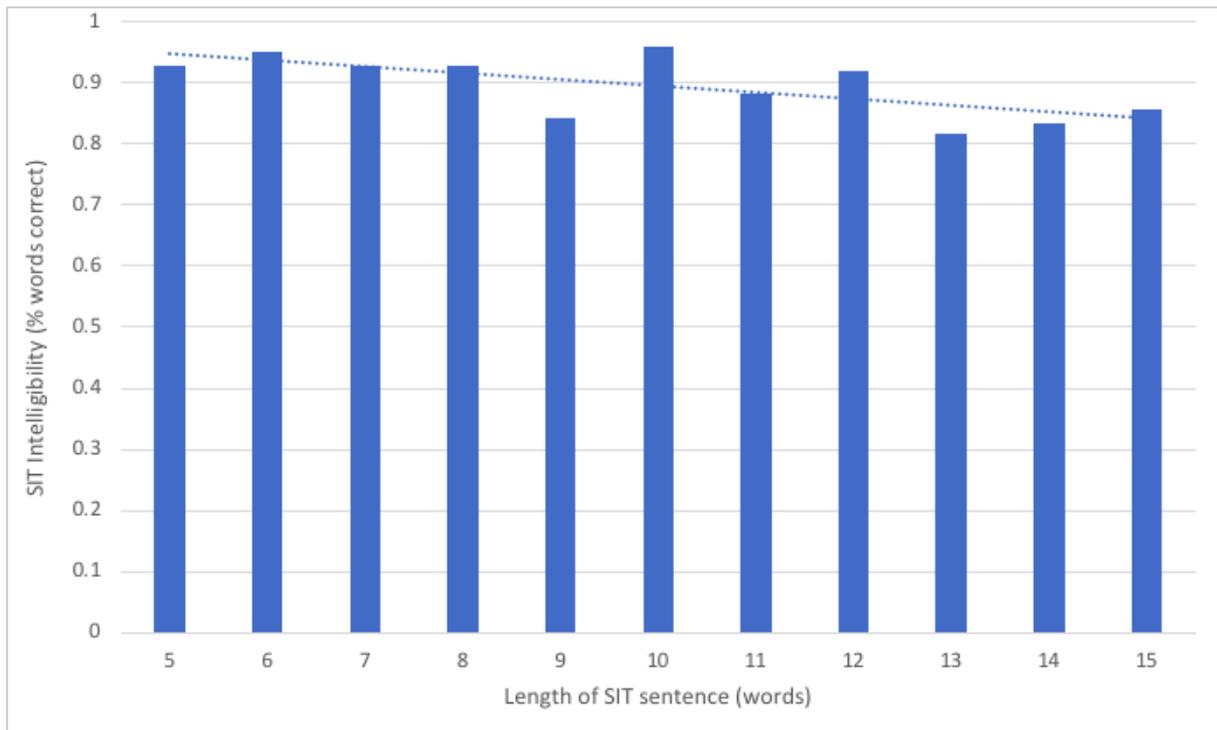


Figure 9. Relationship between length of Speech Intelligibility Test sentence stimulus and percentage words correct.

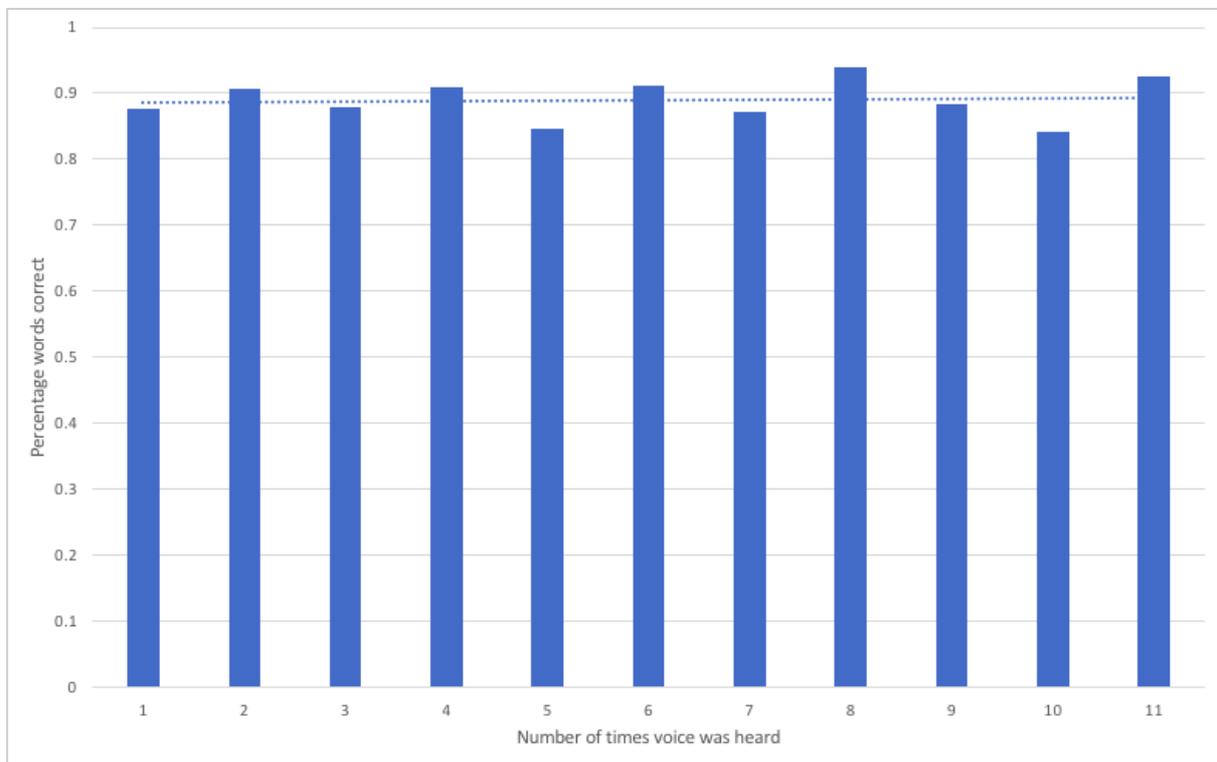


Figure 10. Relationship between exposure of synthesised voices and percentage words correct.

## 6.2. Intelligibility visual analogue scale

Figure 11 shows the average ratings of intelligibility from the visual analogue scale, including error bars of one standard deviation from the mean. The male child's voice has the lowest ratings of intelligibility with a mean of 5.61 out of a possible 10 (standard deviation = 2.18). The female child, Māori male, young female, Māori female, middle-aged male, and older male's voices had mean scores of 7.97, 7.65, 7.79, 7.39, 7.86 and 7.76. Standard deviations ranged from 1.79 (middle-aged male) to 2.16 (Māori male). The young male, middle-aged female, and older female showed highest average ratings of intelligibility with means of 8.52, (standard deviation = 1.77), 8.52 (standard deviation = 1.75), and 8.37 (standard deviation = 1.59). The error bars for one standard deviation above the mean for the young male and middle-aged female's voices extend beyond the maximum score of 10.

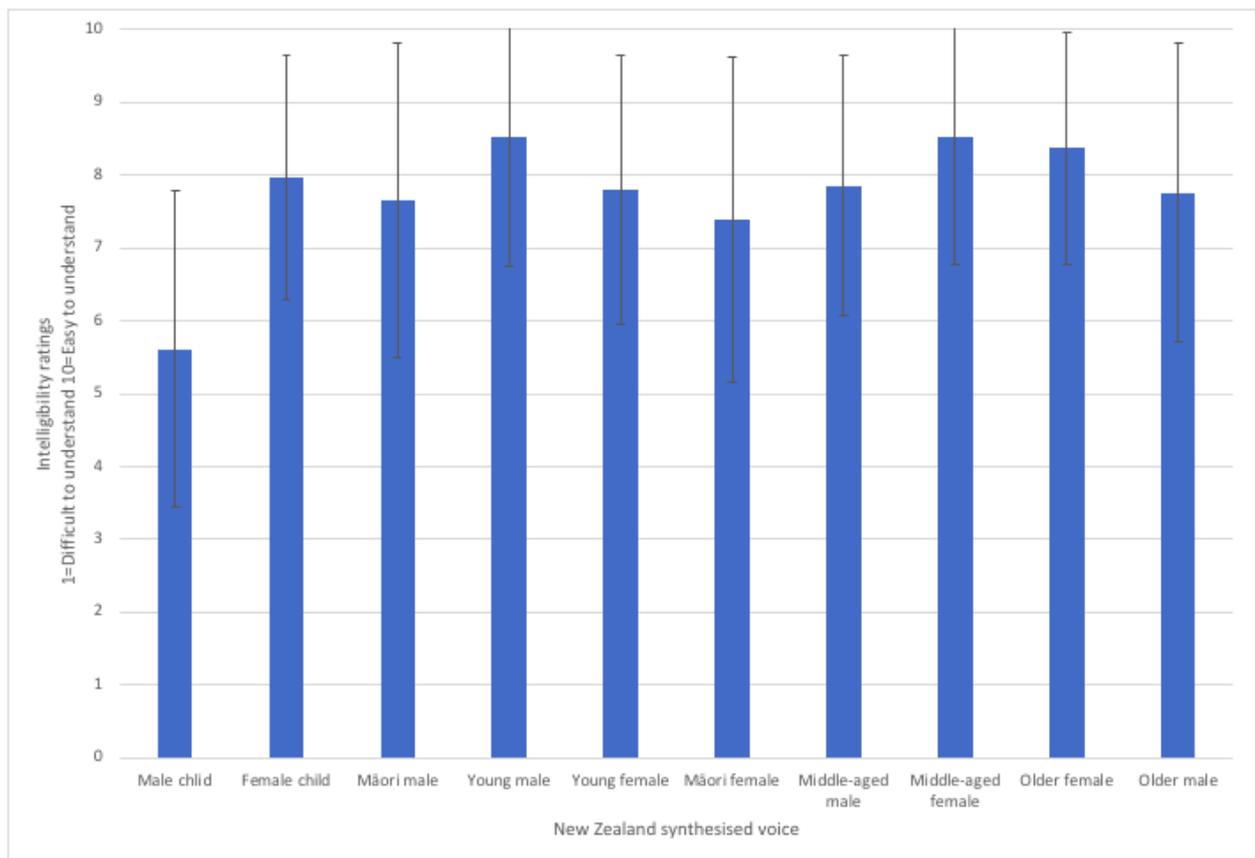


Figure 11. New Zealand synthesised voices average intelligibility visual analogue ratings.

Note. Error bars are one standard deviation.

Participants rated each voice twice on the intelligibility visual analogue scale. The average within listener difference in ratings were 1.57 points. The minimum within listener ratings were 0.23 points and the maximum was 2.63 points. A *t*-test was performed to investigate if there was a difference in ratings between sentence A “*The new kitchen shelves were mounted to the wall*” and sentence B “*The facility is open to the public, and you may visit.*” The means, standard deviations, *t*-values, and *p*-values are shown in Table 11. The results show there was no significant difference at  $p \leq 0.05$  between sentence A and B.

Table 11.

*T-test statistics for sentence A and B in the intelligibility visual analogue measure.*

	Sentence A		Sentence B		<i>t</i> -value	<i>p</i> -value
	Mean	Standard deviation	Mean	Standard deviation		
Intelligibility rating	7.51	1.15	7.98	0.931	-0.993	0.167

A *t*-test was performed to investigate if there was a difference in ratings between the sentence heard first compared the sentence heard second for each synthesised voice. The means, standard deviations, *t*-values, and *p*-values are shown in Table 12. The results show there was no significant difference at  $p \leq 0.05$  in intelligibility ratings between the two exposures to the synthesised voices.

Table 12.

*T-test statistics for first heard sentence and second heard sentence in the intelligibility visual analogue measure.*

	First heard		Second heard		<i>t</i> -value	<i>p</i> -value
	Mean	Standard deviation	Mean	Standard deviation		
Intelligibility rating	7.75	1.56	7.73	0.709	0.0447	0.418

Figure 12 shows the correlation between the two measures of intelligibility used in the present study: SIT, and the intelligibility visual analogue scale. The linear trend line shows the  $R$ -squared value of 0.62 indicating a moderate positive correlation between average SIT scores and average intelligibility visual analogue ratings.

