Development of Bisyllabic Speech Audiometry Word Lists for Adult Malay Speakers

PhD Thesis

Marina L. Alisaputri

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Allied Health Sciences

De Montfort University

June 2016

Declaration

I declare that this is my own original work and that all sources used have been cited.

Name: Marina L. Alisaputri

Signed:

Date:

ABSTRACT

Standardised speech audiometry material is essential in assessing hearing for speech; however, material in Malay language, particularly for speech reception threshold test, is limited and not thoroughly validated. This thesis examines the development of standardised, phonemically-balanced bisyllabic Malay speech reception threshold (SRT) test word lists for Malay-speaking adults. The effect of having a mixture of familiar and nonsense words on speech recognition is also explored. The processes of developing the word lists include selecting and compiling the words using content analysis research method, testing for homogeneity and consistency and validating the acoustic content, both using correlational research method, and assessing the clinical validity using concurrent validity method. The familiar words were selected from a corpus of familiar words extracted from daily newspapers while the nonsense words were formed based on linguistic properties of Malay. The preliminary set consisted of fifteen lists with 10 familiar words and 5 nonsense words in each. The analyses of the findings show consistency of speech discrimination using the word lists using Friedman test to have statistically no significant difference in correct scores achieved using any of the word lists, $X^2 = 19.584$, $p>0.05$. Homogeneity test for all lists using Cronbach's alpha showed a value of 0.78, indicating a strong agreement and good homogeneity among the lists. When five lists with inter-item correlation ≤0.8 were excluded from the homogeneity analysis, the alpha value for the remaining 10 lists increased to 0.88. Consistency analysis of acoustic content using repeated measures ANOVA showed no significant difference between the list and the LTASS, $F=1.229$, $p>0.05$. All 15 lists were then tested for clinical validity. Two versions of list content were assessed, an all-words version (AWL) containing all 15 words each list, and a meaningful-words only version (MWL) containing 10 meaningful words for each list. Correlation analyses between half peak level (HPL) of the speech recognition curve and pure tone (PT) thresholds showed that, in consideration of both normal hearing and hearing impaired listeners, the HPL correlated best with PT average of 250, 500, 1000, 2000 and 4000 Hz for both AWL ($r = 0.67$ to 0.95) and MWL ($r = 0.65$ to 0.95). A comparison between HPL and PT average of 250, 500, 1000, 2000 and 4000 Hz showed mean differences of 4 dB (SD = 3) and 3 dB (SD = 4) with the range of tolerance (95% confidence) of ±7 dB and ±8 dB for AWL and MWL respectively. Sensitivity, specificity, and positive and negative predictive values, when set at tolerance level of ±10 dB, were mostly >0.90 for normal hearing and hearing loss listeners using either versions. It was concluded that the addition of

nonsense words does not significantly affect SRT. The correlation between the SRT obtained using the bisyllabic Malay word lists and the PT thresholds suggested that the word lists were robust enough to be used in assessing speech hearing clinically. In conclusion, the current study has achieved to develop and produce a standardised, phonemically balanced bisyllabic Malay speech audiometry (BMSA) word lists for assessing speech reception threshold and discrimination in adult Malay speakers.

ACKNOWLEDGEMENT

Alhamdulillah, all the praises and thanks be to Allah for the strength and His blessing in completing this thesis.

I would like to express my gratitude to my supervisors, Professor Lixian Jin and Dr. Qin Xu, whose expertise, understanding, and patience, contributed greatly to the success of this research.

I would like to express my appreciation to the Dean, Kulliyyah of Allied Health Sciences, Dr. Wan Azdie b. Mohd. Abu Bakar, and also to the Head, Department of Audiology and Speech-Language Pathology, Dr. Noraidah bt. Ismail, for their tremendous support and help towards my postgraduate affairs. My acknowledgement also goes to all audiologists at the IIUM Hearing and Speech Clinic and Hospital Tengku Ampuan Afzan as well as the staff of Kulliyyah of Allied Health Sciences for their assistance, especially during data collection.

I would also like to thank my parents, Lamri b. Ali and Maimunah bt Ibrahim, and my siblings and sisters-in-law, Shamsul, Syaiful, Sarina, Bahrin, Ika and Wan, for the immense love and support they provided me, without which I would not have finished this thesis. Special thanks to my friends, Marul, Daus and Nina, for their constant support and encouragement.

This research would not have been possible without the financial assistance of the Ministry of Higher Education, Malaysia and the International Islamic University of Malaysia; I express my gratitude to these agencies.

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABR	Auditory brainstem response
AWL	All-word list
CHL	Conductive hearing loss
dB HL	decibel hearing level
HPL	Half peak level
LTASS	Long-term average speech spectrum
MHL	Mixed hearing loss
MSRS	Maximum speech recognition score
MWL	Meaningful-words list
OAE	Otoacoustic emissions
PI function	Performance-intensity function
PTA	Pure tone threshold averages
SNHL	Sensorineural hearing loss
SRT	Speech reception threshold

CHAPTER 1 INTRODUCTION

1.1 Overview

Speech audiometry is a group of hearing tests that use speech as stimuli. As much as pure tone audiometry provides the information regarding the level and type of hearing loss, information on how the speech is heard is limited. Speech stimuli are thought to provide closer representation to the speech used in daily conversations. By using speech as stimuli, inferences can be made on how the hearing loss affects the hearing for speech.

The motivation behind this research is the need to produce a standardised speech audiometry material that can assess speech intelligibility and speech discrimination in Malay-speaking adults. There are several speech audiometry materials that have been developed in Malay language; among them Malay Hearing In Noise Test (MyHINT) (Quar et al., 2008), Malay Speech Intelligibility Test (MSIT) (Yusof et al., 2013), disyllabic Malay word lists (Hong, 1984) and Malay speech audiometry (Mukari and Said, 1991). MyHINT utilises sentences as stimuli and is intended to measure hearing in noise (Quar et al., ibid.). MSIT, which is aimed to assess speech intelligibility in children, uses nonsense syllables as test items (Yusof et al., 2013). Both sets of word lists developed by Hong (ibid.) and Mukari and Said (ibid.) made use of bisyllabic words in their material; however, the weakness of the sets was that there was no well-defined verification and validation of the test items in assessing speech hearing. Thorough verification and validation of the word lists ensure that the test material is standardised and fit to be used in the practice.

The type of speech that is used as stimuli may range from a simpler form of speech, such as syllables, to a more complex structure, such as sentences. The type of speech chosen is dependent on the test objective; shorter speech structures such as syllables and words are mainly used to assess threshold of intelligibility for speech and speech discrimination abilities (Egan, 1948; Hudgins et al., 1947, Hirsh et al., 1952, Boothroyd, 1968). On the other hand, assessments of the more complex aspects of speech hearing, for example, speech processing skills, linguistic-situational skills and hearing in noise, usually utilise longer forms of speech, such as sentences (Kalikow, et al., 1977; Bochner, et al., 1986; Nilsson et al., 1994).

Speech reception threshold (SRT) tests are tests that are used to determine the lowest level of speech at which the listener can hear and recognise it. Examples of recognition of speech are correct repetition of the speech stimuli and correct pointing at pictures or objects that represent the speech stimuli. Katz (2015) listed several clinical uses of SRT tests: measuring communication disability, particularly speech intelligibility and speech discrimination, cross-checking of pure tone thresholds, and as a reference point for further speech tests. Typically, SRT tests employ spondees and bisyllabic words (Mendel and Danhauer, 1997); however, since the development of speech reception threshold test materials in other languages, the word structures now vary from monosyllables to trisyllables, depending on the linguistic properties of the language.

It is highly recommended that the material for speech recognition testing to be familiar to the listener (Hudgins, et al., 1947; Hirsh, et al., 1952; Webster, 1972); therefore, it is best that SRT tests to be employed in the listener's native language, even for a bilingual person (Hapsburg, et al., 2004). Due to that reason, SRT test materials can be found in many languages, such as Arabic, Russian, Swedish, Mandarin, Cantonese and Malay (Alusi et al., 1974; Ashoor and Prochazka, 1982; Hong,1984; Lau and So, 1988; Mukari and Said, 1991; Magnusson, 1995; Nissen et al., 2005a; Harris et al., 2007; Nissen et al., 2007; Wang et al., 2007; Han et al., 2009).

There are two SRT test materials previously developed in Malay (Hong, 1984; Mukari and Said, 1991). There are many similarities between both sets; both materials used bisyllables for their test items, both employed phonetic balance for their words, and the method for testing is similar. However, there are several weaknesses in the sets of word lists, among them the contents of the word lists, the scoring system and the verification processes, which can be improved on. These SRT test materials will be further reviewed in the following chapters.

The reason for using familiar words in SRT test materials is to increase the homogeneity of the test items (Webster, 1972). Highly familiar words, normally judged through their frequency of occurrence in daily speech or through rating by native speakers of the language, ensures that the test items produce comparable results within and between listeners. However, there is a limitation to how the familiar words represent normal conversation; normal conversation contains words with a much wider range of familiarity compared to the test items in SRT test word lists. On the other end of the familiarity spectrum, there are speech audiometry materials that are based on nonsense words (Levitt and Resnick, 1978; Gelfand, et al., 1992; Cheesman and Jamieson, 1996).

However, the aim of these tests is more directed towards phoneme identification rather than speech intelligibility and/or discrimination.

This thesis explores the feasibility of using a mix of familiar and nonsense words as test items for SRT test. Furthermore, the thesis attempts to improve on the previously developed Malay SRT test word lists, particularly on the phonetic balance as well as the verification and validation processes.

There are three main parts in this study; development of the word lists, and verification and clinical validation of the SRT test material, all in chronological order. The thesis is therefore arranged according to the same order with each section of the study presented in a chapter.

1.2 Thesis outline

This thesis is divided into six chapters, including the current first chapter on Introductions. The chapters in the thesis are arranged chronologically (Figure 1.1). This section outlines the chapters in the thesis.

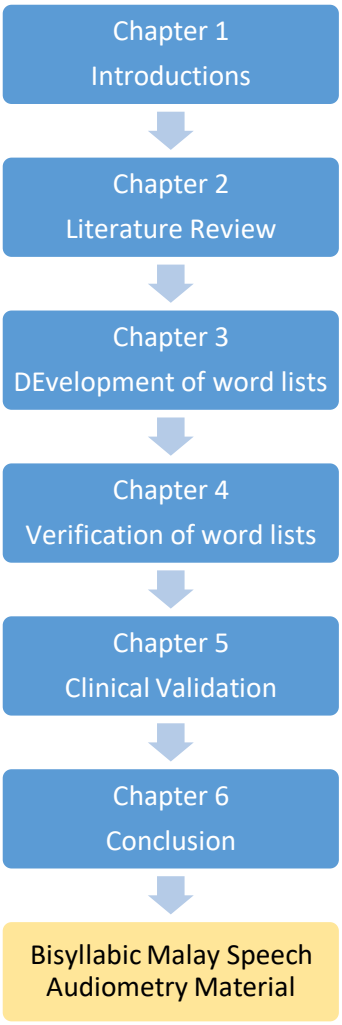


Figure 1.1 Thesis chapters based on the flow of study

Chapter Two, Literature Review, presents the reviews on past studies regarding SRT tests and the various factors that were considered in the development of the SRT test materials. The chapter also includes discussions on the issues related to the hearing system and its assessment and the linguistic properties of Malay language, particularly the phonetics, phonology and word structure. The gaps in knowledge as well as the research questions, aims and objectives of the study are also presented in the chapter.

Chapter Three presents the first part of the study, which was the development of the bisyllabic Malay word lists. The process of compiling the word lists that would be used in the following sections of the study is discussed in this chapter. Therefore, this chapter contains the review of methods used in past studies, research design, research findings and discussions related to the development of the word lists.

Chapter Four presents the verification process of the bisyllabic Malay word lists. The process includes an assessment of acoustic homogeneity through the speech acoustic spectrum of the lists, as well as analysis of variance (ANOVA) and internal consistency assessment through intraclass correlation coefficient based on the listeners' performance during the SRT test. Similar to Chapter Three, this chapter contains the review of methods, research design, research findings and discussions related to the process.

Chapter Five presents the clinical validation process for the bisyllabic Malay word lists. This section of the study involves utilising the bisyllabic Malay word lists for SRT tests on normal hearing and hearing impaired listeners. The clinical validity is assessed using the performance-intensity (P-I) function, particularly the SRT and the maximum speech recognition score (MSRS). The results of several select cases of hearing loss are discussed in detail. The predictive analyses based on the comparison between the SRT and the pure tone threshold are also presented.

Chapter Six consists of the conclusion, theoretical implications, research outcome, future research and limitations of the research. To accompany the conclusion, the actual research outcome in the form of prototype test kit is also included in Appendix I.

CHAPTER 2 LITERATURE REVIEW

This study aims to develop a clinically valid speech audiometry material to test the hearing for speech and assess speech recognition in adult Malay speakers. This chapter discusses and reviews the background of hearing impairment and its effects on hearing for speech, types of speech hearing assessments available in the market, particularly in Malay language, and the general approach to developing a speech audiometry material. The chapter also discusses the structure of Malay phonetics, phonology and word structure and its relation to the development of speech audiometry material.

2.1 Hearing loss

Hearing impairment or hearing loss is defined as a condition that causes a person to be unable to hear as well as someone with a normal hearing (WHO, 2015). According to the World Health Organisation (2015), over 5% of the world's population has hearing loss, with higher prevalence in adults than in children. Prevalence studies done in several Southeast Asian countries recorded prevalence of hearing loss between 13.6% to 19.5%, much higher compared to the global average (Prasansuk, 2000; Stevens, et al., 2013).