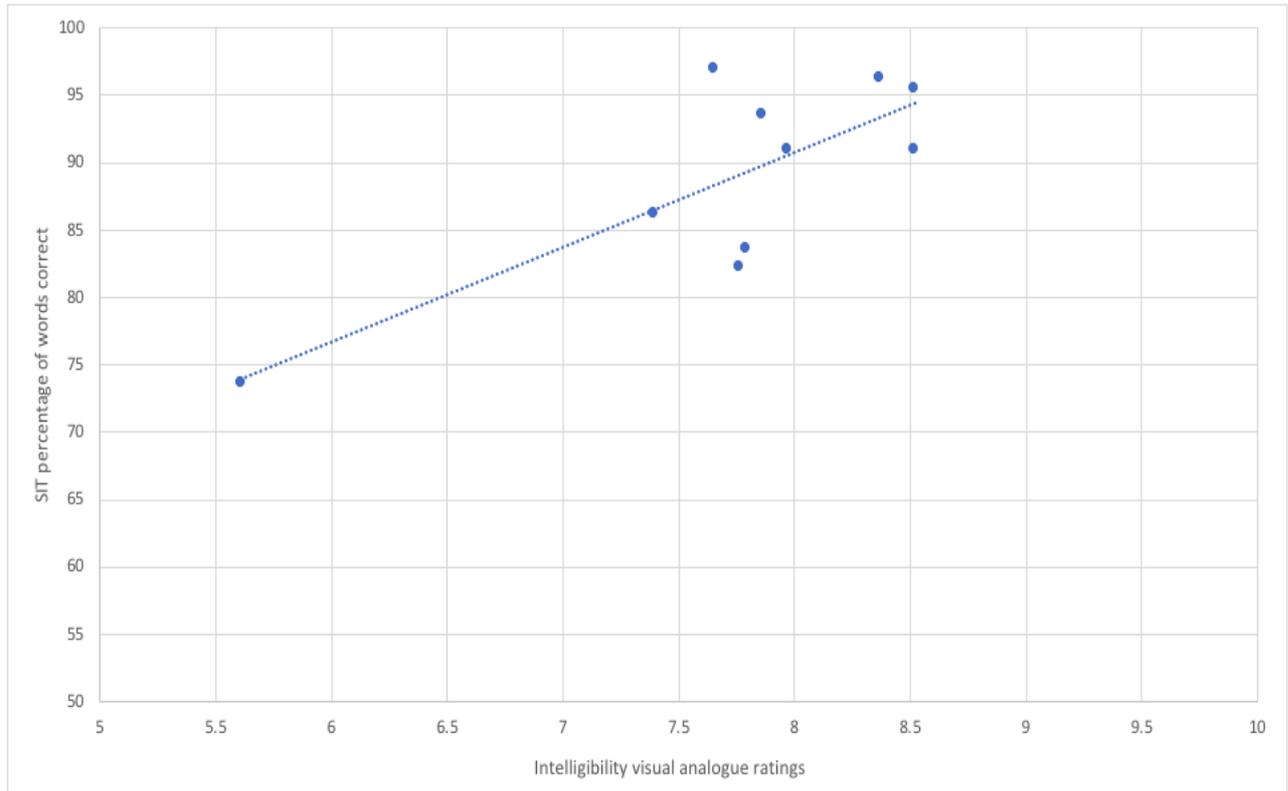


Figure 12. Correlations between Speech Intelligibility Test percentage of words correct and average intelligibility visual analogue ratings.

### 6.3. Naturalness visual analogue scale

Figure 13 shows the New Zealand synthesised voices' average ratings for naturalness. The female child voice has the lowest mean score of 3.90 out of a possible 10 (standard deviation = 2.68). The male child, middle-aged male, and older male's voices have means below 5 (the half way mark on the scale) with scores of 4.37 (standard deviation = 2.76), 4.85 (standard deviation = 2.46), and 4.59 (standard deviation = 2.58). The Māori male, older female, Māori female, young male, and middle-aged female's voices have means in the 5s and 6s with 5.25, 5.59, 6.07, 6.17 and 6.27. Standard deviations ranged from 2.29 (older female) to 3.07 (middle-aged female). The young female's voice had the highest mean of 7.75 and the lowest standard deviation of 1.75 across the ten voices.

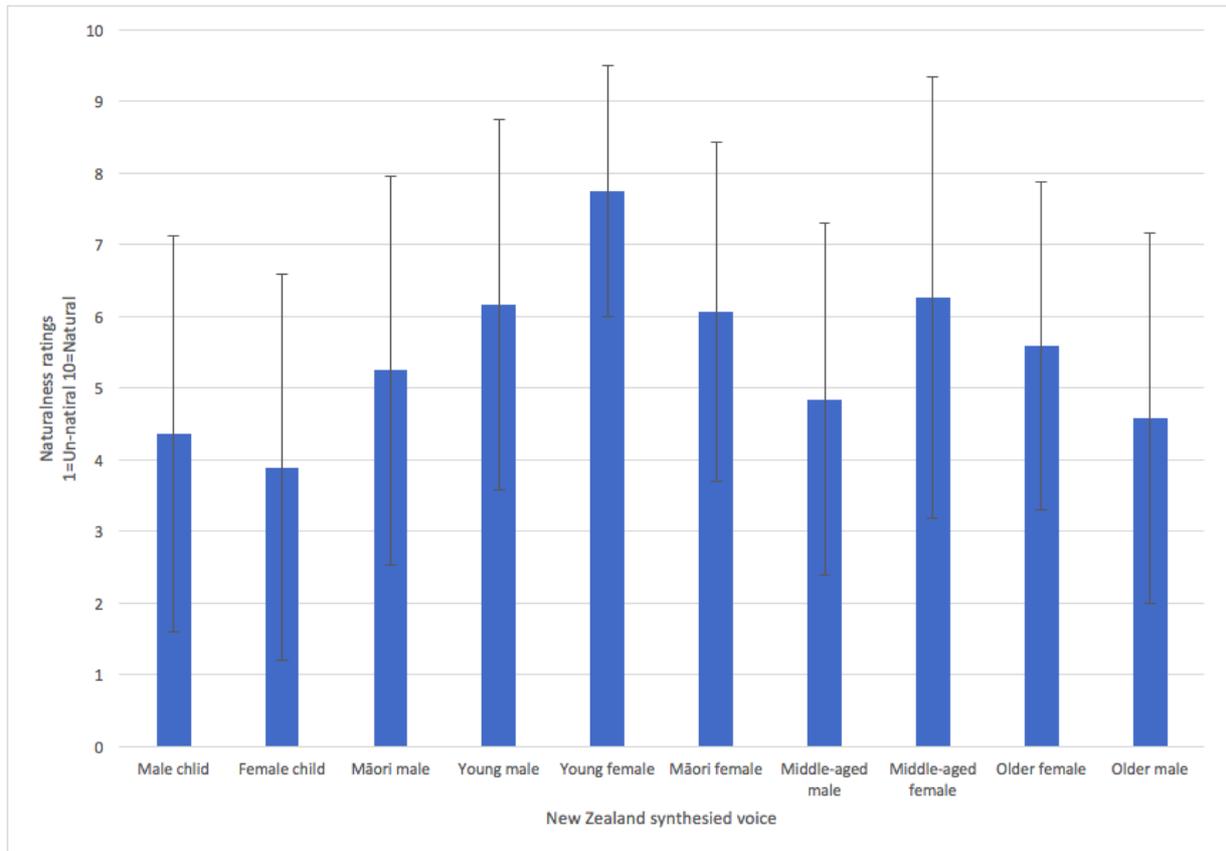


Figure 13. New Zealand synthesised voices average naturalness visual analogue ratings

Note. Error bars display one standard deviation.

The average unfamiliar listeners score for sentence A “*Leave the rest on the table for later*” was higher than sentence B “*He was in no way prepared for what might happen*” for all ten voices. A t-test was performed to investigate if there was a difference in ratings between the two sentences. The means, standard deviations, *t*-values, and *p*-values are shown in Table 13. The results show there was a significant difference at  $p \leq 0.05$  between sentence A and B.

Table 13.

*T*-test statistics for sentence A and B in the naturalness visual analogue measure.

	Sentence A		Sentence B		<i>t</i> -value	<i>p</i> -value
	Mean	Standard deviation	Mean	Standard deviation		
Naturalness rating	7.98	0.931	7.51	1.15	-3.70	0.000887

The correlations between the ratings for the synthesised voices using the two visual analogue scales were investigated. There were no strong correlations between the mean intelligibility visual analogue scores and mean naturalness visual analogue scores with an *R*-squared value of 0.127.

#### **6.4. Age estimation**

Figure 14 shows unfamiliar listeners' age estimations for the New Zealand synthesised voices. The male child and female child's voices were the most accurately estimated with mean deviations from the speakers' chronological age being -0.33 years (standard deviation = 1.26) and 0.67 years (standard deviation = 1.73). The two child voices had the smallest standard deviations. The Māori male voice's mean age estimation was 11.50 years (standard deviation = 9.29) older than his chronological age, with one standard deviation below the mean falling above zero indicating that the listeners perceived the voice to be older than his chronological age. The young male's voice had a mean estimate of 4.50 years above his chronological age. The young female's voice had a mean age estimate of 6.83 years (standard deviation = 5.94) older than her chronological age, with one standard deviation below the mean falling above zero indicating that the listeners perceived the voice to be older than her chronological age. The Māori female's voice had a mean age estimate of -7.67 years (standard deviation = 7.16) indicating that listeners perceive the voice to be younger than her chronological age. The middle-aged male's voice had a mean estimation of 2.33 years above his chronological age; however, the standard deviation of 7.85 years, created a one standard deviation error range above and below zero. The middle-aged female and older female had mean estimates of -4.50 and -7.00 years respectively, which are below their chronological ages. Both these voices had standard deviations (7.11 and 9.43 respectively) where the one standard deviation above the mean fell above zero. The older male's voice had a mean age estimation of -19.50 years (standard deviation = 8.03), indicating that the listeners perceived the voice to be younger than his chronological age.

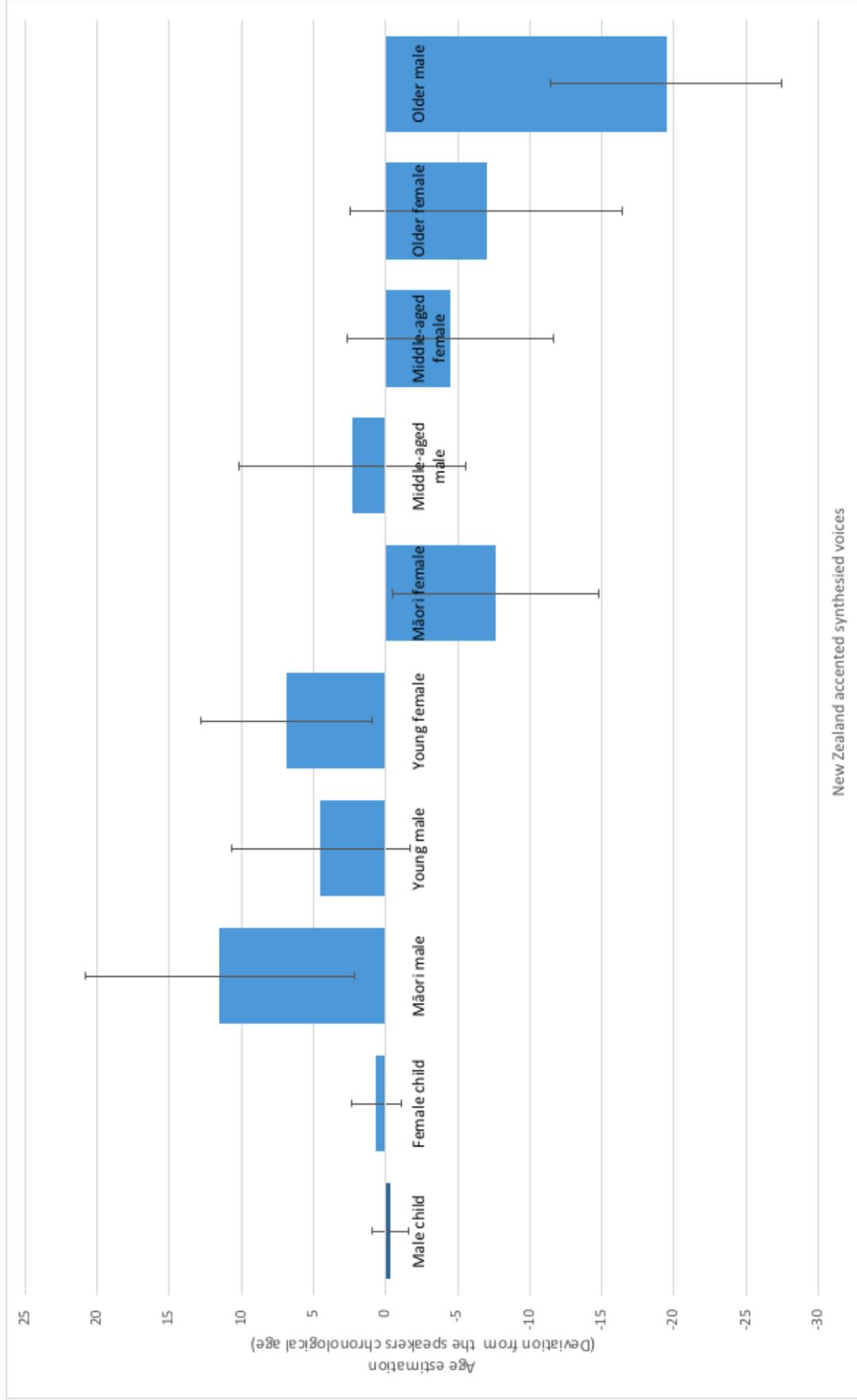


Figure 14. Unfamiliar listeners age predictions for the New Zealand synthesised voices.

Note. Error bars for each voice are calculated at one standard deviation. Voices are arranged in chronological order from left to right.

## 6.5. Gender identification

Figure 15 shows the unfamiliar listeners' gender identifications for the synthesised voices. The male child, Māori male, young male, middle-aged male, and older male's voices were all identified as male voices for 100% of the listeners' responses. The young female, Maori female, and middle-aged female's voices were identified as female voices for 100% of the listeners' responses. The older female's voice was identified as female 97% of the time. The female child's voice was identified as female in 3% of responses, with the majority of listeners perceiving the female child's voice as a male's voice.

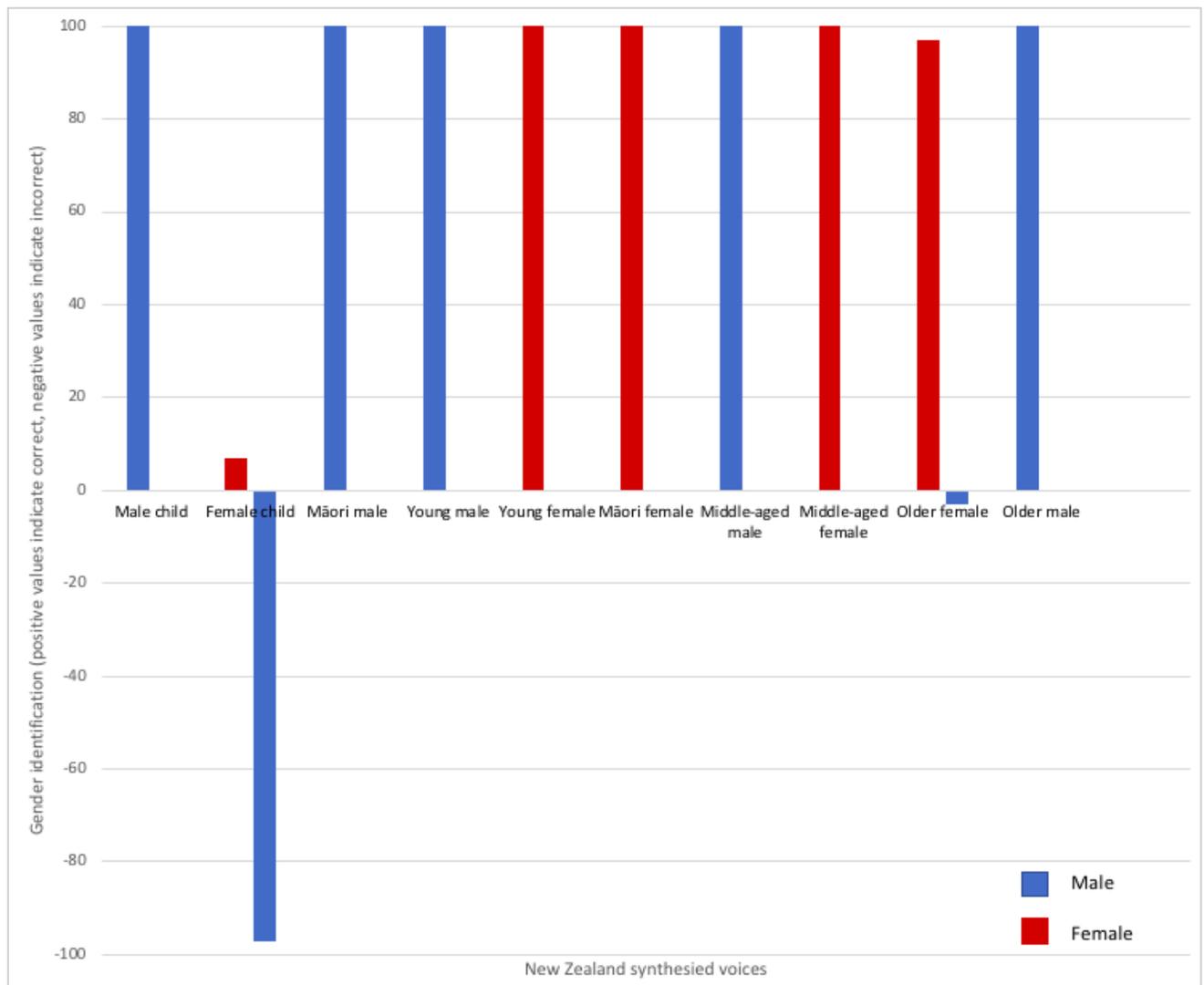


Figure 15. Gender identification of the New Zealand synthesised voices.

Note. Positive scores indicate correct responses, negative scores indicate incorrect responses.

## Chapter 7. Discussion

This study examined the need and suitability of New Zealand-accented synthesised voices created using ModelTalker voice banking technology. This chapter discusses the main findings of the present study and is organised according to the two aims of the study: (1) to explore the effectiveness of the ModelTalker voice banking protocol for New Zealand speakers, and (2) to create and evaluate the New Zealand-accented synthetic voices for speech generating devices. This discussion will begin by examining the results from the voice donor participant experiences of using ModelTalker to voice bank. Following this, results from the unfamiliar listener perceptual experiment are considered in relation to the intelligibility, naturalness and age and gender characteristics of the New Zealand-accented synthesised voices. The strengths and limitations of the study are incorporated throughout the discussion. This chapter ends with a brief discussion regarding applications for voice banking with clinical populations and future directions for research.

### **7.1. Aim One: To explore the effectiveness of the ModelTalker voice banking protocol for New Zealand speakers**

To investigate the use of ModelTalker voice banking technology with New Zealand speakers, ten voice donors underwent the voice banking process and completed a questionnaire relating to their experiences.

#### **7.1.1. What is the experience of healthy voice donors during the ModelTalker voice banking process?**

This research question was explored through hypotheses I and II.

Hypothesis I: *Voice banking with ModelTalker technology is a positive experience.* The results from the voice donor questionnaire yielded multiple themes and subthemes regarding the experiences of voice banking. The first theme was gaining awareness. Participants increased their knowledge about voice banking. Many of the participants did not previously know about voice banking technologies, and the project expanded their own awareness of services that support people with conditions of progressive speech loss. Education and awareness is vital for clients and their families, but also for those in the community who otherwise may not know of services that are available. It is important for the process of voice banking to be discussed and be available early for those with progressive speech loss to minimise the effect of dysarthria on the synthesised voice (Bunnell et al., 2017; Costello, 2016; Jackson et al., 2017). The voice donor participants appreciated and recommended voice banking as an opportunity to help other people. Being able to take part in a project which

involved donations to others was a positive factor for most of the participants. Generosity research indicates that people who make generous choices show an increase in self-reported happiness (Park et al., 2017). This may link to the positive ratings of this experience for many of the voice donors. The voice donors also reflected on what their voice meant to them and what it would be like if they were unable to speak which has similarities to Nathanson (2017).

Hypothesis II: *Children and adults can voice bank using ModelTalker technology with ease.* Consistent with Creer et al. (2009) and Jackson et al. (2017), the participants summarised that the ModelTalker process was easy. One limitation when comparing the present study to the two mentioned studies is that the primary researcher of the present study did the majority of the behind-the-scenes set up for the participants. The voice donors were not involved in creating a login and were not engaged in email communication with ModelTalker staff. Instead, for consistency, the primary researcher completed all these steps without the involvement of the voice donors. This allowed the primary researcher to give standard explanations of the instructions for each voice donor, however it reduced the complexity of the tasks each voice donor had to complete in order to go through the whole voice banking process. Hyppa-Martin et al. (2017) indicated that some of the ModelTalker instructions were unclear and required clarification, and some communication was redundant. Hyppa-Martin et al. (2017) also commented that computer skills of people voice banking should be considered, as although many parts of the process were very straightforward, there were a few intermediate computer skills such as installing programs and checking sound settings that may cause some difficulties for less confident computer users. The present study did not gather perspectives on the above factors due to the primary researcher setting up the behind-the-scenes aspects of voice banking. It is unlikely that speech-language therapy funding would allow for every patient in the real-world to be able to receive on-on-one assistance to setup and record all 1,600 sentences for the ModelTalker process. That being said, the actual process of recording, i.e., listening back to recordings and re-recording, were independently completed by all participants except the youngest male child. The female child and eight adult participants were independent in the physical process of voice banking, indicating that it would be plausible for speech-language therapists to assist clients to set up the voice banking process and then have the clients complete the recordings independently while checking in with their speech-language therapist if necessary.

Voice donors deemed the speed and volume dials to be useful. The speed and volume dials provided feedback for the voice donors which could have helped to keep them on track for the repetitive task of recording. Das (1982) found that quantity and quality feedback helped

workers of repetitive tasks increase the quantity and quality of their output. Related to the present study, quantity feedback was provided in the form of the number of sentences completed displayed on the ModelTalker interface and the speed and volume dials may have acted as quality measures. The nine-year-old child required additional motivation in the form of goals set by the primary researcher to assist his endurance to continue recording. This indicates that the ModelTalker system may need additional external motivation factors to be successfully completed by younger children. The female child was older at thirteen years and was able to record independently. She was a strong reader and had great awareness of what her voice donation would be used for, which may have increased her motivation. The ModelTalker feature of listening to the progression of the synthesised voice likely acted as a motivational factor and a form of feedback for many of the participants. The male child disliked hearing his own voice to begin with, so this feature was not a source of motivation for him at the initial stages of recording. The voice donors also identified barriers that could prevent other people from voice banking which included time constraints, limited attention span, poor eyesight, being literate, awareness of services, and access to recording areas. Interestingly, many of the factors discussed by participants may contribute to why the male child experienced more difficulty with the voice banking process. He had a noticeably decreased attention span compared to the adult voice donors, he required assistance to read the sentences, and recording sessions interfered with his usual afterschool activity routines.

Challenges of ModelTalker voice banking identified by the voice donors were similar to those reported by Hyppa-Martin et al. (2017) and Jackson et al. (2017). Time constraints and number of sentences recorded were common subthemes, even though the mean recording time of 5 hours 30 minutes was lower than the six to eight hours reported by other literature (Hyppa-Martin et al., 2017; Jackson et al., 2017; ModelTalker, 2018). It should be noted that the recording duration figures for the current study include total time at recording sessions, therefore time spent re-recording and listening back to the preview synthesised voice is included in the duration data. It is expected that all people who voice bank would re-record and listen back to their recordings to preview their synthesised voice. With the exception of the identification of the youngest participant's longer recording duration, detailed calculations on recording duration related to age and gender were not carried out because each voice donor was seen as a case study for each age and gender group.

Due to the nature of the research methodology, voice donors attended set sessions which were planned in advance. Clients completing voice banking who have progressive speech loss are likely to have the option to record in their home environment and could

complete the recordings at their own pace. This may alter the perspectives on recording duration. It could also alter challenges such as energy levels, fatigue, and ergonomics identified by the voice donors and consistent with Hyppa-Martin et al. (2017) and Jackson et al. (2017). Clients would be able to voice bank from the comfort of their own home and could alter recording sessions to match their energy and fatigue levels. However, background noise and distractions at home may cause increased challenges for people, as indicated by some voice donors. As with the University of Minnesota Duluth voice banking clinic described in Hyppa-Martin et al. (2017), options for clients to use sound booths to complete their voice banking could be a factor to consider. This would allow increased options for recording locations if clients find their home is unsuitable due to noise levels, distractions, or other factors. Advances in technology may also assist in reducing the recording duration, as well as increasing the interest and engagement in the sentences read aloud. Costello and Dellea (2017) discussed future technologies where patients' phrase-banked recordings could be analysed and combined to create synthetic voices. Instead of patients doubling up on recordings by needing to record phrase-banked sentences and the specific voice banking sentences, future technologies could use the phrase-banked recordings to create the personalised voices. Using the same recordings for two purposes would be a huge advantage for people with progressive speech loss or dysarthria (Bunnell et al., 2017; Costello & Dellea, 2017). A second change to voice banking technologies discussed by Costello and Dellea (2017) was that people who voice bank could choose the sentences that they record from a selection of written materials. It was suggested that materials such as fiction stories could increase the engagement and enjoyment of the recording process through reading aloud sentences that make up a story. This could help to decrease the repetitive feel of the voice banking process.

#### **7.1.2. Are there any alterations that are required to make ModelTalker voice banking suitable for New Zealanders?**

This research question was explored through hypothesis III: *ModelTalker can be applied to a New Zealand context*. As discussed in the methodology and consistent with Creer et al. (2009), voice donors were instructed to ignore the American pronunciation meter on the ModelTalker interface. Voice donors indicated that the American model voice on the system influenced their own pronunciation when recording sentences and many participants opted to mute the model voice. Muting the model voice likely contributed to the lower recording durations compared to previous studies as voice donors did not need to wait for the auditory prompt before recording (Hyppa-Martin et al., 2017; Jackson et al., 2017; ModelTalker, 2018). The model voice

however was deemed helpful for use as a pronunciation guide for unfamiliar words. As discussed previously, advances in technology to allow voice bankers to choose what materials they record could also reduce the number of unfamiliar words. Having New Zealand vocabulary within the sentences was also suggested by the voice donors. Alongside New Zealand themed sentences, ideally there would be a New Zealand English model voice on the system to allow auditory cues to match the user's accent. A New Zealand English pronunciation meter would further assist the voice banking software to be suitable for New Zealanders. For the time being, muting the model voice and ignoring the pronunciation meter appears to be an appropriate alteration to allow voice banking success for New Zealand speakers. The inventories of the te reo Māori words were effective in phrase banking a small sample of everyday te reo Māori words. All participants recorded 20 common te reo Māori words, and the two te reo Māori speakers recorded an additional 65 core vocabulary words. A fully synthesised te reo Māori voice is needed to support SGD users who wish to communicate in te reo Māori. Having a synthesised te reo Māori voice would enable te reo Māori use by those who already are familiar with the language, and would also provide a platform for new speakers to learn the language. This would support te reo Māori language revitalisation, something which is important for New Zealand culture (New Zealand Ministry for Culture and Heritage, 2018).

## **7.2. Aim Two: To create and evaluate the New Zealand-accented synthetic voices for speech generating devices**

To evaluate the ten New Zealand synthesised voices created by ModelTalker technology, fifteen unfamiliar listeners completed a perceptual experiment which included measures of intelligibility, naturalness, and age and gender identification.

### **7.2.1. How do unfamiliar listeners perceive the intelligibility and naturalness of the New Zealand-accented voices created using ModelTalker technology?**

This research question was explored through hypotheses IV and V.

Hypothesis IV: *The synthesised ModelTalker voices are intelligible to native New Zealand listeners.* The SIT was used to gather data related to the intelligibility of the New Zealand synthesised voices. The mean intelligibility of the adult voices ranged from 86.00% to 98.47% of words intelligible. Studies within the last decade report synthesised speech intelligibility between 85.00% and 97.60% for adult voices in open response formats, indicating that the New Zealand-accented synthesised voices were comparable to other commercially available voices (Jreige et al., 2009; Von Berg et al., 2009). As with most cases

of synthesised speech, natural speech on average is still likely to be more intelligible with scores of 97.20% to 99.00% correct at word level, but the intelligibility of synthesised speech is continuing to improve over time (Greene et al., 1984; Jreige et al., 2009; Logan et al., 1989; Mirenda & Beukelman, 1987; Von Berg et al., 2009). In the present study the mean intelligibility scores for the male and female child voices were 73.53% and 91.45%. Von Berg et al. (2009) reported mean intelligibility scores of 55.24% for the DECtalk child voice and 71.95% for the VeriVox child voice. Both the child voices in the present study received higher intelligibility scores than to the DECtalk and VeriVox child voices, with the female child's intelligibility comparable to the adult voices. Consistent with Jreige et al. (2009), the present study found no differences between male and female voice intelligibility. One limitation of the present study was that there were no control voices in the SIT. To further investigate the intelligibility of the New Zealand synthetic voices, other synthetic voices available on SGDs could be used as controls to compare the present studies' voices to the synthesised voices that are commercially available.