Hearing loss is mainly categorised according to the level of loss and type of loss, and to a lesser degree, the configuration. The effects of hearing loss on communication are largely dependent on the level and type of loss experienced by the listener. The following sub-sections discuss the different levels and types of hearing loss, and how they affect the hearing. The levels and types of hearing loss are relevant to the study as they present different effects on the performance in speech reception threshold test and, thus, contribute to the diagnostic capability of the test.

2.1.1 Hearing loss in Malaysia

Malaysia is located in south-east Asia, north of the Equator. Geographically, the country constitutes of Peninsular Malaysia, which is a part of mainland Asia, and East Malaysia, formed by the states Sabah and Sarawak in the island of Borneo. These two parts are separated by the South China Sea. Neighbouring countries include Singapore, Thailand, Brunei and Indonesia.

Figure 2.1 Map of Malaysia (Google, 2016)

Malaysia's population is 31.7 million with 28.4 million citizens and 3.3 non-citizens (Department of Statistics Malaysia, 2016). "Bumiputeras" or native-born citizens make up the majority of the population (68.6%), followed by the Chinese (23.4%), Indians (7.0%) and others (1%). Bumiputeras include the Malays, the "Orang Asli" (aborigines) and other indigenous natives of Peninsular Malaysia, Sabah and Sarawak.

According to a nation-wide hearing and ear disorder survey done on 7041 subjects in 2005, the prevalence of hearing impairment in Malaysia is 17.1%, with no significant difference between those living in urban areas and the rural population (Ministry of Health Malaysia, 2007). A lower prevalence was found among the Malays (15.7%) compared to those of Chinese descent (21.0%). It was also noted that the prevalence was higher among men and among the elderly.

A study on hearing loss registries from several government hospitals in Malaysia showed 81.1% of the 1341 patients registered as having hearing loss in the year 2010 and 2011 were adults above the age of 20 (National ORL Registry, 2013). Interestingly, 69.5% of these patients were Malays; conflicting with the lower prevalence of hearing loss found in the 2005 survey. An earlier prevalence study on 1307 primary school students on the east coast of Peninsular Malaysia showed 5.81% of the participants failed the hearing screening test (Elango, et al., 1991), slightly higher than the prevalence estimate of hearing loss given by World Health Organisation (2015).

2.1.2 Levels of hearing loss

British Society of Audiology (BSA) (2011) guideline on audiometric descriptors suggested four levels of hearing loss – mild, moderate, severe and profound. Classification is based on the average hearing threshold levels (HTLs) through pure tone audiometry at frequencies 250, 500, 1000, 2000 and 4000 Hz. Similar classification of hearing loss is also used in other countries around the world (American Speech-Language-Hearing Association, 2011; Stevens et al., 2013). Classification of the level of hearing loss in Malaysia is slightly different from the one devised by BSA. The HTL ranges for mild and moderate hearing losses are the same; however, the ranges for severe and profound hearing losses are 71 to 90 dB HL and in excess of 90 dB HL respectively (International Islamic University Malaysia, 2014). Table 2.1 outlines the level of hearing loss and the corresponding averages.

Table 2.1 Audiometric descriptors for pure tone hearing threshold levels in (a) Britain and (b) Malaysia (British Society of Audiology, 2011; International Islamic University Malaysia, 2014)

Level of loss	**Average HTLs (dB HL)**
Mild	20 - 40
Moderate	41 – 70
Severe	71- 95
Profound	In excess of 95

(b)

Level of loss	**Average HTLs (dB HL)**
Mild	20 - 40
Moderate	41 – 70
Severe	71- 90
Profound	90 and above

2.1.3 Anatomy and physiology of hearing

Hearing loss may rise due to any abnormalities in the anatomy and/or physiology of the hearing system. In the current study, the effect of hearing impairment to the ability in

hearing speech is one of the research interests. To understand how hearing loss happens, an understanding of the anatomy and physiology of hearing is essential.

Sound travels in the form of waves through the external ear canal, eardrum and the ossicles (malleus, incus and stapes) (Figure 2.2) before it enters the inner ear. When the waves hit the eardrum, malleus bone attached to the proximal side of the eardrum vibrates and this in turn vibrates the incus and stapes. During this part of the pathway, the sound energy is transferred mechanically. The inner ear contains the cochlea, a tubular coil-like structure which looks similar to a snail. It contains the Organ of Corti, the sensory organ of hearing. In the Organ of Corti, the energy transferred from the sound that travels through the middle ear displaces the basilar membrane and stimulates the hair cells. The stimulation of the hair cells activates the sensory nerve fibers adjacent to them and produces electrical potentials. These action potentials are then sent to the central auditory pathway through the auditory nerve. The sound information in the form of potentials terminates and is processed in the auditory cortex, located in the temporal lobe of the brain.

The perception of sound is usually manifested by two elements – pitch and loudness. This corresponds to frequency and intensity coded in the auditory pathway. The coding of frequency starts at the cochlear level. The widening of the basilar membrane towards the end of the cochlea causes the basilar membrane to have maximum displacement at different points in along the cochlea depending on the frequency of the sound. This in turn will stimulate different nerves fibres and allow for frequency discrimination. Lower frequencies are known to be detected at the apical end of the cochlea while higher frequencies are detected at the basal end. The mapping of the frequencies continue along the central auditory pathway up to the cortical level, where different nerve fibres and sections of the auditory cortex are known to be more receptive to certain frequencies.

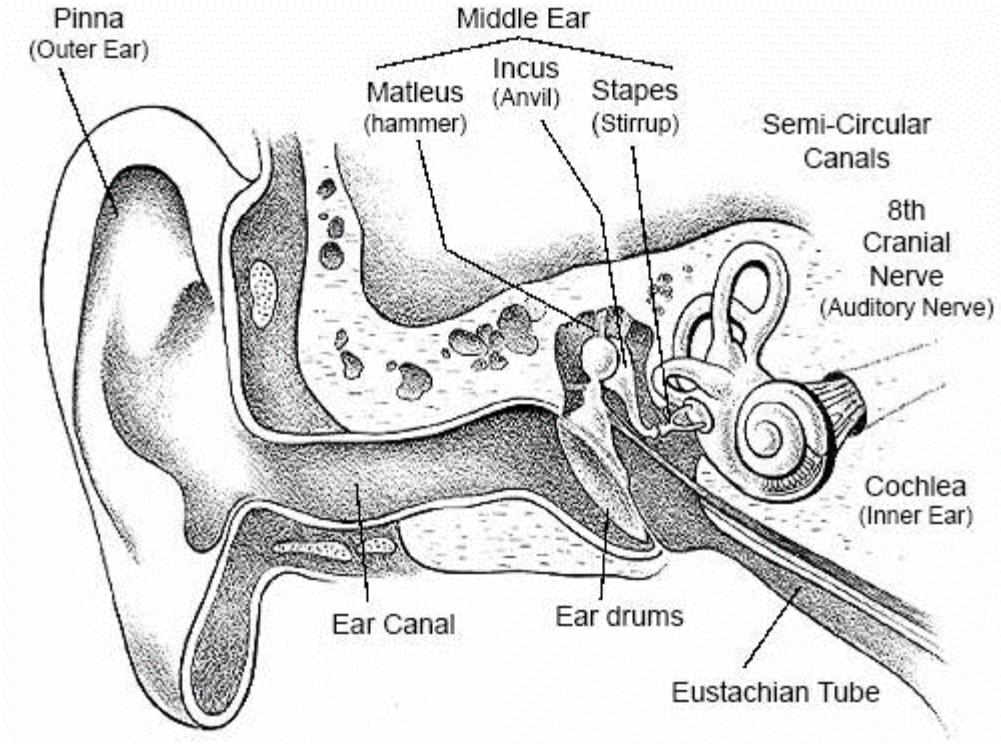


Figure 2.2 Cross section of the peripheral portion of the auditory system (Colorado Hands & Voices, 2013)

2.1.4 Types of loss

Classification of types of hearing loss is made based on the site of disorder and/or the underlying cause of loss. There are two basic types of hearing losses - sensorineural hearing loss (SNHL) and conductive hearing loss (CHL). A combination of the two, e.g. when a person suffers from both SNHL and CHL, produces mixed hearing loss (MHL). To understand the types of hearing loss, it is necessary to study the anatomy and physiology of hearing.

2.1.4.1 Conductive hearing loss (CHL)

Conductive loss occurs when there is any disturbance in the transfer of the mechanical energy along the peripheral auditory system. Structures that are usually involved in conductive hearing losses are the external ear canal, the eardrum and the ossicles. Disturbance can be in the form of a blockage, for example an occlusion in the external ear canal by earwax, or fluid in the middle ear. Disruption of energy transfer at the

ossicles, such as discontinuity/dislocation or stiffening of the ossicular bones, may also cause conductive loss.

In principle, conductive loss involves attenuation of the energy, and therefore the intensity, of the sound that reaches the inner ear. Frequency discrimination is not affected due to the fact that the inner ear, the site where frequency is analysed, is unaffected in CHL. Therefore, increasing the intensity of the sound presented to the ear with CHL up to a level at which it overcomes the attenuation allows the ear to hear without compromising the pitch perception. Due to the nature that causes the disruption of the energy transfer along the outer and middle ear, conductive losses are generally temporary; the hearing improves once the condition is better.

2.1.4.2 Sensorineural hearing loss (SNHL)

Sensorineural hearing loss is caused by disorders in the sensory and neural part of the auditory pathway. It can be further classified into cochlear loss and retrocochlear loss.

Cochlear loss or sensory loss originates from disorders in the cochlea. Causes of damage to the cochlea and hair cells include trauma caused by noise, infection medication, congenital disorders, metabolic and genetic disturbances, and even old age (Moore, 2007; Frisina, 2009). Abnormalities of the hair cells form the majority of cochlear loss (Katz, 2014). This affects frequency discrimination, which means even though the intensity of sound that reaches the inner ear has overcome the attenuation caused by the hearing loss, the ear might still not be able to distinguish sounds of certain frequencies. Cochlear loss can also be characterised by having rapid growth of loudness or recruitment, causing the range between the threshold of hearing to the loudness discomfort level (the level at which sound causes discomfort to the listener) to be lesser than normal hearing people (Gelfand, 2009; Katz, 2014)

Retrocochlear loss or neural loss are types of hearing loss that originates from abnormalities beyond the cochlea, including the auditory nerve and any part of the central auditory nervous system (Gelfand, 2009). Abnormalities may be caused by tumours, such as acoustic neuroma or tumours associated with neurofibromatosis type 2 or degeneration of the nerve fibers. Conditions that cause cochlear loss, such as infection, congenital disorders and genetic disturbances may also cause retrocochlear loss (Gelfand, ibid.). A characteristic of retrocochlear loss that may distinguish it from

cochlear loss is the loudness adaptation, a condition at which the loudness perception of a suprathreshold, continuous sound decays (Katz, 2014).

The characteristics of the sensory and neural hearing losses affect the results of speech audiometry (Townsend and Bess, 1980; Boothroyd, 2008). The effect of reduced frequency discrimination ability is presented in the maximum speech recognition score (MSRS). Frequency discrimination, particularly in the cochlea, enables the ear to discriminate and correctly recognise speech sounds. Therefore, in a speech recognition test, normal hearing listeners usually achieve high scores (>95%) for MSRS. Listeners with sensorineural loss, particularly those with severe to profound loss, will show lower MSRS as a result of the affected frequency discrimination ability (Jerger et al., 1968).

The loudness decay experienced by those with retrocochlear loss will generate a unique pattern of speech audiometry curve called 'rollover' (Jerger et al., 1968; Hannley and Jerger, 1981; Humes, 2002). At high intensity presentations, the loudness decay occur and affect the response of the listener. The sound at higher intensity is somehow perceived as softer than when it is presented at lower intensities (Jerger et al., ibid.). The decreased loudness perception at higher intensities results in lower correct scores. This effect is seen in the speech audiometry curve in the form of rollover where, after a certain presentation level, the correct score decreases as the presentation level decreases. An example of a rollover is shown in Figure 2.3.

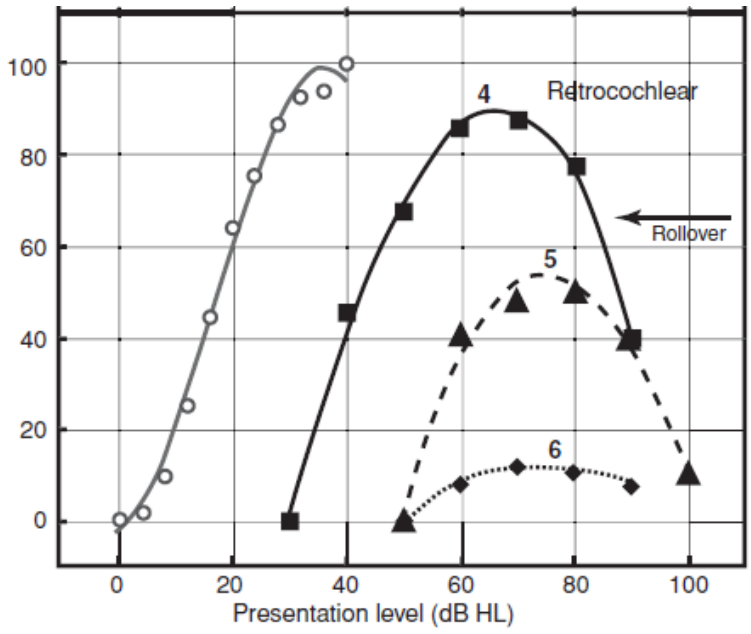


Figure 2.3 Example of speech audiometry curve rollover (McArdle and Hnath-Chisolm, 2015)

2.2 Hearing tests

Hearing tests can be divided into two main categories – behavioural tests and physiological tests. Behavioural tests requires the listener to display a change in behaviour, e.g. pressing the button, raising the hand, putting blocks through holes, when a sound stimulus is heard (NAL, 2015). Physiological tests or objective tests do not require any behavioural changes; instead, they only require passive cooperation from the listener. Responses to sound stimuli are recorded through physiological changes, such as electrical signals from neural activity or sound emissions from the ear (NAL, ibid.). Examples of behavioural hearing tests are pure tone audiometry, speech recognition testing and visual response audiometry. Auditory brainstem response, electrocochleography and otoacoustic emissions are examples of physiological hearing tests.