A relationship was found between sentence length and number of words correct in the SIT; however, no relationship was found between exposures to the synthesised voices and number of words correct. Interestingly the opposite was found in Venkatagiri (1994) where no significant difference was found between sentence length and intelligibility. Sentences with a mean length of five words had the same intelligibility as sentences with a mean length of eleven words (Venkatagiri, 1994). Venkatagiri (1994) had hypothesised that longer sentences would use more short-term memory in order to hold the information due to cognitive strain. However, following the results of the study, Venkatagiri (1994) discussed the idea of linguistic/cognitive 'chunking' of information into groups of grammatically related words. Venkatagiri (1994) had participants write the words that they heard after hearing the whole sentence. This is different from the present study where participants were able to type the words as they heard them; they did not need to wait for the completion of the sentence. Although detailed analysis was not carried out, a trend noticed by the primary researcher when scoring the SIT responses was that some participants omitted more words near the end of longer sentences compared to the start of the sentence. It is possible in the present study when participants performed multiple tasks simultaneously i.e. hearing the sentence and typing the sentence, this influenced the number of words correct in the longer sentences, as the final words were forgotten. The present study also used sentences up to 15 words in length which was longer than Venkatagiri (1994), where the longest sentence stimulus was 11 words. A limitation in the present study was that there were 11 sentence stimuli for each synthesised voice, which is far fewer than Venkatagiri (1994).

Each synthesised voice only had one sentence stimulus for each sentence length. Results for sentence length investigations consisted of only one sample for each length of sentence per voice, which may have influenced results. Future studies should further investigate relationships between sentence length and intelligibility considering participant response methods such as: comparisons between writing, typing, and verbally repeating stimuli; trialling longer sentences; and increasing the sample size for each sentence length.

When people hear novel utterances, it may take time to accommodate to the speech patterns of an unfamiliar accent or to the speech of a young child. It can be much easier to understand heavily accented speech or the speech of a young child once the listener gets used to the speech patterns. Exposure to synthesised speech in contemporary literature follows this trend. Venkatagiri (1994) found an increase in intelligibility in the last 30 sentences that the participants heard compared to the first 30 sentences. Venkatagiri (1994) indicated there may be a practice effect where people understand synthesised speech better over time. This is consistent with McNaughton et al. (1994) who found that intelligibility scores for synthetic speech increased significantly with repeated listening sessions across a number of days for both adult and child listeners. Reynolds, Isaacs-Duvall and Haddox (2002) found that response latencies shortened over time for participants making true or false judgements for statements spoken by synthesised speech. In the present study, no significant relationship was found between exposure to the synthesised voice and intelligibility. However the 11 sentences which were randomly presented for each voice was a considerably smaller number of sentence stimuli compared to Venkatagiri (1994) and the study did not include repeated listening across different days as in McNaughton et al. (1994). Due to these factors it is unlikely that the measures used in the present study accurately captured any repeated listening effects if they were present.

The intelligibility visual analogue scale found that the mean ratings of intelligibility ranged from 5.61 to 8.52 out of a possible score of 10. With the exception of the male child's voice, the nine other voices all had average scores above 7 which indicates that listeners were rating the voices towards *easy to understand* rather than *difficult to understand* on the scale. Moderate positive correlations were found between the averages of the two measures of intelligibility: the SIT and the intelligibility visual analogue scale. The New Zealand synthetic voices which scored higher in intelligibility with the SIT tended to yield higher intelligibility ratings on the visual analogue scale and vice versa. The standard deviations for the average ratings of intelligibility on the visual analogue scale were wide and there was an average within listener variability of 1.57 points. It is unlikely that the intelligibility visual analogue scale used

in the present study would be a consistently good predictor of SIT scores. Drager et al. (2010) suggests that intelligibility measures may give an estimate of a listener's ability to recognise the words in synthetic speech, but they do not adequately assess whether the listener can comprehend the linguistic meaning of the message. An alternative for future studies would be the inclusion of comprehension tasks. Comprehension tasks require the listener hear the speech signal and process the information. This is more representative of everyday communication and would likely yield higher-quality information about the intelligibility of the New Zealand synthetic voices.

Hypothesis V: *The synthesised ModelTalker voices are natural-sounding to native New Zealand listeners.* The second visual analogue scale measured perceived naturalness of the synthesised voices. The results found that the mean ratings of naturalness ranged from 3.90 to 7.75 out of a possible score of 10. Only six of the voices had mean scores over the half way point of 5. Interestingly the female child's voice had a lower naturalness rating than the male child and the lowest rating of all the voices. This was a different trend compared to both intelligibility measures used in the present study which showed the female child's voice comparable to averages given to the adult voices. There was a wide standard deviation of the means for the naturalness scores for each synthesised voice. Jreige et al. (2009) included naturalness ratings where the listeners rated the naturalness of each recording on a Likert-type scale from 0 to 5, where 0 was defined as computerised and unnatural and 5 was defined as natural and human sounding. The average naturalness rating across all voices was 3.5 which is 7 if doubled to match the scale of the present study. This indicates that the VocaliD voices in Jreige et al. (2009) received higher naturalness ratings than most of the voices in the present study. It should be noted that Jreige et al. (2009) had 24 participants who heard 30 sentences for each of the eight synthesised voices, giving a larger participant sample size and sentence stimulus size compared the 15 listeners and two sentences per voice heard in the present study. In the present study, a difference was found in the naturalness ratings between the Sentence A and Sentence B stimuli. Sentence B was consistently scored lower in naturalness for each voice compared to Sentence A. Although detailed analysis was not carried out, the primary researcher noticed that all voices pronounced the word "prepared" in Sentence B with an unnatural exaggerated gap between the syllables. This word might have influenced the listeners ratings of the voice when hearing Sentence B. Measuring naturalness is more difficult than measuring intelligibility as it is subjective; every individual has their own view on what sounds natural depending on their background, age, gender, and dialect (Nerbonne, 2003). Although it is important for synthesised speech to approximate the naturalness of normal speech, for people

who are voice banking it is important for the synthesised speech to sound like them. The present study did not include any measures of similarity between the synthesised speech and the voice donors' natural speech. Future research should include similarity ratings (Jreige et al., 2009; Veaux, Yamagishi & King, 2012). Jreige et al. (2009) had listeners hear a random sample of 232 pairs of sentences, where each pair consisted of a natural recording and a synthesised voice. For each pair, listeners indicated whether the samples were produced by the same or different speakers. Across all voices the average listener accuracy was 79.5%. Veaux, Yamagishi, and King (2012) used a 5-point scale of *1=very dissimilar*, *2=dissimilar*, *3=quite similar*, *4=very similar*, and *5=identical* where participants were asked to listen to the reference voice and synthesised voice alternately saying the same sentence. The mean rating was 3 indicating participants perceived the synthetic and natural voices to be quite similar. A similarity test should be carried out for the New Zealand synthesised voices. This would provide some evaluation into the similarity of natural voices compared to synthesised voices created using ModelTalker technology, since the purpose of voice banking is to create a personalised synthetic voice which matches the vocal characteristics of the speaker (Beukelman & Miranda, 2013; Bunnell et al., 2017; Creer, 2009; ModelTalker, 2018; Nathanson, 2017).

### **7.2.2. Can unfamiliar listeners perceive the age and gender of the New Zealand-accented voices created using ModelTalker technology?**

This research question was explored through hypotheses VI and VII.

Hypothesis VI: *The synthesised ModelTalker voices portray the age of the voice donors.* Consistent with Cerrato et al. (2000) and Waller et al. (2015), the listeners in the current study tended to overestimate the age of young speakers and underestimate the age of the older speakers. The three voices aged in the 18 – 24 age group (Māori male, young male, young female) had average deviations of 11.50 years, 4.50 years, and 6.83 years above the chronological ages of the speaker, which is comparable to the deviations reported by Cerrato et al. (2000) and Waller et al. (2015) of 10 years and 4.8 years. The Māori female's voice was estimated to be younger by 7.67 years, however both Cerrato et al. (2000) and Waller et al. (2015) report overestimation of voices in this age category. The middle-aged male's voice had a mean deviation of 2.33 years older than chronological age which was also different from the trend from Cerrato et al. (2000) and Waller et al. (2015) for this age group's deviation. However, it should be noted that one standard deviation above the mean for the middle-aged male's age estimation sat above the zero mark and one standard deviation below the mean sat

below the zero mark. The middle-aged female, older female, and older male's voices had age estimations younger than the speakers' chronological ages, consistent with the trends of Cerrato et al. (2000) and Waller et al. (2015) for these age groups. The older male had a deviation of 19.50 years younger than his chronological age, which is a larger deviation compared to the reported deviations by Cerrato et al. (2000) and Waller et al. (2015) of 12 years, however still follows the trend of older voices having the greatest deviation. Waller et al. (2015) reported age estimation of younger individuals having greater accuracy than age estimation of older individuals. The average age deviation in the two children's voices were both less than one year different from their chronological ages and the most accurately estimated. One limitation of the current study's age estimation methodology is that the unfamiliar listeners were not fully responding in an open format as with the two mentioned studies, as in the current study participants were restricted to choosing an age from the list of ages provided in five-year intervals from 5 years to 70 years. Literature shows that people are more accurate at estimating ages when further cues about the speaker are present such as their face (Waller et al., 2015). It could be a strength that the young and middle-aged voices in the present study were identified to be a range of ages when hearing the voice alone. This may mean that these voices are acceptable for people of a range of ages who use SGDs, as most communication would be carried out where listeners can use visual cues such as the face of the speaker to match with the synthetic voice.

Hypothesis VII: *The synthesised ModelTalker voices portray the gender of the voice donors.* Aligning with Cerrato et al. (2000), the unfamiliar listeners were able to correctly identify the genders of most of the New Zealand synthesised voices. The only voice which was consistently identified incorrectly was the female child's voice. The majority of participants estimated that the female child was a male speaker. School-aged male and female voices often have similar fundamental frequencies which can make it more difficult to distinguish between a male and female child's voice with only an auditory prompt (Sorenson, 1989).

### **7.3. Clinical implications**

The ten New Zealand synthesised voices are available for use by SGD users in New Zealand. TalkLink provide these voices as options for people who are unable to record their own voice but wish to use a locally-accented voice on their device. The voices were given names as outlined in Table 14. The two Māori speakers' voices were given te reo Māori names. The Māori male's voice, *Mana*, translates to the concept of power and prestige. *Wairua* translates to the concept of the soul or spirit of a person and was the name given to the Māori female's

voice. The ten New Zealand synthesised voices have an interactive demo on the ModelTalker website: <https://www.modeltalker.org/demo/>

Table 14.

*Names of the New Zealand synthesised voices.*

<b>Voice donor</b>	<b>Name of New Zealand synthesised voice</b>
Male child	Sam
Female child	Sarah
Young male	Ben
Young female	Emma
Middle-aged male	Mark
Middle-aged female	Alice
Older male	Jack
Older female	Helen
Māori male	Mana
Māori female	Wairua

As recommended by Jackson et al. (2017), the use of written instruction sheets were helpful to summarise the voice banking process for the voice donors in the present study. This can be applied to clinical settings where speech-language therapists can talk over the process and give written information to the client and their family to look over and keep for future reference. Written information is vital, especially for populations such those with MND who may see many different specialists and have large amounts of information given to them at once (Costello & Della, 2017). Early education and starting voice banking early in the span of the progressive speech loss is also vital to ensure the highest quality recordings with minimal dysarthria influence (Bunnell et al., 2017). Visual aids used in the current study were also useful to act as reminders about voice banking instructions and would be appropriate for use with clinical populations who voice bank. It is also recommended that clinical populations make use of the Sennheiser PC-36 headset or other USB headsets recommended by ModelTalker. The current study had no difficulties with headset quality, unlike in Jackson et al. (2016) which used built-in laptop microphones. Speech-language therapists can guide clients when initially setting up the ModelTalker software and finding appropriate recording spaces. Ideally collaboration with facilities that have sound booths would provide options for

environments with minimal background noise and distractions, when home environments may not be as suitable (Hyppa-Martin et al., 2017). These booths would need to be accessible for clients in terms of logistics such as mobility and availability of scheduling the space. Portable sound proofing methods as suggested by Hyppa-Martin et al. (2017) could be worth pursuing if recording at home is necessary. The computer skills of the clients should be acknowledged. Having a speech-language therapist set up the login details and including the clinician's email address allows both client and speech-language therapist to be involved in communication with ModelTalker. The speech-language therapist could demonstrate the recording process and assist the client for the screening session. Because the full recording sessions have an identical process to the screening session, clients can become familiar with the process and may be able to complete the full recordings independently. The speech-language therapist can be available to aid if necessary. Equally a therapy assistant would be able to do this whole process after gaining familiarity with ModelTalker technology. The role of the speech-language therapist or therapy assistant could also include assisting motivation levels of the clients by sharing techniques used by other clients to get through the potentially daunting task of recording 1,600 sentences when one's speech may be fatigued already. By showing clients the function to preview the synthetic voice or showcasing other completed voices, the client can see how the software captures people's voices. Talking about communication options early on is vital because it gives clients, such those with MND, the chance to think over options and it provides some control over otherwise seemingly uncontrollable conditions such as MND (Costello & Dellea, 2017).