Audiology services in Malaysia are mainly offered in government and private-owned hospitals and clinics as well as hearing aid clinics. No published studies can be found on the audiology services in Malaysia. However, review of audiology clinic websites in Malaysia and visits to several audiology clinics in both government and private settings provide some information on the services offered. Routine assessment for adults comprise of otoscopy, pure tone audiometry and tympanometry. Diagnostic tests such as auditory brainstem response (ABR), auditory steady state evoked response (ASSR) and otoacoustic emissions (OAE) are also widely available (Perfect ENT Hearing and Speech Centre, 2011; Universiti Malaya Medical Centre, 2014; Loh Guan Lye Specialist Centre, 2016; Jensen Hearing, 2016; Tab a Doctor, 2016).

Although there are speech audiometry materials available in Malay, speech audiometry is not administered extensively in Malaysia. The review of audiology clinic websites in Malaysia did not list speech audiometry as part of their services. Limitations in conducting speech audiometry possibly arise from the limited access to published speech audiometry materials in Malay. One speech test that was commercialised was Malay Hearing-In-Noise Test (MyHINT); however, at the point of writing MyHINT is not commercially available anymore.

2.3.1 Hearing thresholds

Hearing threshold level (HTL) is universally defined as the lowest intensity level at which the listener responds to the stimulus at least 50% of the time (Gelfand, 2009). Depending on the stimulus, the definition of threshold may rely upon the nature of test itself; for example, pure tone audiometry defines pure tone HTL as the lowest intensity at which the responds two out of two, three or four responses on the ascending trials (British Society of Audiology, 2011), while speech audiometry that utilises the performance-intensity (PI) function defines the speech reception threshold as the level at which the listener scored 50% correct (Hudgins et al., 1947; Hirsh et al., 1952; Boothroyd, 1968).

Threshold seeking can be done by measuring the level of stimulus needed to evoke a response from the patient, also called audiometry. The lowest stimulus intensity to evoke a response from a person is considered as his hearing threshold for that particular stimulus. Convention has agreed to define hearing threshold of a sound as the lowest level of intensity of the sound at which the patient responds at least 50% of the time (British Society of Audiology, 2011; American Speech-Language-Hearing Association, 1978). This definition of threshold applies across the range of stimuli used in audiometry.

Pure tones are the commonest stimuli used in audiometry. They are more preferable over other types of sound due to their discrete frequencies, which makes the stimuli easy to calibrate and be standardised across clinics. However, there is an argument that pure tones do not represent actual daily listening abilities, which is the ability to hear and understand speech (Gelfand, 2009). Therefore, samples of speech, as alternatives to pure tones, are used as stimuli in hearing tests.

An audiometry is conducted using an audiometer (Figure 2.4). There are 4 essential components in an audiometer: sound generator to generate stimulus to be used in the audiometry. Dependent on the type of audiometry, signals can be pure tone generated by oscillators, or speech sounds provided through live voice or compact disk. Amplifier serves to amplify the tone to a maximum of around 110 dB HL, attenuator to manipulate the intensity level of the stimulus, and air conduction and/or bone conduction transducers to transmit the stimulus to the listener (Martin and Clark, 2010). Figure 2.5 shows block diagrams for (a) pure tone audiometer and (b) speech audiometer.

Figure 2.4 An Itera II diagnostic audiometer (GN Otometrics, 2016)

2.4 Speech audiometry

Speech audiometry is a group of behavioural hearing tests that uses speech as stimuli. Carhart (1951) defined speech audiometry as a technique in which standardised samples of a language are presented through a calibrated system to measure some aspect of hearing ability. It was designed to provide information on a person's ability to hear and understand speech as well as to overcome the limited information of speech hearing given by pure tone audiometry (Gelfand, 2009).

Speech sounds contain complex acoustical characteristics; the temporal, frequency and amplitude modulations vary widely in speech (Zatorre et al., 2002). A study on speech processing suggested involvement of different cortical areas between simple noise and speech (Zatorre et al., 1992), therefore, suggesting that the information on hearing acuity using pure tones might not reflect the hearing acuity for speech.

2.4.1 Earlier Speech Audiometry in English

The speech audiometry using word lists that formed the basis for later speech audiometry materials were developed and published by Hudgins et al. in 1947 (American Speech-Language-Hearing Association, 1988). Hudgins et al. (1947) developed two recorded tests to measure the loss of hearing for speech, one of the tests used words

as the stimuli, the other, short sentences. The loss of speech is defined as the difference between the level at which the hearing impaired person scores 50% correct of all test items and that of a normal hearing person.

Hudgins et al. (1947) outlined the essential characteristics of the test materials, all of which are taken into consideration by subsequent developers of speech audiometry, including those in other languages. The criteria given by Hudgins et al. (1947) for the selection of test items are 1) word familiarity, 2) phonetic dissimilarity, 3) normal sampling of English speech sounds (for speech audiometry in English) and 4) homogeneity with respect to basic audibility. These criteria has since been used as basis for the development of speech audiometry materials (Lehiste and Peterson, 1959; Tillman and Carhart, 1966; Boothroyd, 1968; Ashoor and Prochazka, 1982; Ousey et al., 1989; Mukari and Said, 1991; Nissen et al., 2007;; Fu et al., 2011) .

The aim of the tests are to explore problems that might involve in the development of a test assessing hearing loss for speech, produce a test material that is suitable to be used in a wide range of intensity and explore the possibility of developing a speech audiometry that is able to differentiate a high-frequency hearing loss from a flat loss. To achieve the second aim, recorded test materials, which can be played back at the intensity required to measure the loss, were introduced. To attain the third aim, Hudgins et al. (ibid.) first defined the term 'high-frequency hearing loss' in relation to the test. This is to overcome the problem of the unlimited configurations of high frequency losses. High-frequency losses were grouped into two: 1) high frequency losses with low- to high-frequency slope of 5 to 15 dB or more per octave and 2) high frequency hearing losses with an abrupt slope at or above 1000 Hz. For this, the developers had considered two solutions – devise a test that is composed of items made up of high-frequency phonemes, or have a set of test items consisting of normal sampling of English phonemes put through a high-pass filter to give them a high frequency emphasis. The test item were designed to have high-frequency emphasis by inserting a 4000Hz high-pass filter with attenuation of 17dB/octave for frequencies lower than 4000Hz. Limited lower frequency speech cues forced the listener to depend on the higher frequency cues to detect/recognise the stimulus. The second approach showed better discrimination towards high-frequency loss. It also was able to differentiate a uniformly sloping loss (loss with low- to high-frequency slope) and a loss with high-frequency cut-off, as long as the cut-off frequency is lower than 3000 Hz. However, Hudgins et al. (ibid.) concluded that the use of speech

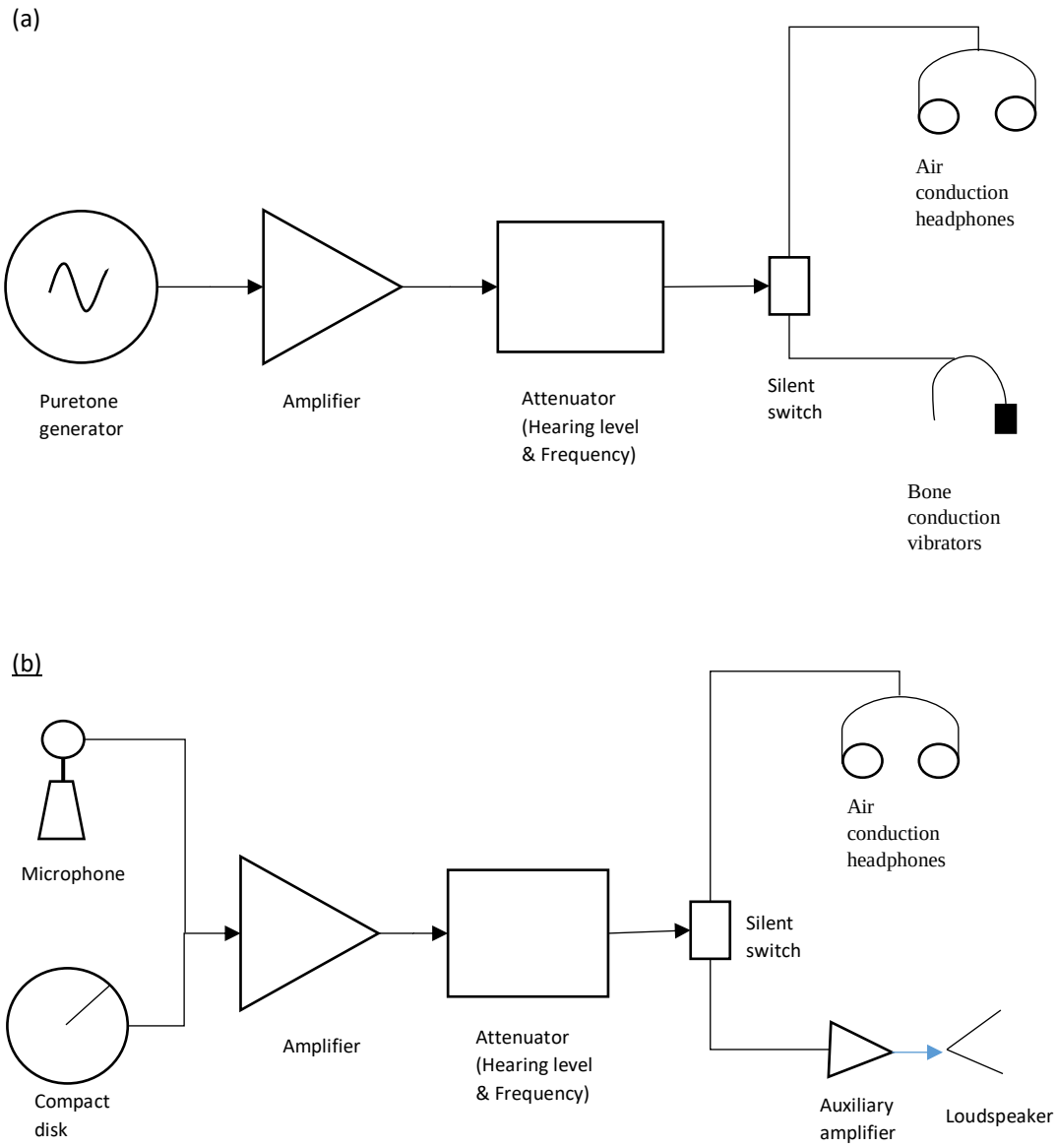


Figure 2.5 Block diagrams of (a) a pure tone audiometer, and (b) speech audiometer (Martin and Clark, 2006)

audiometry is not as efficient as pure tone audiometry as a measure for high-frequency hearing loss.

For the construction of word lists, Hudgins et al. (1947) decided to choose spondees as they showed high audibility, steeper articulation function and therefore greater relative homogeneity compared to trochees, iambs and monosyllabic words. All in all, 84 spondees were chosen, making up two lists and 6 scrambled versions each. In each list,

the 42 words were divided into 7 groups of 6 words, and recorded at decreasing intensity levels of 4 dB interval. This allowed each version of tests to cover a range of 24 dB. For the sentence lists, Hudgins et al. (ibid.) chose 28 short questions that can be answered with single words (e.g. 'What number comes before 10?'). Homogeneity of the sentences was achieved by adjusting the intensity of the test items during the final recordings. The intensity level of recording was also similar to the word lists, by which seven groups of four sentences were recorded with decreasing intensity of 4dB per interval.

The concept of phonetic balance had not been introduced during Hudgins et al.'s (1947) construction of speech audiometry materials. They concluded that although normal representation of English speech sounds should be a criterion in constructing the test items, it was not as important the other criteria. They supported the notion by giving an example of an earlier speech audiometry, the Western Electric 4C Test, which was able to provide reliable measurement despite having only seven digits as its test item. The varied recording intensities for each group of test items reflect the level of technology at the time.

Hirsh et al. (1952) attempted to improve on earlier speech audiometry materials. They developed a new test material to add to the objective of PAL Auditory Test No. 9 developed by Hudgins et al. (1947). The aim was to measure discrimination loss on top of measuring hearing loss for speech instead of measuring hearing loss for speech and loss of discrimination separately. Two major improvements were made; the test items were selected based on their fulfilment of a stricter criterion for familiarity and phonemic balance, and the recording of test items on magnetic tapes and therefore recording it only once and allowing repeated use of individual test item instead of recording each item repeatedly for each of the lists as per Hudgins et al. (ibid.).

Discrepancies in hearing thresholds for speech when certain PAL Auditory Test No.9 word lists were used indicated that the PAL word lists were not standardised. Another previously established material, the PB-50 word lists, contained a large vocabulary which was found to be unsuitable for some patients. The speech audiometry materials developed by Hirsh et al. (1952) produced three sets of word lists: CID Auditory Test W-1, Test W-2 and Test W-22. The differences between the sets are the purpose of the test, the type of words used for the test items and the way the test items are presented to the listener.