#### **7.4. Future directions**

Based on the limitations discussed in the previous sections there are three main points for future research to address. Firstly, future studies should investigate the voice banking experiences of clinical populations such as those with MND. It is expected there would be different benefits and challenges for people using ModelTalker who have conditions of progressive speech loss. Recording rates are expected to be considerably longer, as fatigue may reduce recording session durations (Costello, 2016). Motivational factors for completing the voice banking process are likely to be different to those of voice donors. It is possible that completing the recordings would be an emotional journey as people may view the process of voice banking as a type of legacy, with the creation of the synthetic voice preserving their own voice and identity as they face their upcoming loss of speech (Creer, 2009; Nathanson, 2017). Secondly, the ModelTalker software should be personalised for the New Zealand accent if many New

Zealanders are going to go through the voice banking process. With the increasing prevalence of neuro-progressive conditions over time and advances in technology meaning that SGDs are widely used with many clinical populations, personalised synthetic voices are going to continue to increase in demand (Creer et al., 2016). A New Zealand model voice should be developed in order to allow auditory cues to match the accent of the speaker. Options for New Zealand themed voice banking sentences would be a supportive addition to the advances in voice banking technologies and sentence options over time as discussed by Costello and Dellea (2017). The third direction for future research is focused on the creation of a fully synthetic te reo Māori voice. The phonemic aspects of the te reo Māori language would need to be mapped out in order for a voice banking system to recognise all the vowels and consonants that make up the language. Alterations would need to be made for ModelTalker software to recognise the phonetic significance of the macron diacritic. A fully synthetic te reo Māori voice would allow SGD users access to the language, something that is difficult at the current time as there are no systems which are in the language nor allow correct pronunciation. The te reo Māori language is important in New Zealand culture and a synthetic voice would allow SGD users to support the revitalisation of this language. The Ka Hikitia strategy is in place in the New Zealand education system (Ministry of Education, 2018). This strategy aims for all Māori students to gain the skills, qualifications, and knowledge they need to achieve educational success as Māori (Ministry of Education, 2018). Access to a synthesised voice capable of communication in te reo Māori is important to allow children and young adults who use SGDs to participate in this area of the curriculum.

### **7.5. Conclusions**

Personalised synthetic voices were effectively created using the ModelTalker technology by healthy adult and child voice donors. The voice donors found the voice banking process to be a positive experience and identified multiple strengths of the ModelTalker voice banking system, as well as a handful of challenges consistent with previous literature (Creer et al., 2009; Hyppa-Martin et al., 2017; Jackson et al., 2017). The New Zealand synthesised voices were found to have intelligibilities similar to those previously reported (Jreige et al., 2009; Von Berg et al., 2009). Age estimations followed patterns reported in the literature such as in Cerrato et al. (2000) and Waller et al. (2015). The majority of the gender predictions were accurate. The New Zealand-accented synthetic voices created by the healthy voice donors performed at acceptable rates for each measurement used. Future studies should investigate the voice banking experiences of clinical populations such as those with progressive speech loss, and

continue to work towards further personalisation of the voice banking process with the New Zealand accent and the creation of a te reo Māori synthetic voice. The voices created in this study are available for New Zealand SGD users who want to use a local accented voice on their device. With the availability of these voices, this thesis has addressed the lack of New Zealand-accented synthetic voices available for SGD users.

## References

- A-Soft Software. (2018). Speech Assistant. Retrieved February 8, 2018, from <http://www.a-soft.nl/speechassistant-tips.html>
- Acapela Group. (2017). My-own-voice. Retrieved February 8, 2018, from <http://www.acapela-group.com/voices/voice-replacement/faq-my-own-voice/>
- Acapela Group. (2018). Voices. Retrieved January 29, 2018, from <http://www.acapela-group.com/voices/>
- Anand, S., & Stepp, C. (2015). Listener Perception of Monopitch, Naturalness, and Intelligibility for Speakers With Parkinson's Disease. *Journal of Speech, Language, and Hearing Research, 58*, 1134–1144.
- Aphasia New Zealand Charitable Trust. (2010). About Aphasia. Retrieved January 26, 2018, from <http://www.aphasia.org.nz/public/about/public-about/>
- Apple. (2018). Languages supported by VoiceOver. Retrieved from <https://support.apple.com/en-us/HT206175>
- ASHA. (2011). Speech-language pathology medical review guidelines. Retrieved March 2, 2018, from <http://www.asha.org/practice/reimbursement/SLP-medical-review-guidelines/>
- ASHA. (2018). Augmentative and alternative communication: Key issues. Retrieved March 2, 2018, from [http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773&section=Key\\_Issues](http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773&section=Key_Issues)
- AssistiveWare. (2017). Proloquo2Go Voices. Retrieved January 29, 2018, from [http://www.assistiveware.com/product/proloquo2go/voices#en\\_AU](http://www.assistiveware.com/product/proloquo2go/voices#en_AU)
- Autism New Zealand. (2018). About Autism. Retrieved January 26, 2018, from [https://www.autismnz.org.nz/about\\_autism](https://www.autismnz.org.nz/about_autism)
- Bailey, R., Parette, H., Stoner, J., Angell, M., & Carroll, K. (2006). Family members' perceptions of augmentative and alternative communication device use. *Language, Speech and Hearing Services in Schools, 37*, 50–60.
- Ball, L. J., Beukelman, D., & Pattee, G. (2004). Acceptance of Augmentative and Alternative Communication Technology by Persons with Amyotrophic Lateral Sclerosis. *Augmentative and Alternative Communication, 20*(2), 113–122.
- Baxter, S., Enderby, P., Evans, P., & Judge, S. (2012). Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: A systematic

- review and qualitative synthesis. *International Journal of Language and Communication Disorders*, 47(2), 115–129.
- Bedrosian, J., Hoag, L., Calculator, S., & Molineux, B. (1992). Variables influencing perceptions of the communicative competence of an adult augmentative and alternative communication system user. *Journal of Speech and Hearing Research*, 35, 1105 – 1113.
- Beukelman, D., Fager, S., Ball, L., & Dietz, A. (2007). AAC for adults with acquired neurological conditions: a review. *Augmentative and Alternative Communication*, 23(3), 230–242.
- Beukelman, D., & Mirenda, P. (2013). *Augmentative and alternative communication: Supporting children and adults with complex communication needs* (4th ed.). Baltimore, MD: Brookes.
- Beukelman, D. R., Ball, L. J., & Fager, S. (2008). An AAC personnel framework: adults with acquired complex communication needs. *Augmentative and Alternative Communication*, 24(3), 255–267.
- Blischak, D., Lombardino, L., & Dyson, A. (2003). Use of Speech-Generating Devices: In Support of Natural Speech. *Alternative and Augmentative Communication*, 19(1), 29–35. <https://doi.org/10.1080/0743461032000056478>
- Bourgeois, M. (2007). *Memory books and other graphic cueing systems: Practical communication and memory aids for adults with dementia*. Health Professionals Press Inc. Baltimore.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Britton, D., Cleary, S., & Miller, R. (2013). What is ALS and what is the philosophy of care? *Perspectives on Swallowing Disorders*, 13, 4–11.
- Bunnell, T., Lilley, J., & McGrath, K. (2017). The ModelTalker project: A web-based voice banking pipeline for ALS/MND patients. *Interspeech*, 1–2.
- Carr, D., & Felce, J. (2007). Brief report: Increase in production of spoken words in some children with autism after PECS teaching to phase III. *Journal of Autism and Developmental Disorders*, 37(4), 780–787.
- Cerebral Palsy Society of New Zealand. (2018). Cerebral Palsy. Retrieved January 19, 2018, from [http://www.cerebralpalsy.org.nz/Category?Action=View&Category\\_id=88](http://www.cerebralpalsy.org.nz/Category?Action=View&Category_id=88)
- CereProc. (2018). CereProc Voices. Retrieved February 8, 2018, from <https://www.cereproc.com/en/products/voices>
- Cerrato, L., Falcone, M., & Paoloni, A. (2000). Subjective age estimation of telephonic

- voices. *Speech Communication*, 31(2), 107–112.
- Costello, J. (2000). AAC intervention in the intensive care unit: The Children’s Hospital Boston model. *Augmentative and Alternative Communication*, 16(3), 137–153.
- Costello, J. (2016). ALS Augmentative Communication Program. Retrieved February 8, 2018, from <http://www.childrenshospital.org/centers-and-services/als-augmentative-communication-program>
- Costello, J., & Dellea, P. (2017). ALS and AAC: Early engagement in assessment, system design and implementation. *Seminar presented at the annual convention of the American Speech-Language-Hearing Association, Los Angeles.*
- Crabtree, M., Mirenda, P., & Beukelman, D. (1990). Age and gender preferences for synthetic and natural speech. *Augmentative and Alternative Communication*, 6(4), 256–261.
- Creer, S. (2009). *Personalising synthetic voices for individuals with severe speech impairment* (unpublished doctoral thesis). University of Sheffield, Sheffield, England.
- Creer, S., Cunningham, S., Green, P., & Yamagishi, J. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language*, 27, 1178–1193.
- Creer, S., Enderby, P., Judge, S., & John, A. (2016). Prevalence of people who could benefit from augmentative and alternative communication (AAC) in the UK: determining the need. *International Journal of Language & Communication Disorders*, 51(6), 639–653.
- Creer, S., Green, P., & Cunningham, S. (2009). Voice banking. *Advances in Clinical Neuroscience and Rehabilitation*, 9(2), 16–18.
- Crystal, D. (2003). *The Cambridge encyclopedia of the English language* (2nd ed.). Cambridge University Press.
- Darley, F., Aronson, A., & Brown, J. (1975). *Motor speech disorders* (3rd ed.). Philadelphia, PA:W.B. Saunders Company.
- Das, B. (1982). Effects of Production Feedback and Standards on Worker Productivity in a Repetitive Production Task. *A I I E Transactions*, 4(1), 27–37.
- Desktop Technology Services Limited. (2018). LAMP Words For Life. Retrieved February 8, 2018, from <http://assistive.dtsl.co.nz/products/17658-lamp-words-for-life.aspx>
- Drager, K. D. R., Reichle, J., & Pinkoski, C. (2010). Synthesized Speech Output and Children: A Scoping Review. *American Journal of Speech-Language Pathology*, 19(7), 259–273.
- Duffy, S., & Pisoni, D. (1992). Comprehension of synthetic speech produced by rule: A

- review and theoretical interpretation. *Language and Speech*, (35), 351–389.
- Eriksen, G. (2018). *Comparison of voice banking protocols for creating New Zealand-accented voices for children and youth who use speech generating augmentative and alternative communication* (unpublished honours report). University of Canterbury, New Zealand.
- Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York: Harper & Row.
- Google. (2018). Google Text-to-speech. Retrieved February 8, 2018, from <https://play.google.com/store/apps/details?id=com.google.android.tts&hl=en>
- Gorenflo, C., & Gorenflo, D. (1991). The effects of information and augmentative communication technique on attitudes toward nonspeaking individuals. *Journal of Speech and Hearing Research*, 34, 19 – 34.
- Greene, B., Manous, L., & Pisoni, D. (1984). Perceptual evaluation of DECtalk: A final report of version 1.8. In *Research on speech perception: Bloomington, Speech Perception Laboratory, Psychology Department, Indiana University*.
- Harpo Software. (2018). Nuance Voices. Retrieved January 29, 2018, from <https://harposoftware.com/en/2-main/s-1/brand-nuance>
- Headway: The Brain Injury Association. (2018). Language impairment. Retrieved January 19, 2018, from <https://www.headway.org.uk/about-brain-injury/individuals/effects-of-brain-injury/communication-problems/language-impairment-aphasia/>
- Hollingworth, H. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17), 461–469.
- Howey, K. (2017). The Lived Experience of Speaking Through a Speech-Generating. *Seminar presented at the annual convention of the American Speech-Language-Hearing Association, Los Angeles*.
- Hustad, K. C. (2011). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria, *51*(3), 562–573.
- Hux, K., Knollman-Porter, K., Brown, J., & Wallace, S. (2017). Comprehension of synthetic speech and digitised natural speech by adults with aphasia. *Journal of Communication Disorders*, 69, 15–26.
- Hyppa-Martin, J., Friese, J., & Barnes, C. (2017). Voice banking for individuals with ALS: Tips and pointers for successful, low-cost voice banking. *Poster session presented at the annual convention of the American Speech-Language-Hearing Association, Los Angeles*.
- Ivona Software. (2018). Ivona Voices. Retrieved February 8, 2018, from

<https://www.ivona.com/us/about-us/voice-portfolio/>

- Jackson, S., Bode, A., Lyon, J., Beck, K., Alexander, K., & Lamb, H. (2017). Using the ModelTalker software to create a personalised synthetic voice: A hands-on student learning experience. *Poster session presented at the annual convention of the American Speech-Language-Hearing Association, Los Angeles.*
- Jackson, S., Foutch, A., Roberts, M., Duff, J., & Collins, J. (2016). Voice banking using the ModelTalker system: Graduate students' experiences of creating a synthesised voice. *Poster session presented at the annual convention of the American Speech-Language-Hearing Association, Philadelphia.*
- Jreige, C., Patel, R., & Bunnell, H. T. (2009). VocaliD: personalizing text-to-speech synthesis for individuals with severe speech impairment. *Conference on Computers and Accessibility, 259–260.*
- Keegan, P. (2018). Maori Language Information. Retrieved January 29, 2018, from [http://www.maorilanguage.info/mao\\_phon\\_desc1.html](http://www.maorilanguage.info/mao_phon_desc1.html)
- Koul, R. (2003). Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication, 19(1), 49–58.*
- Koul, R., & Allen, G. (1993). Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in the presence of noise. *Journal of Speech and Hearing Research, 36, 790 – 798.*
- Lasker, J., & Bedrosian, J. (2001). Promoting acceptance of augmentative and alternative communication by adults with acquired communication disorders. *Augmentative and Alternative Communication, 17(3), 141–153.*
- Light, J., Page, R., Curran, J., & Pitkin, L. (2007). Children's ideas for the design of AAC assistive technologies for young children with complex communication needs. *Augmentative and Alternative Communication, 23(4), 274–287.*
- Logan, J., Greene, B., & Pisoni, D. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, (86), 566–581.*
- Marthy, P., Yorkston, K., & Gutmann, M. (2000). AAC for Individuals with Amyotrophic Lateral Sclerosis. In J. Beukelman, D., Yorkston, K. Riechle (Ed.), *Augmentative Communication for Adults with Neurogenic and Neuromuscular Disabilities* (pp. 183–229). Baltimore: P.H. Brookes.
- McCall, F., Marková, I., Murphy, J., Moodie, E., & Collins, S. (1997). Perspectives on AAC systems by the users and by their communication partners. *European Journal of Disorders of Communication, 32(3), 235–256.*