CID Auditory Test W-1 was designed to measure the threshold of speech intelligibility. The test items are made of 36 spondees, compiled in a single list. The spondees were

sourced from PAL Auditory Tests No. 9 and 14, and were filtered in terms of their familiarity and homogeneity to achieve equal intelligibility among test items. Six lists with varied word sequence were made. To further attain equal intelligibility, the more difficult test items were presented at an elevated intensity as compared to the easier words.

To measure the threshold of speech intelligibility, the six lists are presented to the listener at levels +4 to –6 dB SL in 2 dB steps. The scores are plotted in a performance-intensity function, similar to the speech curve used in current speech audiometry tests. The intensity level at which the listener scored 50% correct is considered as the hearing threshold for speech.

CID Auditory Test W-2 was designed for rapid estimate of the threshold of intelligibility. It utilised the same lists as in Test W-1 but with a different presentation procedure. Instead of presenting each list of the 36 spondees at a constant level, the test items in Test W-2 is presented at 3dB steps every three words. This allowed one list to cover a wider range of intensity, 33 dB, compared to Test W-1, which only covered a range of 10 dB. The estimated threshold of intelligibility is also defined as the level at which the listener scored 50% correct. When compared with the threshold measured through Test W-1, it showed that Test W-1 presented an average threshold of 3.5 dB better than Test W-2.

The third test, W-22, is a measure of loss of discrimination of speech instead of a measure of speech intelligibility. Discrimination loss is calculated by comparing percentage of correct scores to the full score, which is 100% correct. Four phonetically balanced lists of 50 monosyllabic words made up the material, with each list having six scramblings of word order. The score of a patient is measured at the level at which any further increase of stimulus intensity does not result in any increase of correct responses.

These earlier tests form the basis of much of contemporary speech recognition tests in both English and non-English languages. However, the downside of these tests is that the measurement of threshold of intelligibility and the measurement of speech discrimination require two separate tests. This requires more time and labour. A test that could combine both measurements was desirable.

American Speech-Language-Hearing Association (ASHA) has produced a guideline for determination of threshold for speech in English (American Speech-Language-Hearing Association, 1988). The recommended test material is a combination of the spondaic test items developed by Hudgins et al. (1947), also known as PAL Auditory Test No. 9,

and word lists developed by Hirsh et al. (1952). Although these tests have been around for decades, they are still being used in the clinics. This raises several questions; “Would speech recognition test be beneficial for assessing speech hearing in Malay population?”, “Are spondees applicable in Malay language, and if not, what is the word structure most suitable for Malay speech recognition test?” and “Would a speech audiometry material in Malay language be able to measure both speech reception threshold and discrimination loss?”

2.4.2 Speech reception threshold tests in English Language

A more desirable speech audiometry material is the one that allows speech recognition and discrimination measurements in one test. Following the speech reception threshold tests designed by Hudgins, et al. (1947) and Hirsh, et al. (1952), later developers of speech recognition tests tried to combine the measurements of speech intelligibility and speech discrimination into a single test.

In 1959, Lehiste and Peterson introduced a set of phonemically-balanced word lists called the CNC word lists. The set was made up of ten lists of 50 monosyllabic, consonant-nucleus-consonant (CNC) words each. Phonemic balance meant that the phonemes occur with the same frequency in each list. The set was revised for a more uniform word frequency, as it was thought that word frequency might affect the response (Peterson and Lehiste, 1962). Causey et al. (1984) utilised the revised CNC lists to produce normative data on the performance-intensity (PI) functions of normal hearing and hearing impaired participants by presenting the lists at several intensity levels between 4 and 40 dB SL. The result was a PI curve depicting the correct scores at the presentation levels. The 50% word recognition score was taken as the threshold while the maximum speech recognition score represented speech discrimination. Comparison of the PI function between normal hearing and the hearing impaired showed that there were significant differences in speech reception thresholds and maximum speech recognition scores. This showed that the measurements for speech intelligibility and speech discrimination could be combined in one test.

McCormick developed a speech test that is designed for children above the mental age of 2 years (1977). Although the test is not strictly a discrimination test, McCormick Toy Test incorporates discrimination of phonemes using 14 paired words, accompanied by matching ‘toys’. The test was developed with the aim of assessing the integrity of

neurological hearing pathways beyond what is used in pure tone measurements. The task involves identifying test items, therefore challenges the cortical processing in discriminating and interpreting the stimulus given. The identification task will also be able to detect children with good hearing acuity but poor discrimination ability. As the test is primarily aimed for children, the vocabulary of the test items are limited and, therefore, not suitable to be used for assessing hearing acuity of older children and adults. However, the concept of paired items with matching vowels or diphthongs but differing consonants is applicable in the design of speech discrimination tests.

The concept of paired words used by McCormick (1977) in order to measure discrimination can be applied in the current study. However, having minimal pairs for each of the test item in speech recognition test aimed for adults would require much longer list of words in order to cover the variety of phonemes. In the current study, several adjustments can be considered; the use of multisyllable words will cover more phonemes without jeopardising the length of word list, and the use of nonsense words as pairs to match the meaningful words reduces the need to find matching word pairs which would limit the number of words suitable to be included.

2.4.3 Speech reception threshold tests in non-English languages

Recent development on speech reception threshold test involves more material in non-English languages. Developers of speech audiometry word lists identified the need for having material in the native languages of the listeners, as speech reception thresholds are poorer in those tested not in their native languages (Hapsburg et al., 2004). This poses the need for having speech audiometry material in Malay language.

A psychometrically-equivalent bisyllabic speech discrimination material in Mandarin to measure speech discrimination was developed by Nissen et al. (2005a). Although monosyllables in Mandarin do carry lexical meaning, bisyllables were selected as the test items based on two reasons: the evidence that they are stored as whole-word representations, and it is easier to provide written responses to bisyllabic words as monosyllable may be written in many different characters. The material consists of 4 psychometrically-equivalent lists containing 50 words each which can be made into 8 half-lists with 25 words each. The bisyllabic words were sourced from a frequency dictionary in Mandarin and an online news corpus. Unsuitable words were taken out of the selection with the help of judges prior to listener evaluation. The recordings of the test items were made with one female talker and one male talker. The recorded words

were evaluated in terms of perceived quality of production and quality of recording. Preliminary testing was done to evaluate the psychometric equivalence of the lists. This was done by grouping the initial 240 words randomly into 10 lists and presenting the lists at 10 intensity levels between -5 and 40 dB HL. The test was done on a group of 10 participants, each with the order of presentation of the words and the order of presentation of the lists randomised. The exercise was repeated using a different group of participants and word content of each list randomised again. The preliminary test began with giving the instructions to the participants followed by presenting the test items. No use of carrier phrase reported. The response was made by writing the perceived word on the response sheets. To study the psychometric equivalence of the words, regression slopes for the 240 words were calculated and ranked. Two hundred words with the steepest slopes were selected and randomly divided into 4 groups of 50 words. The full lists (with 50 words) were then divided again into 8 half lists of 25 words. All lists were represented by a female and a male talker. The regression slopes of all lists were calculated using a formula in order to predict the percentage of correct performance at a specified intensity level, and thus the threshold, the slope at threshold and the slope from 20 to 80% correct responses. The result showed that the psychometric function slopes from the female talker were steeper than that of the male talker, although the differences were not statistically significant. Adjustment on the intensity in order to have the performance from each talker equal showed that the psychometric functions to be very similar.

The development of the speech discrimination material by Nissen et al. (2005a) was very well-thought in terms of its psychometric equivalence. However, the test protocol described in the paper is more reflective of a speech recognition test. As the correct response for each item is described as matching the lexical tone and pronunciation of both syllables, the discrimination ability of the participants are not being assessed clearly. The written response method would also limit the participants to those who are able to write in Chinese characters.

Another group of researchers has also come up with another set of bisyllabic speech audiometry material in Mandarin (Wang et al., 2007). The main difference between the material developed by Wang et al. and Nissen et al. (2005a) is that the one by Wang et al. (ibid.) was phonologically balanced while the one by Nissen et al. (ibid.) was not. Although the authors did report that the issue of phonetic balance was still debatable,

they felt that phonetic balance was '…the important and frequently used method of ensuring content validity in word lists' (Wang et al., ibid.).

Similar to Nissen et al. (2005a), Wang et al. (2007) had sourced the words from publications on the most common words as well as word frequency in Mandarin. To select the words suitable to make up the phonetically balanced word lists, the authors had analysed the phonetic characteristics of the Mandarin language in terms of its phonemic content (consonants and finals) and tones. Ten lists with 50 words each were produced following several criteria – all words are familiar, syllables of the words are equally stressed, combination of the monosyllables to produce the word does not alter the tone and the words make up a phonetically balanced list. Unlike the method of recording done by Nissen et al. (ibid.), Wang et al. (ibid.) had employed only one male talker to record the test items. There was no carrier phrase at the beginning of the test items in the current study.

To study the homogeneity of the lists, Wang et al. (2007) presented all 10 lists to each of the 60 participants at the presentation level of average pure tone threshold at 500, 1000 and 2000 Hz plus 2 dB. The order of the lists was changed for each participant in to avoid effects caused by unfamiliarity with the test, practice, and fatigue. A score was given for every correctly repeated word and the percentage of correct response formed the overall individual score. The lists were found to have equal difficulty except for list 5, and therefore list 5 was taken out from the set.

Wang et al. (2007) had also studied the performance-intensity (PI) function of normal hearing and hearing-impaired participants using the material they had developed. Thirty-five normal hearing participants and forty participants with mild to moderate hearing loss participated in this part of the study. The normal hearing participants were presented with the lists at 15, 12, 9, 6, 3 and 0 dB HL while the hearing-impaired participants were presented with the material at SRT+15 dB, SRT+10 dB, SRT+5 dB, SRT, SRT-5 dB and SRT-10dB (SRT being the speech reception threshold and established using an adapted Monosyllabic Adaptive Speech Test (MAST) procedure using 2 dB steps). Comparison between the speech reception threshold (intensity level at 50% correct scores) and average pure tone threshold at 500, 1000, 2000 and 4000 Hz showed that both are in good agreement for both normal hearing participants and hearing-impaired participants, and therefore suitable to be used in clinical situations.

The difference between the method applied by Wang et al. (2007) and Nissen et al. (2005a) could be seen in the equivalence analysis. While Nissen et al. (ibid.) made their

lists psychometrically equivalent by first selecting words with the most similar psychometric functions and then digitally adjusting the intensity to reach psychometric homogeneity, Wang et al. (ibid.) first built the set of lists, tested the lists for equivalence and then eliminated the lists with significant difference in equivalence, keeping the rest of the lists for the set. The method used by Nissen et al. (ibid.) preserved the words that was included in the lists, while method by Wang et al. (ibid.) required having a larger number of preliminary lists and posed the risk of having much shorter final lists.

In the current study, several improvements can be made on the psychometric equivalence of the word lists. To facilitate the equivalence of the lists, preliminary selection of words can be made more stringent by assigning several criteria to it. The selection criteria may include restricting the familiarity level of the words, thus increasing the probability of the words to be psychometrically equivalent. In addition, the distribution of phonemes can be determined in advance in order to be able to set a filter on the phonemes to be included in the list.

2.4.4 Speech reception threshold tests in Malay

To identify the areas in Malay speech reception threshold test materials that needed improvement, the published Malay speech reception threshold test word lists are reviewed. There are two previously published studies on the development of speech audiometry materials in Malay language (Hong, 1984; Mukari and Said, 1991). The first set of word lists, developed by Hong (ibid.), aimed to develop a short word lists that reflect the natural usage of Malay-speaking population in Singapore, Malaysia and Brunei for the purpose of speech recognition audiometry. The set consists of 10 different lists with 10 bisyllabic words. The words were chosen following a criterion on familiarity, equal average difficulty, equal range of difficulty, representation of Malay language, and common usage. Hong decided not to have his word lists phonetically balanced. However, he claimed that each is list has 20 phonemes and contains a ‘rough approximation’ of phonetic balance found in everyday spoken Malay (Hong, 1984). There was no indication of the source or materials used to build the list. The method of measuring familiarity, and language usage and representation were not indicated as well. Upon reading the lists, there is a possibility that Hong (ibid.) meant to have 20 syllables instead of 20 phonemes in each list. The lists contain bisyllabic words which are consonant-vowel-consonant-vowel (CV-CV), CV-CVC, V-CV and V-CVC in nature. However, not all structures are

included in each list. The phonemes included in the set are vowels [a], [i], [u], [ə] and [o], and consonants [b], [č], [d], [g], [h], [j], [k], [l], [m], [n], [p], [r], [s], [t], and [ʔ]. All lists failed to incorporate each of the phonemes; therefore the representation of phonemes between each list is not the same.

A speech discrimination curve for normal hearing participants were established. The curve follows the shape of P-I function curve established in earlier speech audiometry studies (Hudgins et al., 1947; Boothroyd, 1968; Ashoor and Prochazka, 1982). The lists were tested through analysing the mean scores heard at a constant, near threshold intensity of each list. Three objectives were tested – interlist difficulty, equal difficulty between Malay and non-Malay group, and test-retest reliability. There were no significant differences between the lists in the interlist difficulty analysis, no significant differences in terms of difficulty between the Malay and non-Malay groups and no significant differences in term of scores between the test and re-test results. The lists were also tested on a case of bilateral moderate-to-severe conductive loss, a mild sensorineural hearing loss accompanied by vertigo and tinnitus case and a case of suspected functional loss. The first two cases reflected the audiogram obtained from the patients, while the patient with suspected functional loss showed an inconsistent speech curve. The extent of the validation of the material is limited to the number of normal participants and the three patients, therefore it seemed too naïve to validate the material's use in speech audiometry.

The second set of word lists were developed by Mukari and Said (1991). The set of bisyllabic word lists was intended for speech recognition audiometry. The word lists, unlike Hong's (1984), were phonemically balanced, which means that the distribution of phonemes is equal among the lists. There are 25 lists with 10 words each. All of the words are of CV-CV structure, as it was reported to be the most common syllable structure in Malay language. Four hundred and fifty commonly-used words were pre-selected from the Dewan Bahasa dictionary. The familiarity of these words was assessed by 150 Malaysian adults of different racial backgrounds, resulting on 196 being shortlisted for further use.

To develop the word lists, Mukari and Said (1991) calculated the frequency of occurrence of 25 consonants (26 when counted), 6 short vowels and 6 diphthongs. Several phonemes were excluded as they were considered either low in occurrence or having the same sound as the included phonemes. The term 'diphthong' may have been misrepresented here as diphthong is defined as 'a vowel sound, occupying a single

syllable, during the articulation of which the tongue moves from one position to another, causing continual change is vowel quality' (Collins English Dictionary, n.d.) whereas [ia], [ua], and [io] are actually vowel sequences (Teoh, 1994). Due to low usage in Malay, [f], [ʃ], [z], [oi] and [io] were excluded. Mukari and Said (ibid.) had also excluded the phonemes [ð], [ɣ], [x] and [θ] on the basis that they are actually pronounced as [d], [g], [k] (also [h]) and [s] respectively. This indicated that the 38 'phonemes' that was listed were actually grapheme, which is the unit of writing that represents one phoneme.

Contrary to the word lists developed by Hong (1984), the developers of these word lists had included the 28 selected phonemes in each in almost the entire list. Twenty-five lists were compiled, each with 10 CV-CV words selected from 196 words shortlisted earlier. Here, the phonemic balance is defined by the occurrence of all 'phonemes' in almost the entire list. The frequency of occurrence, however, does not reflect the real-life usage. The interlist intelligibility difference between the lists was also tested. The prerecorded word lists were presented to 25 normal-hearing participants at a constant intensity of 10 dB above their average hearing thresholds at 500, 1000 and 2000 Hz. The mean scores between lists were then compared and no significant differences were detected.

Several improvements can be made on the Malay speech audiometry material. First, the word lists need to be validated and verified thoroughly. Verification that the word lists are equivalent, interchangeable and generate similar results was present in both studies by Hong (1984) and Mukari and Said (1991) but improvements can be made in terms of verifying the phonemic balance of the list as claimed by Mukari and Said (ibid.), as well as verifying the representation of Malay language (Hong, ibid.). The speech audiometry material can also benefit from having more comprehensive study on the performance of listeners with hearing loss in order to validate the use of the word lists in determining speech hearing level.

2.4.5 Speech audiometry using nonsense syllables

Speech audiometry using nonsense words are not as widely utilised in the clinic as its familiar and meaningful words counterpart. This can be seen in the limited number of available nonsense syllables test available in the market; furthermore, most of the tests are not designed to assess speech recognition (Mendel and Danhauer, 1997).

Mendel and Danhauer (1997) provide a review and compiled several nonsense syllable tests that have been published. Among nonsense syllable tests summarised by Mendel

and Danhauer (ibid.) are Fletcher and Steinberg Nonsense Syllable Tests, Manchester Nonsense Syllable Tests, Closed-response Nonsense Syllable Test (CUNY-NST), Nonsense Syllable Test (NST) and Distinctive Feature Difference (DFD) Test. The objectives of the tests include assessing finer discrimination skills, allowing a larger number of speech sound being tested with lower learning and practice effects and without having to use large number of stimuli, and assessing the clinical efficacy of nonsense words in speech audiometry.

Gelfand et al. (1992) modified the existing City University of New York (CUNY) Nonsense Syllable Test (NST) to allow resolution of consonant confusion errors and construction of single confusion matrix for the test items in speech audiometry. The modified NST is made up of 22 consonant-vowel (CV) syllables and 16 vowel-consonant (VC) syllables as test items. The differences between the original and the modified CUNY were the absence of carrier phrase and the use of only vowel /a/, instead of /a/, /i/ and /u/, in the modified version. The participants were normal hearing adults. Gelfand et al. (ibid.) found that there was no significant difference between the performance using the modified NST and the original NST, therefore, the use of carrier phrase and the choice of vowels did not affect speech audiometry results. They also found out that there were more correct responses made with VC as compared to CV at low presentation levels, possibly due to the higher energy of the initial vowel (compared to the lower energy consonants) which served to 'capture' the listener's attention to the test item.

Cheesman and Jamieson (1996) developed a closed-set, nonsense word, Distinctive Features Differences (DFD) test to provide a measurement of the listener's ability to identify consonant sounds and to identify confusion errors made by the listeners. The design of the test material were aimed to contain 22 consonant sounds, intervocalic (vowel-consonant-vowel) in nature, having four speakers (two male and two female), digitised and neutral in terms of pronunciation, intonation and accent pertaining to central Canadian English. The test items employed the form /^**C**ɪl/ in which **C** is one of the 22 consonants. The test items underwent screening for audibility, category appropriateness and homogeneity through a pilot test. One item, /^ɵɪl/ was taken out of the list due to its high level of confusion error. The test is closed-set and the subjects were required to select one out of 21 possible responses shown on a video monitor. All 84 items (21 items x 4 talkers) were presented in two conditions – speech-in-noise (SIN) and filtered speech. In the SIN condition, noise was kept constant while speech signal were varied. Filtering was done by applying 15 different filter conditions with a fixed signal to noise ratio (SNR).

The performance/intensity (P/I) function obtained from the SIN condition produced a curve with a shallow slope as compared to the usual steep slope seen in P/I function of other speech audiometry.

A study by Humes et al (1987) utilised nonsense syllables to compare the performance of listeners with sensorineural hearing loss and normal hearing listeners that were presented with masking noise in order to simulate hearing loss. The finding showed that the errors made by those with true hearing loss were significantly different from those with simulated hearing loss through masking noise, suggesting that subjects with simulated hearing loss might not be suitable substitutes to hearing impaired subjects.

The findings of these studies (Humes, 1987; Gelfand et al., 1992; Cheesman and Jamieson, 1996) would be useful in the consideration of the research design, especially in the design of the speech audiometry material. Most importantly, the use of nonsense syllables or words are usually aimed for the measurement of phoneme confusion errors instead of the measurement of speech hearing thresholds. The exclusion of carrier phrase prior to the presentation of test items would reduce testing time; the current study would benefit from not including carrier phrases in the speech audiometry material, considering that the current study aim to use bisyllabic nonsense words which are longer than the test items used by Gelfand et al. (ibid.). The choice of vowels to be used for the nonsense words in the proposed word lists should have no effect on the performance of the lists, thus, allowing more combinations of syllables to be included in the proposed set. As speech-in-noise test generate significantly different results from speech audiometry in quiet, inclusion of noise in the current study may not be suitable as it detracts from the objective of measuring speech hearing threshold.

2.5 Development of speech audiometry word lists

Speech audiometry is a hearing assessment tool used to measure the hearing for speech. As our daily mode of communication is usually in speech, it is believed that having speech as stimuli would be more representative of someone's hearing abilities (Hirsh et al., 1952; Martin and Clark, 2010). Similar to having universally accepted test frequencies in pure tone audiometry, speech audiometry utilises standardised and established sets of speech stimuli to ensure reliable comparison across test centres. The speech stimuli can come in the form of phonemes, words or sentences (Hudgins et al., 1947; Hirsh et al., 1952; Nilsson et al. 1994; Ling (1989) as cited in First Years, 2011).

Many earlier developers of word lists used for word recognition testing in speech audiometry believed that the content of the word lists should represent real conversational speech (Hirsh et al. 1952; Boothroyd, 1968; Zakrzewski et al., 1976). In order to simulate the presence of speech sounds in everyday conversation, the construction of these word lists applies the concept of phonetic balance. A phonetically balanced word list contains the phonetic elements with approximately the same percentage of occurrence in the language that they represent (Hirsh et al., 1952).

Another concept that is being used in developing a set of word lists is phonemic balance. While phonetically-balanced word lists take into consideration the relative frequencies of speech sounds (and their distribution) in the language, a phonemically-balanced word list reflects the phonemes of the relevant language, and their distribution (Zakrzewski et al., 1975; Gelfand, 2010). To differentiate between phonetic elements and phonemic elements, a review of phonetics and phonology are given below. Phonetics and phonology, particularly in Malay, are especially important in the current study as they form the foundation for the words and word structure that is to be included in the list. In addition, the formation of nonsense words in the current study is also based on the phonetics and phonology of Malay.

In addition to phonetic balance, there are several other factors that have to be considered in developing speech audiometry word lists. The following sections discuss these factors in detail.

2.5.1 Phonetics

Phonetics is defined as the study of speech sounds, relating to their production, perception, and characteristics (Hyman, 1975; Fromkin and Rodman, 1998). Phonetics provides an inventory and description of phonetic segments, which are the sounds produced by our anatomic and physiologic activities (Hyman, 1975).

Phonetics can be categorized into two – articulatory phonetics and acoustic phonetics. Articulatory phonetics studies the production of speech sounds – the muscular effort, positions of the organs of speech, shape of the oral cavity and airstream mechanisms, among others. Acoustic phonetics defines the physical properties of the sound that are produced, for example, frequency content, tone, duration and stress (Fromkin and Rodman, 1998)

The sounds that we are able to produce are limited by the organs of speech. Our organ of speech consists of the vocal tract and the articulators. The vocal tract is made up of the vocal cords, situated in the larynx, the oral cavity and the nasal cavity (Radford et al., 1999). The vocal cord vibrates as the air is pushed out of the lungs, and the various vocal tract configurations 'shapes' it into different sounds (Fromkin and Rodman, 1998; Radford et al., 1999).

As these speech sounds are in fact acoustic signals, they contain frequencies and intensities. Frequency content and intensity level of a sound are critical in our ability to hear, and therefore are vital in speech perception. Each speech sound is produced differently by the vocal tract and the articulators, and therefore each has its unique frequency content. This unique characteristic enables us to discriminate between speech sounds (Gelfand, 2010). The manner of production determines the level of intensity of the speech sound as well. Vowels show higher intensity levels compared to consonants (Gelfand, ibid.), and this is crucial in determining which sounds someone with a hearing loss can hear.

In phonetics, speech sounds are divided into two major categories – consonants (C) and vowels (V). Consonants are produced by changing the place of the articulators, which are the tongue, lips and teeth, while vowel production is determined by the shape of the vocal tract.

Consonants are further divided into subcategories, depending on the placement of the articulators that produce them. They can also be grouped further into their manners of articulation. These two subcategories, place of articulation and manner of articulation, for consonants are tabled in Tables 2.2 and 2.3.

Vowels are voiced sounds shaped by the positions of the tongue body and the lips (Radford et al., 1999; Gelfand, 2010). They can be described through articulatory phonetics – by tongue positions and lip rounding. To produce different vowels, the front or back part of the tongue can be raised or lowered, hence the term 'front', 'back', 'high' and 'low' vowels (Fromkin and Rodman, 1998). These positions and the vowels they produce can be summarized in a diagram called 'quadrilateral', with the vertical axis representing the height of tongue elevation and the horizontal axis the position of the tongue body horizontally (Radford et al., 1999). Figure 2.6 shows an example of a quadrilateral of English vowels.

Table 2.2 Consonants and their place of articulation

Subcategory	Place of Articulation	Example of consonants
Bilabials	Both lips are brought together	[p], [b] and [m]
Labiodentals	Bottom lip touching upper teeth	[f] and [v]
Interdentals	Tip of tongue inserted between upper and lower teeth	[θ] and [ð]
Alveolars	Part of tongue raised towards, or touching the alveolar ridge	[t], [d], [n], [s], [z], [l] and [r]
Palatals	Front part of tongue raised towards the hard palate	[ʃ], [ʒ], [č] and [ǰ]
Velars	Back of tongue raised towards the velum (soft palate)	[k], [g] and [ŋ]
Uvulars	Back of the tongue raised towards the uvula	[ʀ]*, [q]* and [ɢ]*
Glottal	For [h] - open glottis, no modification of articulators in the mouth For [ʔ] (glottal stop) – tightly closed glottis, so that airstream could not pass through	[h] and [ʔ]

(Based on Fromkin et al. (2003), pp. 242-243)

Consonants marked * do not occur in English

Table 2.3 Consonants and their manner of articulation

Subcategory	Manner of Articulation	Examples
Voiceless	Vocal cords apart, allowing the airstream from the lungs to go through unobstructed	[p], [t], [k], [s]
Voiced	Vocal cords together, forcing the airstream from the lung to vibrate the cords in order to go through the vocal tract	[b], [d], [g] and [z]
Nasal	Airstream escapes through both the oral and nasal cavities	[m], [n], [ŋ]
Oral	Airstream escapes only from the oral cavity (the raised velum prevents the air from flowing through the nasal cavity)	[p], [d]
Stops	Airstream is stopped completely in the oral cavity	[p], [b], [m], [k], [g], [ŋ]
Fricatives	Air at high pressure is forced through a narrow opening	[f], [v], [s], [z]
Affricates	Airstream is stopped by closure, but followed immediately by a slow release of the closure	[tʃ], [dʒ]
Liquids	Airstream is forced through an obstruction that is not narrow enough to cause any friction or constriction	[l], [r]
Glides	No obstruction to airstream, and tongue glides towards or away from the preceding or following vowel	[j], [w]

(Based on Fromkin et al. (2003) pp.244- 250)

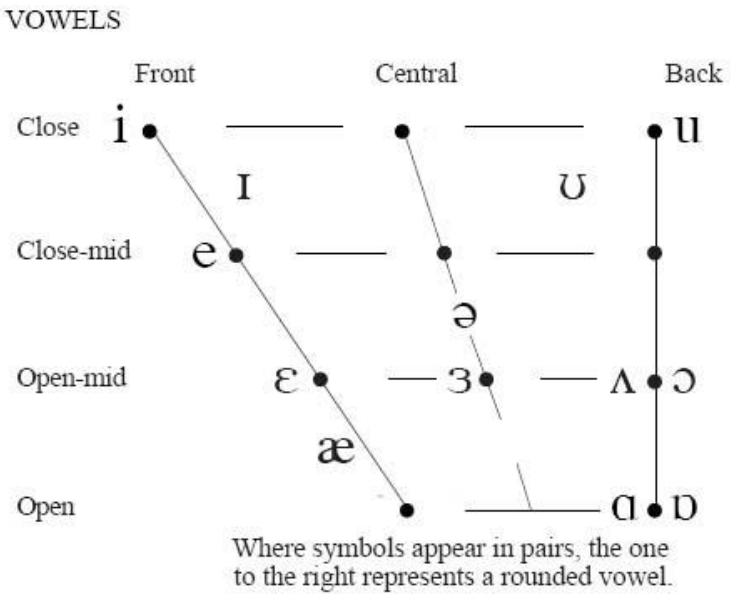


Figure 2.6: A quadrilateral of English vowels (Gramley, 2010)

2.5.2 Phonology

Even though the human organ of speech is capable to produce such many speech sounds, a certain language contains only a certain select speech sounds, which may or may not occur in other languages. Learning a language includes learning the speech sounds of that language.