- McCall, F., & Moodie, E. (1998). Training staff to support AAC users in Scotland: Current status and needs. *AAC: Augmentative and Alternative Communication*, 14(4), 228–238.
- McCord, M., & Soto, G. (2004). Perceptions of AAC: An Ethnographic Investigation of Mexican-American Families. *Augmentative and Alternative Communication*, 20(4).
- McNaughton, D., Fallon, K., Tod, J., Weiner, F., & Neisworth, J. (1994). Effect of repeated listening experiences on the intelligibility of synthesised speech. *Augmentative & Alternative Communication*, 10(3), 161–168.
- Miller, N., Noble, E., Jones, D., & Burn, D. (2006). Life with communication changes in Parkinson's disease. *Age and Ageing*, 35(3), 235–239.
- Mills, T., Bunnell, H. T., & Patel, R. (2014). Towards Personalized Speech Synthesis for Augmentative and Alternative Communication. *Augmentative and Alternative Communication*, 30(3), 226–236.
- Ministry of Education. (2018). Māori Education in New Zealand. Retrieved February 19, 2018, from <https://www.education.govt.nz/quick-links/maori/>
- Ministry of Health and Education. (2008). New Zealand Autism Spectrum Disorder Guideline. Retrieved March 2, 2018, from [www.moh.govt.nz/autismspectrumdisorder](http://www.moh.govt.nz/autismspectrumdisorder)
- Mirenda, P., & Beukelman, D. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*, (5), 84–88.
- Mirenda, P., Eicher, D., & Beukelman, D. R. (1989). Synthetic and Natural Speech Preferences of Male and Female Listeners in Four Age Groups. *Journal of Speech and Hearing Research*, 32(2), 175–183.
- ModelTalker. (2018). Creating Personal Voices For All. Retrieved March 20, 2017, from <https://www.modeltalker.org/>
- Mullennix, J., & Stern, S. (2010). Important Issues for researchers and Practitioners using Computer Synthesised Speech as an Assistive Aid. In J. Mullennix & S. Stern (Eds.), *Computer Synthesised Speech Technologies - Tools for Aiding Impairment* (pp. 1–7). IGI Global.
- Murphy, J. (2004). “I Prefer Contact This Close”: Perceptions of AAC by People with Motor Neurone Disease and their Communication Partners. *Augmentative & Alternative Communication*, 20(4), 259–271.
- Nathanson, E. (2017). Disability and Rehabilitation Native voice, self-concept and the moral case for personalized voice technology. *Disability and Rehabilitation*, 39(1), 73–81.
- National Institute for Health and Care Excellence. (2016). Motor Neuron Disease:

- Assessment and Management. Retrieved March 2, 2018, from <http://www.nice.org.uk/guidance/ng42>
- Nerbonne, J. (2003). Linguistic variation and computation. In *Proceedings of European Chapter of the Association for Computational Linguistics* (pp. 3–10).
- New Zealand Ministry for Culture and Heritage. (2009). New Zealand Speech. Retrieved January 29, 2018, from <https://teara.govt.nz/en/1966/new-zealand-speech/page-6>
- New Zealand Ministry for Culture and Heritage. (2018). Te reo Māori recognised as official language. Retrieved January 27, 2018, from <https://nzhistory.govt.nz/maori-becomes-an-official-language>
- O’Keefe, B. M., Brown, L., & Schuller, R. (1998). Identification and rankings of communication aid features by five groups. *AAC: Augmentative and Alternative Communication*, 14(1), 37–50.
- Palmer, R., Enderby, P., & Hawley, M. (2010). A voice input voice output communication aid: what do users and therapists require? *Journal of Assistive Technologies*, 4(2), 4–14.
- Park, S. Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., & Tobler, P. N. (2017). A neural link between generosity and happiness. *Nature Communications*, 8(May), 1–10.
- Parkinson’s New Zealand. (2018). What is Parkinson’s? Retrieved January 26, 2018, from <http://www.parkinsons.org.nz/what-parkinsons>
- Pennington, L. (2008). Cerebral palsy and communication. *Paediatrics and Child Health*, 18(9), 405–409.
- Powell, T. W. (2006). A comparison of English reading passages for elicitation of speech samples from clinical populations. *Clinical Linguistics & Phonetics*, 20(2), 91–97.
- Prentke Romich Company. (2018). AAC Communication Solutions. Retrieved February 8, 2018, from <https://store.prentrom.com/>
- Pullin, G., & Hennig, S. (2015). 17 Ways to Say Yes: Toward Nuanced Tone of Voice in AAC and Speech Technology. *Augmentative and Alternative Communication*, 31(2), 170–180.
- Ratcliff, A., Coughlin, S., & Lehman, M. (2002). Factors influencing ratings of speech naturalness in augmentative and alternative communication. *Augmentative and Alternative Communication*, 18(1), 11–19.
- Reynolds, M. E., Isaacs-Duvall, C., & Haddox, M. L. (2002). A Comparison of Learning Curves in Natural and Synthesized Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, 45(4), 802–810.
- Salttillo Corporation. (2018). NOVA chat. Retrieved February 8, 2018, from

- <https://saltillo.com/products>
- Schlosser, R. W. (2003). Roles of speech output in augmentative and alternative communication: Narrative review. *Augmentative & Alternative Communication, 19*(1), 5–27.
- Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior. *American Journal of Evaluation, 22*(2), 127–160.
- Shennon, P. (2016). Voice Banking Update. Retrieved March 2, 2018, from <https://www.talklink.org.nz/index.php/2016/11/21/november-2016-voice-banking-update/>
- Simon, N. G., Huynh, W., Vucic, S., Talbot, K., & Kiernan, M. C. (2015). Motor neuron disease: Current management and future prospects. *Internal Medicine Journal, 45*(10), 1005–1013.
- Smartbox Assistive Technology. (2018). Voices. Retrieved February 8, 2018, from <https://thinksmartbox.com/product/grid-3/voices/>
- Sorenson, D. N. (1989). A fundamental frequency investigation of children ages 6-10 years old. *Journal of Communication Disorders, 22*, 115–123.
- Speak For Yourself. (2015). The Ever-Evolving AAC Voice Options. Retrieved February 8, 2018, from <http://www.speakforyourself.org/uncategorized/the-ever-evolving-aac-voice-options/>
- Statistics New Zealand. (2013). 2013 Census QuickStats about culture and identity. Retrieved January 27, 2018, from <http://archive.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-culture-identity/languages.aspx>
- Stern, S. E. (2008). Computer-Synthesized Speech and Perceptions of the Social Influence of Disabled Users. *Journal of Language and Social Psychology, 27*(3), 254–265.
- Sutherland, D. E., Gillon, G. G., & Yoder, D. E. (2005). AAC use and service provision: A survey of New Zealand speech-language therapists. *Augmentative and Alternative Communication, 21*(4), 295–307.
- Therapy Box. (2018). ChatAble. Retrieved January 29, 2018, from <https://www.therapy-box.co.uk/chatable>
- Tobii Dynavox. (2016). Communicator 5. Retrieved February 8, 2018, from <http://www2.tobiidynavox.com/support/communicator-5/>
- Tobii Dynavox. (2018). Communication Devices. Retrieved February 8, 2018, from <https://www.tobiidynavox.com/products/devices/>
- TouchChat. (2018). Voices included in TouchChat. Retrieved January 29, 2018, from

- <https://touchchatapp.com/support/articles/faq/voices-included-in-touchchat>
- Veaux, C., Yamagishi, J., & King, S. (2012). Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders. *Interspeech*, 967–970.
- Venkatagiri, H. (1994). Effect of sentence length and exposure on the intelligibility of synthesised speech. *Augmentative & Alternative Communication*, 10(June), 96–104.
- VocaliD. (2018). About Us - VocaliD. Retrieved March 20, 2017, from <https://www.vocalid.co/about>
- Von Berg, S., Panorska, A., Uken, D., & Qeadan, F. (2009). DECTalk and VeriVox: intelligibility, likeability, and rate preference differences for four listener groups. *Augmentative and Alternative Communication*, 25(1), 7–18.
- Waller, S., Eriksson, M., & Sorqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology*, 6, 1–11.
- Yorkston, K., Beukelman, D., Strand, E., & Hakel, M. (2010). *Management of motor speech disorders in children and adults* (3rd ed.). Austin, TX: Pro-Ed.
- Yorkston, K., Beukelman, D., & Trice, R. (1996). *Speech Intelligibility Test for Windows*. Lincoln: Communication Disorders Software, Distributed by Institute of Rehabilitative Engineering and Science at Madonna Rehabilitation Hospital, Lincoln, NE.

### Appendix A

Department of Communication Disorders  
Telephone: +64 3 364-2987 ext 95090  
Email: [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



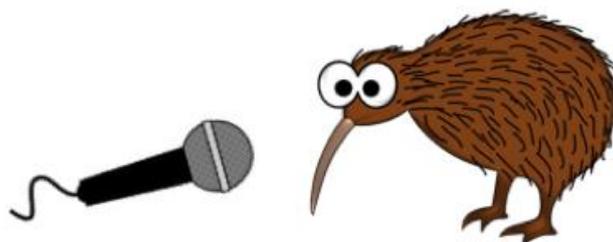
### **Creation and evaluation of ‘kiwi’ voices and the voice banking process: Information sheet for participants**

I am a speech-language therapist and master’s student researching communication disorders at the University of Canterbury. My project aims to look into how we can support adults and children who are unable to use their own voice e.g. people with cerebral palsy, motor neuron disease, Parkinson’s disease, dementia and many other conditions that reduce the ability to speak.

People who are unable to speak often use communication devices with speech-generating abilities to communicate. Currently there are no New Zealand-accented voices for people to use on their computerised devices. Many people in New Zealand use British or American accented voices, and children do not have many age appropriate voices to choose from.

Recent technology has enabled people to record their own voice before they lose the ability to speak in a process called ‘voice banking.’ This process involves a person recording many different sentences which are analysed and combined to create a ‘synthesised’ voice. The synthesised voice can include many of the personal and identifying factors that make up one’s voice, and the synthesized voice can then be used on communication devices.

My project involves healthy children and adults ‘donating’ their voices by trialing a voice banking recording software called ModelTalker. I am wanting to find out more about people’s perspectives around the process of voice banking. The donor voices will also enable creation of authentic New Zealand accented voices (“kiwi” voices) to allow greater choice for adults and children who use communication devices.



If you choose to take part in this study, your involvement in this project will begin with attending one screening session at the speech recording facility at the New Zealand Institute of Language, Brain and Behaviour (NZILBB) at the University of Canterbury. This is to familiarise you with the recording equipment, the sound-treated room environment and to create a sample recording of your voice using the ModelTalker voice banking software. This sample will be analysed by an independent Speech Language Therapist at the University of Canterbury and the ModelTalker technicians to determine selection for involvement in the full voice banking process.

If you are invited to continue, you will be asked to attend five recording sessions at the NZILBB, each approximately two hours in duration. The recording sessions will involve wearing a headset microphone and reading aloud sentences that will be displayed on a computer screen. You will be instructed to keep your speech at a constant loudness and talk at a similar speed, with the help of two dials displayed on the computer screen. Each session you will record around 400 sentences and will be asked to pause every half hour to take a drink of water. You will be encouraged to take a break anytime you require. The ModelTalker process will require you to record 1,600 sentences, however there is a possibility that some sentences may need be re-recorded following feedback from ModelTalker technicians. This is to ensure we have high quality recordings.

As a follow up to this process, once the recordings have been completed, a voice will be created by ModelTalker and you will be asked to attend a one hour follow up session. You will be asked some questions about your perspectives and experiences with the voice banking process and you will have the chance to listen and identify your own “kiwi” voice. On completion of the follow up session, you will receive a \$100 voucher as reimbursement for the approximate ten hours you will contribute.

Participation is voluntary and you have the right to withdraw without penalty. You may ask for your raw data to be destroyed. If you withdraw, I will remove information and recordings relating to you. However, you cannot withdraw once your ‘synthesised’ voice has been completed and the data has been analyzed or submitted for publication.

The results of the project may be published, but you may be assured of the complete confidentiality of data gathered in this investigation: your identity will not be made public without your prior consent. In the future, the “kiwi” voices may be available for public use for people who use speech generating devices, however, your voice will remain anonymous. ***No personal details will be associated with your voice with the exception for gender and age group.***

Because voice banking aims to create synthesized voices that have characteristics that match the vocal identity of the donor voice, aspects of your voice will be present in the created voices. The voices that may be available for public use will remain anonymous and your identification through the voice is unlikely, but may be possible in exceptional situations of chance.

To ensure anonymity and confidentiality, only myself, Grace Erisken (an honour's student) and our supervisor will have access to the identifying information. ModelTalker engineers will have access to your recordings, however all recordings will be identified with a data code. All paper copies will be stored in a locked filing cabinet in the researcher's office for the duration of the study and moved to an archive cabinet for five years before shredding. All data coded with participant number only will be stored digitally on a password protected computer in a locked office. A thesis is a public document and will be available through the UCLibrary. Please indicate to the researcher on the consent form if you would like to receive a copy of the summary of results of the project.

The project is being carried out as a requirement for a Master's project, under the supervision of Dr. Dean Sutherland, who can be contacted at [dean.sutherland@canterbury.ac.nz](mailto:dean.sutherland@canterbury.ac.nz). He will be pleased to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch ([human-ethics@canterbury.ac.nz](mailto:human-ethics@canterbury.ac.nz)).

If you agree to participate in the study, you are asked to complete the consent form and return to the researcher.

## Creation and evaluation of ‘kiwi’ voices and the voice banking process

### Consent form for participants

- I have been given a full explanation of this project and have had the opportunity to ask questions.
- I understand what is required of me if I agree to take part in the research.
- I understand that this first session is to determine selection to continue with the project. Only the voice donors chosen to continue with the voice banking process will receive the \$100 voucher on completion of the voice banking.
- I understand that participation is voluntary and I may withdraw at any time before the creation of the synthesised voice without penalty. Withdrawal of participation will also include the withdrawal of any information I have provided should this remain practically achievable.
- I understand that any information or opinions I provide will be kept confidential to the researcher and that any published or reported results will not identify the participants. I understand that a thesis is a public document and will be available through the UC Library
- I understand that all data collected for the study will be kept in in password protected electronic form and will be destroyed after 5 years.
- I understand that the “kiwi” accented voice created with my voice will be available for public use for those who use speech generating devices but no personal details will be associated with my voice **with the exception for gender and age group**. Identification though my donated voice is unlikely and only possible in exceptional situations of chance.
- I give permission for the engineers at ModelTalker to listen and work with my recordings to create the synthesised voice. I understand they will not receive any identifying information about me.
- I understand that I can contact the researcher Michelle Westley ([michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)) or supervisor Dean Sutherland ([dean.sutherland@canterbury.ac.nz](mailto:dean.sutherland@canterbury.ac.nz)) 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](mailto:human-ethics@canterbury.ac.nz))
- I would like a summary of the results of the project.
- By signing below, I agree to participate in this research project.

Name: \_\_\_\_\_ Signed: \_\_\_\_\_ Date: \_\_\_\_\_

Email address (for report of findings, if applicable): \_\_\_\_\_

*Please return this consent form to the researcher.*

## Appendix B

Department of Communication Disorders  
Telephone: +64 3 364-2987 ext 95090  
Email: [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



### Recording 'kiwi' voices Information sheet for child participants

My name is Michelle Westley and I am student at the University of Canterbury. I am looking for children and young adults to help me create recordings of their voices.

Some people cannot talk using their own voice like you and I do. They might use a computer or iPad to help them to talk to their friends and family, like the one in this picture.



The computer talks when the person touches the picture of the word they want to say. The voice that the computer uses is often an adult's voice or from another country. We want to create some voices that children can use that sound like New Zealand children and young adults.

If you want to help create a New Zealand voice for other people to use, you will need to come along to eight sessions here at our quiet room at the University of Canterbury. At our first session, we are going to record you saying ten sentences wearing a headset that goes on your head and over your ears. A speech therapist and the people who make the voices will listen to your sentences and you might be asked back to do more recordings.

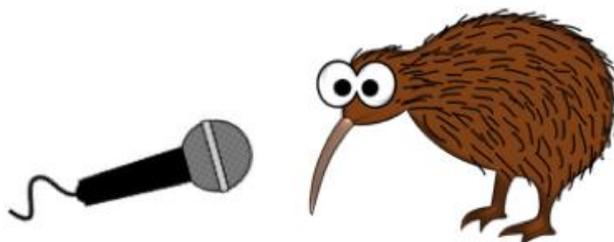


If you are asked to continue, you will have lots more sentences to say and we will record them over seven different days. You will be able to stop to have a rest or a drink of water whenever you like. To be able to create a voice we can use on people's computers, you will need to say 1,600 sentences. We might tell you to talk louder or quieter and faster and slower and we might need to repeat some of the sentences a few times. This is to make sure we get the best recordings of your voice.

When we have finished recording your voice you will be asked come back to our quiet room and tell us what you thought about when you recorded your voice. You will be able to listen and see if you can pick which voice you helped to make. At the end of this you will receive \$100 voucher to say thank you for all your help.

The information you give us will not be shared with anyone outside of our project. We want to give these voices to other children and we need to tell them how old you are and that it is a girl or a boy who is talking.

If you have any questions, ask your parents or caregivers to email Michelle Westley at [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



Department of Communication Disorders  
Telephone: +64 3 364-2987 ext 95090  
Email: [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



## **Recording 'kiwi' voices Consent form for child participants**

I would like to have my voice recorded so that other children and young adults who can not talk can use my voice.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Date of birth: \_\_\_\_\_ Ethnicity: \_\_\_\_\_

## Appendix C

Department of Communication Disorders  
Telephone: +64 3 364-2987 ext 95090  
Email: [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



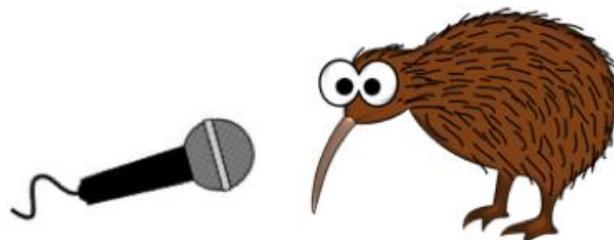
### **Creation and evaluation of ‘kiwi’ voices and the voice banking process Information sheet for listeners**

I am a speech-language therapist and master’s student researching communication disorders at the University of Canterbury. My project aims to look into how we can support adults and children who are unable to use their own voice e.g. people with cerebral palsy, motor neuron disease, Parkinson’s disease, dementia and many other conditions that reduce the ability to speak.

People who are unable to speak often use communication devices with speech-generating abilities to communicate. Currently there are no New Zealand-accented voices for people to use on their computerised devices. Many people in New Zealand use British or American accented voices, and children do not have many age appropriate voices to choose from.

Recent technology has enabled people to record their own voice before they lose the ability to speak in a process called ‘voice banking.’ This process involves a person recording many different sentences which are analysed and combined to create a ‘synthesised’ voice. The synthesised voice can include many of the personal and identifying factors that make up one’s voice, and the synthesized voice can then be used on communication devices.

The first part of my project involved healthy children and adults ‘donating’ their voices by recording their voice. I am wanting to find out more about the synthesised voices we have created by having people listen and evaluate a few factors relating to the voices.



If you choose to take part in this study, your involvement in this project will be involve one session at the University of Canterbury. The session will be approximately one hour in duration and involve listening to speech though headphones. You will be told about different measurements that relate to speech and you will be asked to rate the different “kiwi” voices. On completion of the evaluation session, you will receive a \$10 voucher.

Participation is voluntary and you have the right to withdraw at any stage without penalty. You may ask for your raw data to be destroyed at any point. If you withdraw, I will remove information and responses relating to you.

The results of the project may be published, but you may be assured of the complete confidentiality of data gathered in this investigation: your identity will not be made public without your prior consent. In the future, the “kiwi” voices you will be listening to may be available for public use for people who use speech generating devices.

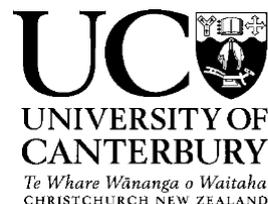
To ensure anonymity and confidentiality, only myself, Grace Erisken (an honour’s student) and our supervisor will have access to the identifying information. All paper copies will be stored in a locked filling cabinet in the researcher’s office for the duration of the study and moved to an archive cabinet for five years before shredding. All data coded with participant number only will be stored digitally on a password protected computer in a locked office. A thesis is a public document and will be available through the UCLibrary. Please indicate to the researcher on the consent form if you would like to receive a copy of the summary of results of the project.

The project is being carried out as a requirement for a Master’s project, under the supervision of Dr. Dean Sutherland, who can be contacted at [dean.sutherland@canterbury.ac.nz](mailto:dean.sutherland@canterbury.ac.nz). He will be pleased to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch ([human-ethics@canterbury.ac.nz](mailto:human-ethics@canterbury.ac.nz)).

If you agree to participate in the study, you are asked to complete the consent form and return to the researcher.

Department of Communication Disorders  
Telephone: +64 3 364-2987 ext 95090  
Email: [michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)



## Creation and evaluation of ‘kiwi’ voices and the voice banking process Consent form for listeners

- I have been given a full explanation of this project and have had the opportunity to ask questions.
- I understand what is required of me if I agree to take part in the research.
- 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 any information or opinions I provide will be kept confidential to the researcher and that any published or reported results will not identify the participants. I understand that a thesis is a public document and will be available through the UC Library
- I understand that all data collected for the study will be kept in in password protected electronic form and will be destroyed after 5 years.
- I understand that the “kiwi” accented voices that I will be listening to will be available for public use for those who use speech generating devices.
- I understand that I can contact the researcher Michelle Westley ([michelle.westley@pg.canterbury.ac.nz](mailto:michelle.westley@pg.canterbury.ac.nz)) or supervisor Dean Sutherland ([dean.sutherland@canterbury.ac.nz](mailto:dean.sutherland@canterbury.ac.nz)) 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](mailto:human-ethics@canterbury.ac.nz))
- I would like a summary of the results of the project.
- By signing below, I agree to participate in this research project.

Name: \_\_\_\_\_ Signed: \_\_\_\_\_ Date: \_\_\_\_\_

Ethnicity: \_\_\_\_\_ DOB: \_\_\_\_\_

Email address (for report of findings, if applicable):  
\_\_\_\_\_

*Please return this consent form to the researcher.*

## Appendix D

### Voice banking participant recording sample

1. Conversation about weekend
2. *My Grandfather* passage

You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

3. Example sentences from ModelTalker

I say this to my friends.

The grizzled old fellow could see on one side.

He couldn't even find a place.

Behind them was the dark forest.

He had rides in the wheelbarrow.

I play with my baby sister.

The whole circle was agitated.

Is he made of tin or stuffed? asked the Lion.

The wolves surged to meet him.

There's another way you can get a tooth out.

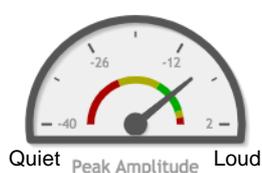
Jo's gentlemanly demeanor amused and set him at his ease.

The forefront of the pack was a large grey wolf.

## Appendix E

### Poster and handout for voice banking participants

#### *Voice banking reminders*



Aim for two green bars, ignore pronunciation

Speak in an everyday voice and focus on saying each of the sounds in the word

The sentences don't need to make sense

Re-record if you make a mistake

Remember to take breaks and drinks whenever you need

### Reminders for voice donors

- 1) Listen to the phrase that the model voice says. When you are ready click the record button and speak the sentence aloud. The computer will stop recording when you are quiet.
- 2) You can click the repeat arrow to hear the model voice say the sentence again
- 3) Make sure you talk in a consistent, almost boring, speaking style. Try not to pause between words. You should NOT make an effort to make the phrases sound interesting or meaningful. The sentences are all about the speech sounds they contain and not about their meaning!
- 4) Ignore the pronunciation checker as this is for American English.
- 5) Focus on the left and the middle dials. The left is showing how loud you are speaking and the middle one shows how fast you are talking. Keep both of these dials in the green on the meters on the screen.
- 6) Let us know if you mix up words or make a mistake, it is easy to re-record any sentence
- 7) If you need a break, take one! It can be tiring to speak for long periods of time and regular breaks will create a better quality recording. The software will automatically save where you are up to so you can pause when you need to.
- 8) Take regular sips of water throughout the session to keep your voice quality high. We will make sure you have a break and a drink every half an hour, but take more breaks if you want
- 9) If you are unsure of anything please don't hesitate to ask the researcher with you.



Peak Amplitude



Slow - Speech Rate - Fast



Pronunciation

# 1347

Why not? she replied.