The study of sounds of a language is termed phonology. Phonology also describes the sounds that occur in a language. Similar to phonetics, in phonology speech can be broken down to smaller segments of individual speech sounds like [a], [θ] and [ŋ]. Individual speech sounds that carry distinctive meanings in a language is called *phonemes* (Fromkin and Rodman, 1998).

One way to determine whether a sound is a phoneme of a language is to compare two words that sound almost the same. An example of these words in English is *sip* and *zip* which only differs in the initial consonants, [s] and [z]. Changing one speech sound, [s], to [z] changes the meaning of the word completely. Similarly with crook and croak, these words differ only in the vowels [ʊ] and [o] respectively. The sound segments, [s], [z], [ʊ] and [o], carry distinctive meanings, and therefore, are phonemes in English. These pairs of words, which differ only in one sound segment, are called minimal pairs. A group of three or more words that differ in one phoneme is called a minimal set. However, there

is a limitation to what can be called as a minimal pair/set. When the pair of words has one contrasting phoneme but it occurs in a different place in the string of phonemes (e.g. *make* [meɪk] and *kale* [keɪl]), or the pair is varied in more than one phoneme (e.g. *seed* [sid] and *sack* [sæk]), they will not be considered as a minimal pair. A pair of words may also have one differing phoneme but still carries the same meaning (e.g. either [iðer] vs [ajðer]); this pair is not considered as minimal pair, but is a *free variation* instead (Fromkin and Rodman, 1998).

For the purpose of comparison, English phonology that will be discussed in this review refers to the Received Pronunciation English phonology. Received Pronunciation English is chosen as a reference as it is the accent on which the phonemic transcriptions in dictionaries are based (Robinson, n.d.).

Malay has a different set of phonemes compared to English. As Malay, too, has many regional accents, the following discussion on Malay phonology will refer to the Standard Malay. Standard Malay (SM) is based on the Malay spoken in the southern part of Peninsular Malaysia, whose accent is termed *Bahasa Melayu Johor-Riau Suluh Budiman* (Onn, 1988). SM is the accent used by examiners during oral examinations and by newscasters in the national television.

Abdul Rahman (1988) and Teoh (1994) have provided a good summary of SM phonology. There are six vowels in SM, /i/, /u/, /e/, /ə/, /o/ and /a/, and eighteen consonant phonemes – [b], [d], [g], [p], [t], [k], [ʃ], [č], [m], [n], [ɲ], [ŋ], [s], [h], [l], [r], [w], and [y] (Tables 2.4 and 2.5). The rules of usage of the phonemes will be discussed later under the subtitle 'Word formation and morphology'.

2.5.3 Phonology and its relations to hearing loss

Each phoneme has its own unique sound and it is characterised by the frequency content of the sound. The ability to perceive different speech sounds depends on the ability of the listener to contrast the frequency content of the sounds. This feature is called vowel/consonant contrast (Wright, 1997)

Vowel contrasts are usually given by the frequencies of the first two formants, F1 and F2. The first formant, F1, is associated with the horizontal placement of the tongue, i.e. front or back (Figure 2.6). The second formant, F2, is associated with the height of tongue placement, which is high vs low (or 'open' vs 'close' vowels in Diagram 1) (Wright,

1997). Ladefoged (1982) presented the difference in F1 and F2 between English vowels in Figure 2.7.

Table 2.4 Malay consonants as described by Abdul Rahman (1988)

	Bilabial	Labiodentals	Dental	Alveolar	Palato-alveolar	Velar	Uvular	Post-uvular
Plosive	[p], [b]			[t], [d]		[k], [g]		ʔ
Fricative				[s]				[h]
Affricate					[č], [ǰ]			
Trill				[r]				
Lateral				[l]				
Nasal	[m]			[n]	[ɲ]	[ŋ]		
Liquids	[w]				[j]			

Table 2.5 Malay consonants as described by Teoh (1994)

Stop	[p], [b]	[t], [d]	[k], [g]	[ʔ]
Affricate		[c], [ǰ]		
Continuant		[s]		
Liquid		[l], [r]		
Nasal	[m]	[n], [ɲ]	[ŋ]	
glide		[h], [y] (or [j])	[w]	
fricatives	*[f], [v]*	*[θ], [ð]*	*[z], [s]*	*[x], [ɣ]*

**All in italics are consonants from loan words*

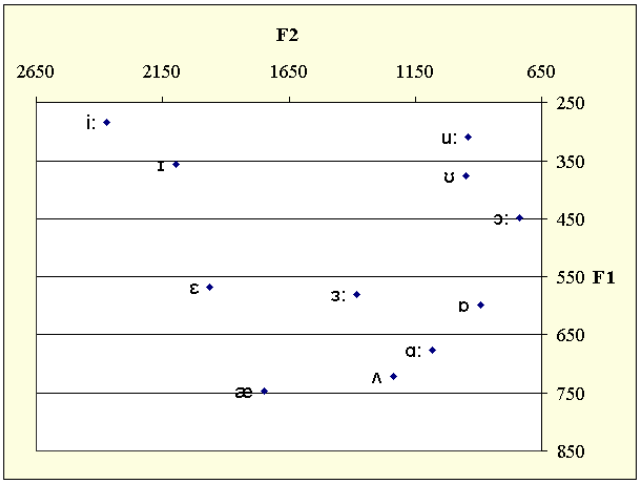


Figure 2.7 English vowels and their F1 & F2 (Ladefoged, 1982)

Consonants are also differentiated by their frequency content. Laryngeal tones are mostly located in the low frequencies, nasals in low- and mid-frequencies, while stops, sibilants and fricatives are in high frequency region (Wright, 1997).

Apart from frequency content, the intensity of which the speech sounds are vocalized is also related to the hearing abilities of the listener. Vowels are basically high in intensity, ranging from 35 to 60 dB. Fricatives and nasals have generally lower intensities, while affricates have relatively higher intensities. Figure 2.8 shows the ‘speech banana’, a map of frequency content and intensity of several speech sounds. The shaded area is concerned in the frequencies and intensities where speech sounds is located and popularly named as the ‘speech banana’.

Due to the different intensity and frequency content of speech sounds, the ability to perceive the speech sounds is highly dependent on the hearing capabilities of the listener.

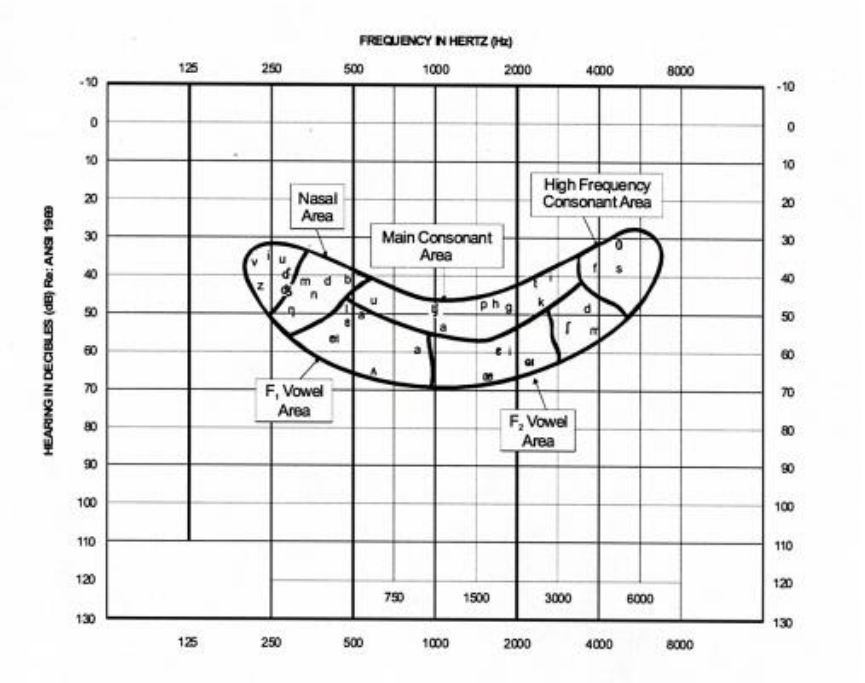


Figure 2.8 Speech banana (EllenBR, 2010)

2.5.4 Word formation/morphology

In developing word lists, it is important to understand the principles of word formation in the language concerned. A speech sound can be identified through intrinsic and contextual factors. Intrinsic factors are contributed by the acoustic properties of the sound and its frequency occurrence given by the phonemes contained in that sound (Zakrzewski et al., 1975). The acoustic properties of neighbouring sounds, the conditional probability relations between the phonemes and the probability of the transmitted word contribute towards the contextual factor of speech sound identification (Zakrzewski et al., 1975). The conditional probability between phonemes of a language can be understood through the principles of word formation of that language. This is particularly important in developing a word list that includes non-sense words, as the words, although carries no meaning, has to sound 'believable' in that language. A study of word formation is also equally important in determining the form of words to be included in the word lists, i.e. monosyllables vs bisyllables; CV vs CVC etc.

2.5.4.1 Word formation in Malay

There are certain rules for word formation in Malay, dependent on the number of syllables (Teoh, 1994). This section discusses the structure and word formation rules in bisyllabic Malay words, the word structure selected in the word lists.

Teoh (1994) had summarised the structures of bisyllabic words in Standard Malay (SM). There are 9 syllable combinations, all of which listed below:

1. V-CV
2. CV-CV
3. CVC-CV
4. VC-CV
5. V-CVC
6. CV-CVC
7. CVC-CVC
8. CV-V
9. CV-VC

V – vowel, C - consonant

(Teoh, 1994, pp. 14-15)

To generate a non-word in Malay, the phonological rules that govern Malay pronunciation should be understood. According to Abdul Rahman (1988), apart from the syllable combinations stated above, there are 17 other phonological rules that can be applied to SM:

1. All vowel phonemes may be present in the initial and middle positions of a word
2. Only vowels [i], [u] and [ə] can be present in the final position of a word. According to the author, the only time [e] appears in the final position is in the word /kole/
3. Only vowels [e], [o] and [a] can support a final closed syllable. Phonemes [i] and [u] may also support final closed syllables, but only in onomatopoeic words.
4. If [i] or [u] are used to support a final open syllable, the preceding syllable are never supported by [e] or [o]
5. If the final closed syllable of a word supported by vowel [a] is preceded by vowels [e] or [o], the vowels [e] and [o] will elect respective allophones [ɛ] and [ɔ] instead.

6. However, allophones [ɛ] and [ɔ] will never substitute [e] and [o] in a final closed syllable
7. All consonants in SM may appear in the initial and middle positions of a word
8. All consonants in SM may appear in the initial position of a syllable ie. underlined consonants in C̲V and C̲VC
9. Only consonants [k], [m], [n], [ŋ], [s], [h], [r] and [w] appear in the final position of a syllable (underlined consonant in CVC̲). Phoneme [ɲ] may close a syllable, but only if the syllable is not the final syllable. On the contrary, phonemes [p], [t], [l] and [y] may only close a syllable in a final syllable of a word
10. Plosives and affricates in the final position of a syllable will elect the stop allophones of themselves
11. Phonemes [b], [d], [g], [ǰ], [č] and[ɲ] do not appear in the final position of a word
12. Phonemes [b], [d], [g], [ǰ] and [č] do not appear as the final consonant in a closed syllable
13. There is no consonant clusters in SM
14. Long vowels are not present in SM
15. Long consonants (as found in Arabic) are not present in SM
16. Nasalisation occur when a vowel is preceded by a nasal consonant

There are a few discrepancies with the rules stated by Abdul Rahman (1988). With respect to the rule regarding the presence of vowels in the final position of a word (point 2), the author missed to include the vowel [a]. Although rare, [a] does appear in final position of a word, as illustrated in the word /bola/ (‘ball’). There is also the case of [k], which, in my opinion, will be replaced with a glottal stop [ʔ] if it is positioned at the final position of a CVC syllable or word. The occurrence of glottal stop was discussed in detail by Maris (1980), who stated that confusion might arise as ‘k’ used in the official orthography symbolises both the voiceless velar plosive consonant [k] and the glottal stop [ʔ]. Examples of glottal stop in the final position of CVC words are ‘rak’ /raʔ/ (‘shelf’) and ‘dek’ /deʔ/ (‘by’). ‘Dek’ can also represent /dek/ (‘deck’), an example of the confusion that may arise from the consonant [k].