Click to hear the sentence again



Click to record

## Appendix F

### 20 te reo Māori custom inventory words and embedded sentences

- |             |               |             |
|-------------|---------------|-------------|
| 1. Maori    | 8. Marae      | 15. Morena  |
| 2. Pakeha   | 9. Iwi        | 16. Tapu    |
| 3. Kia ora  | 10. Kai       | 17. Ka Kite |
| 4. Ka pai   | 11. Hangi     | 18. Wahine  |
| 5. Aroha    | 12. Haka      | 19. Tane    |
| 6. Aotearoa | 13. Tamariki  | 20. Powhiri |
| 7. Whanau   | 14. Haere Mai |             |

Maori means native people from Aotearoa

Pakeha means non Maori

Kia ora means hello

Ka pai means well done

Aroha means love

Aotearoa means New Zealand

Whanau means family

Marae means meeting place

Iwi means extended family

Kai means food

Hangi means method of cooking

Haka means war dance

Tamariki means children

Haere mai means welcome

Morena means good morning

Tapu means sacred

Ka kite means goodbye

Wahine means female

Tane means male

Powhiri welcoming ceremony

Materiki means Maori new year

I am pakeha from birth

He said kia ora before he ate kai

Well done, ka pai for finishing your work

She showed aroha to her extended iwi

I live in Aotearoa with my family

We have a big whanau to feed

The closest marae could host you

I'd like to meet iwi from the South Island

I like to eat kai during class

I like hangi food

His haka performance impressed whanau

The youngest tamariki started preschool

The teacher said haere mai to the youngest tamariki

The teacher said morena to the late wahine

I want to visit Tapu bay

He said ka kite to his friend Aroha

My dog is wahine, but my cat is tane

My friend Tane likes to play rugby

We will have a welcome powhiri tomorrow

## Appendix G

### 65 te reo Māori core vocabulary with part-of-speech information and English translations

Te reo Māori word	Part-of-speech information	English translation
awhina	verb	to help
he aha	wh question	what?
ko wai	wh question	who?
haere	verb	to go
ahea	wh question	when?
whakaaro	noun	idea
etahi	determiner	some
he	verb	to be wrong (oops)
katoa	modifier	all
tena	determiner	that (near or connected with the listener)
tera	determiner	that (away from or unconnected with both the speaker and listener)
haere mai	interjection	come here!
mahi	verb	to work/do
kao	interjection	no
kaua	negative	don't
homai	verb	to give (towards the speaker)
hoatu	verb	to give (away from the speaker)
huakina	verb	to open
katia	verb	to close
a muri	particle	in the future (later)
ano	particle	again
pehea	wh question	how?
e hia	numeral	how many?
etahi atu	determiner	some others (more)
tenei	determiner	this (near or connected to the speaker)
huri	verb	to turn around
inu	verb or noun	to drink or a drink
kai	verb or noun	to eat or food
korero	verb or noun	to say or speech
hanga	verb	to make
inaianei	particle	right now
rite	verb	to be ready
harikoa	noun	happiness
iti	adjective	small
kua pau	verb	to finish

mutu	verb	to stop
noho	verb	to sit
rata	verb	to like
kei hea	wh question	where?
ake	preposition	up
makariri	adjective	cold
ma	adjective	clean
mauiui	adjective	sick
nui	adjective	big
taihoa	interjection	wait
au	pronoun	I
ahau	pronoun	me
koe	pronoun	you
atu	particle	away
orite	verb	to be like (same)
pai	adjective	good
paru	adjective	dirty
tiaki	verb	to look after/protect
titiro	verb	to look
ratou	pronoun	them (three or more)
taku	determiner	my
raro	preposition	down
pouri	adjective	sad
rereke	adjective	different
tapu	adjective	sacred
whakarongo	verb	to listen
tatou	pronoun	we/us (three or more)
taua	pronoun	we/us (two people)
kino	adjective	bad
wera	adjective	hot

## 65 te reo Māori core vocabulary embedded sentences

Awhina means to help and he aha is what

Ko wai means who and ahea is when

Haere means to go and whakaaro is idea

Etahi means some and he is oops

Katoa means all and tena is that

Tera means that and haere mai is come here

Mahi means work and kao is no

Kaua means don't and homai is give

Hoatu means to give and huakina is open

Katia means to close and a muri is future

Ano means again and pehea is how

E hia means how many and etahi atu is more

Tenei means this and huri is turn

Inu means drink and kai is food

Korero means speech and hanga is make

Inaianei means right now and rite is ready

Harikoa means happy and iti is small

Kua pau means finish and mutu is stop

Noho means sit and rata is like

Kei hea means where and ake is up

Makariri means cold and ma is clean

Mauiuui means sick and nui is big

Taihoa means wait and au is I

Ahau means me and koe is you

Atu means away and orite is same

Pai means good and paru is dirty

Tiaki means protect and titiro is look

Ratou means them and taku is my

Raro means down and pouri is sad

Rereke means different and tapu is sacred

Whakarongo means listen and tatou is us

Taua means us and kino is bad

Wera means hot and awhina is help

He aha means what and ko wai is who

Ahea means when and haere is to go

Whakaaro means idea and etahi is some

He means oops and katoa is all

Tena means that and tera is that

Haere mai means come here and mahi is work

Kao means no and kaua is don't

Homai means to give and hoatu is give

Huakina means to open and katia is close

A muri means in the future and ano is again

Pehea means how and e hia is how many

Etahi atu means more and tenei is this

Huri means turn and inu is drink

Kai means food and korero is speech

Hanga means make and inaianei is right now

Rite means ready and harikoa is happy

Iti means small and kua pau is finish

Mutu means stop and noho is sit

Rata means like and kei hea is where

Ake means up and makariri is cold

Ma means clean and mauiuui is sick

Nui means big and taihoa is wait

Au means I and ahau is me

Koe means you and atu is away

Orite means same and pai is good

Paru means dirty and tiaki is protect

Titiro means look and ratou is them  
Taku means my and raro is down  
Pouri means sad and rereke is different  
Tapu means sacred and whakarongo is  
listen  
Tatou means us and taua is us  
Kino means bad and wera is hot  
Mana said awhina  
Wairua said he aha  
Mana said ko wai  
Wairua said ahea  
Mana said haere  
Wairua said whakaaro  
Mana said etahi  
Wairua said he  
Mana said katoa  
Wairua said tena  
Mana said tera  
Wairua said haere mai  
Mana said mahi  
Wairua said kao  
Mana said kaua  
Wairua said homai  
Mana said hoatu  
Wairua said huakina  
Mana said katia  
Wairua said a muri  
Mana said ano  
Wairua said pehea  
Mana said e hia  
Wairua said etahi atu  
Mana said tenei  
Wairua said huri  
Mana said inu

Wairua said kai  
Mana said korero  
Wairua said hanga  
Mana said inaianei  
Wairua said rite  
Mana said harikoa  
Wairua said iti  
Mana said kua pau  
Wairua said mutu  
Mana said noho  
Wairua said rata  
Mana said kei hea  
Wairua said ake  
Mana said makariri  
Wairua said ma  
Mana said mauui  
Wairua said nui  
Mana said tahoā  
Wairua said au  
Mana said ahau  
Wairua said koe  
Mana said atu  
Wairua said orite  
Mana said pai  
Wairua said paru  
Mana said tiaki  
Wairua said titiro  
Mana said ratou  
Wairua said taku  
Mana said raro  
Wairua said pouri  
Mana said rereke  
Wairua said tapu  
Mana said whakarongo

Wairua said tatou  
Mana said taua  
Wairua said kino  
Mana said wera  
Wairua said awhina

Mana said he aha  
Wairua said ko wai  
Mana said ahea  
Wairua said haere

## Appendix H

### Voice banking participant questionnaire

1. What did you know about voice banking before participating in this study?
2. Tell us about your experience of the voice banking process
3. Please rate your overall experience of the voice banking process

1-Negative    2    3    4    5    6    7    8    9    10-Positive  
                                   

4. What did you like best about voice banking and why?
5. What did you like least about voice banking and why?
6. How could voice banking be improved?

7. Please rate the ease of the voice banking process; explain you rating

1-Difficult    2    3    4    5    6    7    8    9    10- Easy  
                                   

8. Please rate the types of sentences that you were required to record (please ignore the te reo Māori recordings for this question); explain your rating

1-Complex    2    3    4    5    6    7    8    9    10- Simple  
                                   

9. Please rate the process of recording the te re Māori words; explain your rating

1-Complex    2    3    4    5    6    7    8    9    10- Simple  
                                   

10. Please rate the overall time taken to complete the voice banking process; explain your rating

1-Little time    2    3    4    5    6    7    8    9    10-Long time  
                                   

11. Please rate the model voice which read aloud the sentences; explain your rating

1-Unhelpful    2    3    4    5    6    7    8    9    10-Helpful  
                                   

12. Please rate the volume and speed dials; explain your rating

1-Useless    2    3    4    5    6    7    8    9    10-Useful

13. Please rate the option to listen to the preview of your synthesised voice; explain your rating

1-Didn't like to hear the preview voice	2	3	4	5	6	7	8	9	10-Liked to hear the preview voice
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

14. Please rate the preview of your voice near the START of the recordings; explain your rating

1-Didn't sound anything like me	2	3	4	5	6	7	8	9	10-Sounded exactly like me
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

15. Please rate the preview of your voice near the END of the recordings; explain your rating

1-Didn't sound anything like me	2	3	4	5	6	7	8	9	10-Sounded exactly like me
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

16. Would you recommended that other people donate their voice? Please explain your rating

Definitely yes	Probably yes	Unsure	Probably not	Definitely not
<input type="radio"/>				

17. Would you recommend voice banking to someone in the process of losing their ability to talk? Please explain your rating

Definitely yes	Probably yes	Unsure	Probably not	Definitely not
<input type="radio"/>				

18. What do you see as any barriers that could prevent people from voice banking? Please explain your rating

19. How successful would voice banking be if we recorded at your home? Please explain your rating

1-Not at all	2	3	4	5	6	7	8	9	10-Very successful
<input type="radio"/>									

20. How successful would voice banking be in you completed the recordings on your own? Please explain your rating

1-Not at all	2	3	4	5	6	7	8	9	10-Very successful
<input type="radio"/>									

21. Finally, is there anything else you would like to share?

## **Appendix I**

### **Auditory sentence stimuli**

#### **Trial Speech Intelligibility Test sentences (primary researcher's synthesised voice)**

You and your family will share experiences that will last a lifetime

There is nothing in this world he cannot do

#### **Speech Intelligibility Test sentences**

*Male child's voice (Group A)*

He took heart and played.

Cover and chill until serving time.

People may find them hard to digest.

None of the stops is in a city.

There is a small additional charge for other attractions.

At certain times, I like being strong for someone else.

The ideal city is a place of crowds, not of highways.

I have seen him modify his own opinions often, because he listens.

Yet, when the flowers do come, the leaves fall and are not replaced.

Why, these are the same questions that you asked when I was a student.

We are still surrounded by mysteries here on Earth which at present cannot be explained.

*Young male's voice (Group A)*

They'll never fit me again.

He erased his mistake and continued.

A large coffee machine was bubbling away.

You have to expect a few bad calls.

Your house tells very much about how you live.

He was in no way prepared for what might happen.

The crew passed out large envelopes of coloring books and toys.

A minister told his congregation that there are seven hundred different sins.

Sir, when my form is filled out, what do I do with it?

Little children came up the ramp, most of them crying and trying to leave.

Successful community gardens start when somebody has a good idea and follows through on it.

*Young female's voice (Group A)*

That wasn't me you heard.

Look for pockets of black sand.

Paper cups of jellybeans were set out.

There are two methods for soaking dried beans.

Afterwards, his view hinted at a more mature attitude.

We were working with that thing for over two weeks.

People see him as a spoiled brat in a gentlemen's sport

He will regard his wife and family as full partners and friends.

Two years ago, he heard encouraging news about a couple of old friends.

Expert car trackers are able to turn up a fair amount of untouched stock.

The inn is an unusual retreat built into the ruins on a long, black beach.

*Middle-aged female's voice (Group A)*

It's the way it was.

How many hours have you worked?

I gave the role everything I had.

It grows faster than anything else on earth.

It gave her an excuse not to leave us.

The patient managed to fall and break his ankle again.

This can be the cheapest way to ship them long distance.

The baby was fed, dressed, and put to sleep in his crib.

For one moment, it seemed as though the mother bear might charge again.

She hung a metal pie plate and a large metal spoon at the door.

They proceeded past the large group to the house and forced their way into it.

*Older male's voice (Group A)*

I've had one unusual request.

You want him to do well.

You're betting that he won't get hurt.

Night after night, they received annoying phone calls.

I didn't think it would be any big job.

You have walked all this distance just to see me?

The guide at the lodge has a supply of trail markers.

Most people find that each passing year leaves skin a little drier.  
A clerk reads the zip code and punches three numbers on his keyboard.  
Just as you outgrew making mud pies, you have outgrown wearing those old clothes.  
If you own gear, almost any time of the year can be good for hiking.

*Female child's voice (Group B)*

He has played very well.  
The dolphins swam around our boat.  
Beans must be soaked and then cooked.  
The process works in other ways as well.  
There is nothing in this world he cannot do.  
You may also recognize the things that are not practical.  
The facility is open to the public, and you may visit.  
I have seen him modify his own opinions often, because he listens.  
If you count the fun I've had, I'm well ahead of my goals.  
Labor costs account for only a small portion of what consumers pay for fruit.  
It was the exact same feeling you get when your knee gives out on you.

*Māori male's voice (Group B)*

I enjoy him very much.  
The wallpaper is green and blue.  
Big muscles are not necessarily strong ones.  
Don't let me know how you do it.  
Eventually, of course, we all got used to it.  
So far, only a handful of wells have been drilled.  
The disaster was only prevented because we had received advanced warning.  
If they are having a difficult time, he does not feel neglected.  
He told the jury to retire and come back with a lawful verdict.  
The portrait of me you painted looks more like me with each passing year.  
No one had to tell what organized labor could do for working men and women.

*Māori female's voice (Group B)*

You are paying their salary.

I mailed the package years ago.

Be prepared for odd behavior from friends.

Leave the rest on the table for later.

Pride can be used to beat down simpler vices.

Most studies of animal behavior do not support this view.

Whatever the reasons, the day of the old car is here.

You and your family will share experiences that will last a lifetime.

Last night, we all went to a music festival they had across town.

The idea of the magic wand may well have begun with the divining rod.

Give your friend a bunch of brightly colored balloons tied with a big, red bow.

*Middle-aged male's voice (Group B)*

The sun died at night.

Cover and chill until serving time.

That's what life is really all about.

The wait for work can be very long.

Today they offer to help amateur gardeners as well.

This is a trend we simply cannot allow to continue.

In an area uncrowded by other peaks, it seems even higher.

Most people find that each passing year leaves skin a little drier.

Growers, for the most part, were united this time in resisting the demands.

I like it, and I like the independence and courage it has given me.

Above all, when she receives a compliment from friends or family, she jots it down.

*Older female's voice (Group B)*

That wasn't me you heard.

Just don't fill them too full.

They're kept out on an open patio.

None of the stops is in a city.

The new kitchen shelves were mounted to the wall.

Things obviously were less tense there than she had pictured.

In that certainty lies a great peace and a great joy.

I figured it worth a try, since I wouldn't have another chance.

A person is more likely to think about job changes when he's unhappy.

Now, I go into the children's section to make sure that I'm fully informed.

If you own gear, almost any time of the year can be good for hiking.

**Intelligibility visual analogue scale (all voices)**

Leave the rest on the table for later.

He was in no way prepared for what might happen.

**Naturalness visual analogue scale (all voices)**

The new kitchen shelves were mounted to the wall.

The facility is open to the public, and you may visit.

**Age and gender estimations (all voices)**

You and your family will share experiences that will last a lifetime.

Sir, when my form is filled out, what do I do with it?