2.5.5 Phonemic and phonetic balance

In the effort to replicate the sounds of everyday language, the construction of word lists was based on the phonetic composition of the language concerned (Eldert and Davis, cited in Martin et al., 2000). Phonetic or phonemic balance is an issue frequently discussed when developing speech audiometry material. According to Causey et al. (1984), phonetic and phonemic balance are two different concepts. Phonetic balance refers to the occurrence of phonemes in each word lists that mimic that of the frequency of occurrence of phonemes in the representative sample of the language. Phonemic balance, on the other hand, indicates that the frequency of occurrence of phonemes is equal for every list in the set.

Phonetic balance is achieved by having a speech audiometry test material to contain the phonemic composition that is equivalent to what is found in everyday speech (Lyregaard, 1997). The phonetically balanced test material should reflect everyday speech qualitatively, i.e. contains all the phonemes found in the spoken version of that particular language; and quantitatively, that is, the percentage of occurrence of the phonemes in the test material should be similar to that in everyday speech of the language (Lyregaard, ibid.). The reason for this balance is that the handicap experienced in hearing (or, more correctly, not hearing) a particularly infrequent phoneme would be less that missing a common phoneme in that particular language. There are many word lists that employ phonetic balance in their set (Egan, 1948; Hirsh et al., 1952; Han et al., 2009; Neilsen and Dau, 2009; Fu et al., 2011)

Phonemic balance is also widely used in speech audiometry materials (Boothroyd, 1968; Hong, 1984; Lau and So, 1988). AB word lists used the term ‘isophonemic’ instead of phonemic balance, however, the concept is the same (Boothroyd, ibid.). An advantage of using phonemic balance instead of phonetic balance is that it allows for shorter lists while still having phonemic content that are representative of the language (Boothroyd, ibid.).

Some speech audiometry material developers, on the other hand, did not regard phonetic or phonemic balance in the design of their word lists (Nissen et al., 2005; Harris et al., 2007). The effect of phonetic and phonemic balance on speech reception threshold was studied by Martin et al. (2000); it was found that phonetic/phonemic balance did not significantly affect the performance of listeners. It was concluded that having similar distribution of phonemes in the word lists as in daily speech does not make speech

recognition easier. The current study opted for the phonetic balance approach, considering that the available speech reception threshold test materials in Malay are either phonemically-balanced (Mukari and Said, 1991) or has debatable phonetic balance in the lists (Hong, 1984). Phonetic balance is also being used as a measure of validity of the word list (Wang et al., 2007); therefore, the employment of phonetic balance in the current study is anticipated to add to the validity of the word lists.

2.5.6 Speech material selection criteria

Almost all speech audiometry material utilised the same selection criteria outlined by Hudgins et al. (1947). As time progresses, the criteria were expanded and/or modified in order to improve the selection strategy and meet certain test objectives. However, the criteria specified by Hudgins et al. (1947) are featured repeatedly in the development of speech audiometry material, even the recent ones.

Hudgins et al. (1947) has given four criteria that were deemed as essential in speech item selection for speech audiometry. These criteria served as a guideline to other developers of speech audiometry material.

a) Item familiarity
b) Phonetic dissimilarity
c) Normal sampling of English sounds
d) Homogeneity with respect to basic audibility

Item familiarity is important in a population where there is a wide range of levels of education and social standards (Ashoor and Prochazka, 1982). Most speech test materials employ familiar items in their lists (Ashoor and Prochazka, ibid.; Wang et al., 2007; Fu et al., 2011). The CNC monosyllables in the NU-6 word lists, however, are built from a compilation of words with 7 stages of familiarity (Tillman and Carhart, 1966).

Item familiarity can be determined in several ways. Wang et al. (2007) and Fu et al. (2011) utilised publications on common words and word frequency in Mandarin. Ashoor and Prochazka (1982) sourced their test items from primary school books, daily newspapers and children's story books to ensure that the test items are familiar to the wide range of social and educational status in Saudi Arabia. The test items were further

refined using a word familiar rating on a four-point scale ('very familiar', 'familiar', 'somewhat familiar' and 'not familiar'), with words which are 'somewhat familiar' and 'not familiar' to a certain degree excluded from the list. The familiarity of CNC monosyllables in the NU-6 word lists was rated differently. Instead of having the familiarity rated according to participants' perception, the rating was based on the frequency of occurrence. This type of familiarity rating was also applied by Hirsh et al. (1947) in their W-22 word list. Hirsh et al. (ibid.) referred the familiarity rating of their test items on publications on word frequency by Thorndike (1932) and Dewey (1923) while Tillman and Carhart (1966) referred theirs on a publication by Thorndike and Lorge (1944).

Words that do not share phonetic elements, such as rhyming, are said to have phonetic dissimilarity (Nissen et al., 2011). Hudgins et al. (1947) argued that, as speech audiometry is actually a test of speech intelligibility and not intelligence or vocabulary, it is necessary that the test item should be simple and familiar. It is also essential to have a list of words that would not elevate the need of the listener's discrimination skills further than necessary, as having a choice of words that are minimally different to each other will increase the test difficulty but not increase the effectiveness of the test. Having similar words in a list will increase the difficulty of test without increasing its effectiveness. Similar words would demand finer discrimination, which is not part of the test objective. This principle was applied in the development of PAL Auditory Test No. 9. Dissyllabic spondees were chosen instead of monosyllables as monosyllables have greater possibility for phonetic similarity and thus make them unsuitable. However, not all speech audiometry developers apply this criterion in their test materials. The W-22 word list by Hirsh et al. (1947) has several minimal pairs ('ten'/'tan', 'an'/'as', 'yet'/'yes').

Hudgins et al.(1947) also introduced the concept of normal sampling of English sounds, defined as normal representation of the speech sounds of English language in speech audiometry material. The concept is further developed into phonetic balance, which is defined as having the frequency of vowels and consonants in word list reflect their frequency in the connected text (Ashoor and Procazka, 1982). This also gives rise to development of speech audiometry in local languages.

It is important to keep the word lists of a speech recognition test to be homogenous as it increases precision of the probability of the performance-intensity function. Basically, it can be achieved by either choosing test items that reaches the ear at the same level of

amplification when spoken in a similar tone of voice, or adjusting the intensity level of test items during recording so that they are heard at the same level of reproduction, or both (Hudgins et al., 1947). Based on this, several techniques have been developed to achieve homogeneity, such as psychometric equivalence technique (Nissen et al., 2011) and digitisation of speech lists (James et al.,1991). These techniques are further discussed in Section 4.2.2.

2.5.7 Lexical category and morphological similarity

Morphological similarity of test items, which indicates the similarity in the lexical category of words, is one of the criteria that had been considered in the selection of words to be included in a set of speech audiometry material (Ashoor and Prochazka, 1982).

Word lists that were developed by Ashoor and Prochazka (1982) were made up of all nouns, as it was thought that nouns "…best serves the purpose of the speech test" (Ashoor and Prochazka, ibid.). However, it seemed that morphological similarity was not a crucial issue as other developers of speech audiometry material did not include it in their criteria for item selection (Peterson and Lehiste, 1962; Boothroyd, 1968; Mukari and Said, 1991). Northwestern University Auditory Test No. 6 (NU-6) which utilised CNC monosyllabic words, used a combination of lexical categories such as nouns, verbs and adjectives in their word list (Tillman and Carhart, 1966). Hirsh et al. (1952) went as far as having conjunctions ('or') and adverbs ('not', 'by') in their lists. There is no study found comparing the efficacy between nouns and other lexical categories. One lexical category that was specifically studied as speech audiometry material is digits. Miller et al. (1951) studied the effects of different types of test material and compared the performance-intensity functions of digits, sentences formed with five words connected by auxiliaries and non-sense syllables. It was found that, at response level of 50% correct, digits require the lowest signal-to-ratio (SNR) (-14dB) while nonsense syllables require the highest (+3dB). It was believed that the response was made through the process of elimination, i.e. the correct responses were determined by the characteristics of the stimuli that could have occurred but didn't, of the alternative answers available in the range of that particular test items. Digits are relatively easy given by the narrow range of possible answers as well as the distinctive phonetic content of each item. Almost all vowels in digits 'one' to 'nine' are different from each other ('five' and 'nine' contain the same vowel). These limits are less applicable to nonsense words; therefore the listener

has to be able to perceive each phoneme in the test item in order to make the correct response.

Based on the findings on the effect of lexical categories on the performance in speech audiometry, it is thought that having more than one lexical category in a word list has no significant effect on the performance in speech audiometry. For a set of speech audiometry material that has more than one word list, having all-digits lists would be impractical. Inclusion of digits in a list made up of words from a mixture of several lexical categories removes the effect of the range of possible answers and, therefore, shall not affect the performance of the listener.

2.6 Considerations in current research

Most speech audiometry materials, especially those targeted to measure speech reception thresholds, employ familiar words. Utilisation of familiar words ensures that the auditory recognition ability is assessed without having to put unnecessary stress on finer discrimination skills. Familiar words are also more suitable for wider range of literacy levels compared to unfamiliar words, in addition to showing better homogeneity.

Nonsense syllable tests on the other hand have their own advantages. They allow assessment of discrimination skills with lesser cues compared to familiar words. Practice and learning effects of are also lower for nonsense words, allowing the lists to be used more times compared to familiar word lists. The flexibility of phoneme combination in nonsense words allow for more speech sound to be included in shorter lists.

Daily speech, unlike most speech recognition test material, is made up of both familiar and unfamiliar words. To test the speech recognition of hearing impaired listeners solely on familiar words is not representative of their daily communication needs. However, information on how listeners, normal hearing and hearing impaired, perform using word lists with mixed familiarity is unknown. How words with mixed familiarity affect the performance of listeners, in terms of speech hearing threshold and speech discrimination, needs to be explored. The clinical feasibility and efficacy of mixed familiarity word lists also need to be studied. The current study considers the addition of nonsense words, instead of unfamiliar words, into the design of the word lists. The use of nonsense words would greatly reduce the uncertainty on the familiarity level of unfamiliar words.

2.7 Research questions, aims and objectives

This study aims to develop and produce a bisyllabic Malay speech recognition test word lists for adult Malay speakers. Several research questions arise following the literature review. What Malay phonological and phonetic features should be included in the Malay speech recognition test word lists? How can the acoustic properties of the word lists be validated further than measuring the difficulty in audibility/difficulty of test? How would adult Malay speakers, both normal hearing and hearing impaired, perform using word lists that contain both meaningful and nonsense words? Would speech audiometry material consisting of meaningful and nonsense words be able to reflect the speech hearing and discrimination abilities of its listener?

The aim of the current study is to develop a bisyllabic speech reception threshold (SRT) test word lists. The word lists consist of a mix of meaningful and nonsense Malay words. Three specific aims are set to help answer the research questions. The first aim is to produce a phonetically balanced bisyllabic Malay word lists. The lists are to contain both meaningful and nonsense words in order to simulate the wide range of word familiarity in everyday speech. In order to achieve this aim, the objectives of the current study are to develop a word corpus, and analyse the distribution of phonemes, word familiarity and phonetic balance in order to assemble the set of word lists. Secondly, the current study aims to verify the word lists to ensure homogeneity and consistency among the lists. Two aspects of homogeneity and consistency are explored, difficulty of test and acoustic content. In previous studies, only homogeneity and/or consistency of difficulty of test were measured (Lau and So, 1988; Mukari and Said, 1991; Nissen et al., 2007; Wang et al., 2007; Nissen et al., 2011). The current study intends to study the feasibility of using acoustic properties of the word lists in consistency measurement. To achieve this aim, pure tone thresholds and speech audiometry thresholds using the developed word lists of normal hearing Malay speakers are used to establish the homogeneity and consistency of difficulty of test. Measurement of the consistency of acoustic content is a new concept that is explored by the current study. The research objectives to achieve this aim are establishing the Long Term Average Speech Spectrum (LTASS) of the Malay language followed by establishing the consistency of acoustic content between lists and between the lists and LTASS. The third aim is to clinically validate the word lists in terms of its ability to distinguish different types and levels of hearing loss. The research objectives in order to achieve the third aim are to recruit normal hearing and hearing-

impaired adult Malay speakers and establish the characteristics of their speech audiometry curves, also known as performance-intensity (PI) functions. The speech audiometry curves are then used to establish the speech reception thresholds and discrimination scores, and establish the relationship between the speech audiometry curves and pure tone thresholds. The summary of the specific aims and objectives are given in Table 2.6.

Table 2.6 Specific research aims and objectives

Specific aims	**Objectives**
Produce a phonetically balanced bisyllabic Malay word lists using combination of meaningful and nonsense words	Develop a Malay word corpus
	Establish the distribution of phonemes based on the word corpus
	Establish the frequency of occurrence of familiar words based on the word corpus
	Establish phonetically balanced Malay word lists
Verify the bisyllabic Malay word lists in two main aspects; the consistency and homogeneity of the word lists in terms of difficulty of test, and the consistency of the word lists in terms of acoustic content	Recruit normal hearing volunteers and native Malay speakers
	Perform hearing assessments – pure tone audiometry and speech recognition test
	Establish the consistency and homogeneity of the word list based on the speech recognition test results
	Establish Malay long term average speech spectrum (LTASS)
	Establish the consistency of the acoustic content of the word lists and the Malay LTASS
Clinically validate the bisyllabic Malay word lists in two main aspects; whether the word lists are able to reflect the different types of hearing conditions, and whether the word lists are able to reflect the hearing level	Recruit normal hearing volunteers and hearing impaired patients
	Perform hearing assessments – pure tone audiometry and speech recognition test
	Establish the characteristics of speech audiometry curve
	Establish correlation between speech audiometry threshold and the pure tone threshold

CHAPTER 3 DEVELOPMENT OF BISYLLABIC MALAY WORDLISTS

3.1 Introduction

The development of word lists consisted of a number of steps; selecting the word sources from which the words would be selected and determining the type or types of words that will be used, deciding whether or not to have phonemic balance, as well as number of words and number of lists in the set. The objectives of this part of the study are to establish the distribution of phonemes and the frequency of occurrence of familiar words in Malay language, both in order to develop a set of 10 phonetically balanced word lists in Malay language. Each list contains a mix of 10 meaningful words and 5 nonsense words to be used to test speech reception threshold and word recognition score in Malay adult speakers. The reason for incorporating nonsense words in the lists is to simulate the variety of word familiarity in everyday speech. In addition, the word lists can be shortened to contain 10 meaningful words only while still maintaining the phonetic balance. These lists will ultimately be used for speech reception threshold test in adult Malay speakers.

The drafted word lists would then be tested for homogeneity in terms of difficulty and phonemic balance.

This chapter includes the review the methods available in the development of speech audiometry word lists, an outline of the method utilised, the results obtained in this study as well as a discussion of the findings.

3.2 Review of methods

This section provides a review on the previous methods used in the development of speech audiometry material. The review includes alternatives on word choice, how phonetic balance is achieved, speech material selection criteria, methods of word selection, and methods of construction of word lists. These issues are important in the design of speech audiometry word lists, and are arranged according to the order of the development process.

3.2.1 Word source

In the construction of word lists, one of the major issues is the selection of the speech material. The type of speech that can be used in the construction of the word list may be phonemes, syllables, words or even sentences. Words can range from monosyllabic ones to polysyllabics. To select which speech structure is the best suited to the objective of the test, several factors have to be considered: target age, redundancies, scoring of responses, relation to 'everyday' or 'real' speech, and test duration (Lyregaard, 1997; Mendel, 2008). The word lists may be constructed from scratch, that is, the words were collated from another word sources; or adapted or revised from previously established word lists (Peterson and Lehiste 1962; Boothroyd 1968; Tillman and Carhart 1966; Hirsh et al. 1952; Alusi et al., 1974; Ashoor and Prochazka, 1982; Nissen et al., 2011).

For SRT tests, common sources include published word corpora, including electronic corpora, conversation studies, dictionaries and reading materials such as textbooks, story books and daily newspapers (Hirsh, et al., 1952; Peterson and Lehiste, 1962; Wilson, et al., 1976; Ashoor and Prochazka, 1982; Mukari and Said, 1991; Nissen, et al., 2011). For developers who were not reliant on sources of words, the test words might be derived from phoneme matrices (Lau and So, 1988). There is no definite source of words for SRT based on the review of research designs; however, development designs are mainly divided into three categories. The first category is the word lists that were built based on readily available word corpora, such as CNC lists (Peterson and Lehiste, 1962). Another group of word lists is adaptations of previously published SRT lists, such as AB word lists and CID Auditory Test W-1 lists (Hirsh et al. 1952a; Boothroyd 1968). The third category of word lists is lists that were built based on words selected by the developers (Ashoor and Prochazka, 1982; Nissen, et al., 2005; Harris, et al., 2007; Nissen et al. 2007; Nissen, et al., 2011). Most of the recent, non-English SRT word lists were built this way as the publication word corpora, especially those that include word frequency, were not as extensive as those found in the English language.

In the current study, the design of the construction of word lists falls in the third category. There is no published word corpus in Malay language; therefore, a selection of words to be included in the word list would have to come from another source. In this case, the sources are daily newspapers. The justification in using daily newspapers was that they are accessible and provide words that are familiar to most people. Several previously developed speech audiometry materials were also constructed based on the words

sourced from daily newspapers or corpus that used daily newspapers as sources (Ashoor and Prochazka, 1982; Nissen et al., 2005b; Nielsen and Dau, 2009).

3.2.2 Phonemic/phonetic balance

There are two types of phoneme distribution in speech audiometry material, phonetically balanced and phonemically balanced.

Phonetic balance is achieved by having a speech audiometry test material to contain the phonemic composition that is equivalent to what is found in everyday speech (Lyregaard, 1997). The phonetically balanced test material should reflect everyday speech qualitatively, i.e. contains all the phonemes found in the spoken version of that particular language; and quantitatively, that is, the percentage of occurrence of the phonemes in the test material should be similar to that in everyday speech of the language (Lyregaard, ibid.). The reason for this balance is that the handicap experienced in hearing (or, more correctly, not hearing) a particularly infrequent phoneme would be less than missing a common phoneme in that particular language.

Phonemic balance refers to the distribution of phonemes that is equal between the lists in a set of speech audiometry material. However, unlike phonetic balance, the distribution of phonemes in the lists does not reflect the actual distribution of phonemes occurring in the language (Mendel and Danhauer, 1997).

There are, however, views that oppose to this issue: Hudgins, et al (1947) stated that a normal representation of the speech sound is not particularly important in the threshold measurement of the hearing of speech. This view was supported by the finding of Martin et al. (2000) which stated that there was no clinically significant difference found between the word recognition scores measured using phonetically balanced word lists and those using lists made up of randomly chosen words.

The current study opted to employ phonetic balance as it adds to the validity of the word lists (Wang et al., 2007).

3.2.3 Speech material selection criteria

Almost all speech audiometry material utilised the same selection criteria outlined by Hudgins et al (1947) and discussed in the previous chapter. As time progresses, the

criteria were expanded and/or modified in order to improve the selection strategy and meet certain test objectives. However, the criteria specified by Hudgins et al (1947) are featured repeatedly in the development of speech audiometry material, even the recent ones (Hong, 1984; Mukari and Said, 1991; Ashoor and Prochazka, 1992).

Item familiarity is set as a criterion due to the objective of speech audiometry, which is to measure speech intelligibility instead of vocabulary or intelligence (Hudgins et al., ibid.). Further discussion on item familiarity and how it is determined in relation to the current study is given in Section 3.2.4. Phonetic dissimilarity refers to having test items that are dissimilar in terms of sound. Having test items that are minimal pairs, such 'sun' and 'bun' or 'eyeball' and 'highball', put extra demand on the test as they require finer speech discrimination. Phonetic dissimilarity is applied in the word selection criteria with the justification that speech audiometry is aimed to measure speech hearing threshold and not finer speech discrimination (Hudgins et al., 1947). Normal sampling of English sounds refers to the representation of English phonemes in the speech item selection. This is supported by Lehiste and Peterson (1959) who extended the premise to phonetic balancing. Homogeneity signifies equal audibility among the words used in a test. Homogeneity of the test items is advised due to two reasons; firstly, the constant amount of random error ensures precision in determining speech audiometry threshold based on the performance-intensity function, and secondly, it allows the words to be grouped into several equally audible lists.

In the current study, phonetic dissimilarity is kept to a minimum. The normal sampling of Malay words is applied in the word lists, with the representation of most of Malay phonemes. The distribution of phonemes included in the word lists in the current study is further discussed in Section 3.5.3. Homogeneity of the word lists is also considered. Homogeneity testing of the word lists is discussed in Chapter 4.

3.2.4 Item familiarity

Item familiarity is important in a population where there is a wide range of levels of education and social standards (Ashoor and Prochazka, 1982). Most speech test materials employ familiar items in their lists (Ashoor and Prochazka, ibid.; Wang et al, 2007; Fu et al., 2011). The CNC monosyllables in the NU-6 word lists, however, are built from a compilation of words with 7 stages of familiarity (Tillman and Carhart 1966).

Item familiarity can be determined in several ways. Wang et al (2007) and Fu et al. (2011) utilised publications on common words and word frequency in Mandarin. Ashoor and Prochazka (1982) sourced their test items from primary school books, daily newspapers and children's story books to ensure that the test items are familiar to the wide range of social and educational status in Saudi Arabia as well as the population originating from different cities in Saudi. The test items were further refined using a word familiar rating on a four-point scale ('very familiar', 'familiar', 'somewhat familiar' and 'not familiar'), with words which are 'somewhat familiar' and 'not familiar' to a certain degree excluded from the list. This method was designed to cater for the variety of literacy levels as well as possible difference in word familiarity for people who come from different areas of the country.

The familiarity CNC monosyllables in the NU-6 word lists were rated differently. Instead of having the word familiarity rated according to participants' perception, the rating was based on the frequency of occurrence (Tillman and Carhart, 1966). This type of familiarity rating was also applied by Hirsh et al (1952) in their W-22 word list. Hirsh et al referred the familiarity rating of their test items on publications on word frequency by Thorndike and Dewey while Tillman and Carhart (1966) referred theirs on a publication by Thorndike and Lorge.

Zakrzewski, et al. (1975) studied the effects of word meaning on the identification and discrimination of speech. Monosyllabic meaningful Polish words and nonsense words, in separate sets of lists, were used as stimuli in the speech audiometry of 297 normal hearing children. Responses were written. The study found that the sound recognition scores for nonsense words were significantly lower than that of the meaningful words. The authors attributed this difference in score to the lesser probability of correct guessing and lower effects of word context in nonsense words. The difference in word recognition thresholds was not noted.

Zakrzewski, et al. (1975) had also studied the errors made in the phoneme recognition in the two sets of lists. They found out that, in meaningful words, most of the mistakes happened within the group of the phonemes' distinctive feature (DF). For example, in consonantal DF when errors in consonantal (+consonant) and nonconsonantal (−consonant) phonemes were studied, most consonants were mistaken for another consonant (within-sign) while none of the non-consonantal phonemes were mistaken for a consonant (across-sign). An analysis of across-sign errors made in nonsense words showed that the patterns of mistakes are different compared to the meaningful words.

While there were no across-sign mistakes in the nasal DF group of meaningful words, the nonsense words showed otherwise. In the voiced DF group, unvoiced phonemes were mistaken for voiced phonemes many more times than the vice versa.

The authors did not provide any conclusions from these findings as the study were still ongoing. It was also not stated whether the nonsense words were adherent to the phonetic rules of the Polish language. The findings have also raised the question of the difference in error patterns between nonwords that follow the languages phonetic rules and the nonwords that do not.

The current study proposes to utilise both familiar and nonsense words in the development of the word lists. The familiarity rating is based on the frequency of occurrence method used by Tillman and Carhart (1966), as the words in current study are sourced from an online corpus of newspaper texts.

3.2.5 Selection of words

There are many methods utilized by speech test developers to construct their word lists. Several criteria are applied in the selection of words that is to be included, among them:

1) familiarity of words (Zakrzewski et al. 1976; Ashoor and Prochazka, 1982; Martin, 1997; Lyregaard, 1997; Harris et al, 2007; Wang et al, 2007; Han et al, 2009;),

2) phonetic balance or phonemic balance (Zakrzewski et al, 1975; Ashoor and Prochazka, 1982; Lyregaard, 1997; Wang et al., 2007; Han et al, 2009)

3) morphological similarity (Hirsh et al. 1952; Ashoor and Prochazka, 1982)

4) representation of normal sampling of spoken language (Ashoor and Prochazka, 1982)

5) homogeneity in reference to intelligibility (Ashoor and Prochazka, 1982; Harris et al, 2007),

6) frequency of use in spoken language (Harris et al., 2007) and

7) structural balance (Zakrzewski et al., 1975).

The current study employs both meaningful and nonsense words in the lists. Familiarity assessment is applied on the meaningful words, using the frequency of occurrence as the mark of familiarity. Normal sampling of the language and phonetic balance is kept in

each list in the set in order to represent the distribution of Malay speech sounds found as well as to increase the validity of the word lists. The morphological similarity, however, is not applied in the current study as it is thought that having one word form, e.g. nouns or verbs, will not affect the effectiveness of the word lists. On the other hand, the structural balance of the words is set throughout the lists; only words with consonant-vowel-consonant-vowel structure are used in the speech audiometry set.

3.2.6 Construction of word lists in non-English languages

This section discusses the approaches in constructing speech audiometry word lists in non-English languages. A number of non-English word lists were constructed quite recently (post-year 2000), and Lau and So (1988) presented a detailed method of constructing speech audiometry material in Cantonese, using phonetic principles. They tried to address the problem. They applied the word selection criteria suggested by many previous developers of speech audiometry material (Hudgins et al., 1947; Zakrzewski et al., 1975; Ashoor and Prochazka, 1982): these include equal average difficulty, equal range of difficulty among individual lists (i.e. homogeneity); inclusion only of words in common usage (word familiarity) and equal phonetic composition in reference to the Cantonese language. They have also included a condition suited to Cantonese language; that is, the words are monosyllables.

Lau and So (1988) outlined the Cantonese linguistic characteristics related to developing a speech audiometry word list. They opted to confine the syllable structure only to consonant-vowel (CV) in their word list. Although CV is not the most frequent syllable structure in Cantonese (consonant-vowel-consonant or CVC is), the selection was made on the basis of the ease of achieving phonemic balance across individual list.

Some Cantonese phonemes only appear in certain positions in a syllable. To overcome this problem of phonemic position, Lau and So (1988) produced an initial consonant-vowel nucleus matrix made of 19 consonants against 17 vowels that can be found in Cantonese (in this case, diphthongs are also considered as vowels). Only the vowels and consonants that can form CV words in the most possible ways are chosen. This eliminates the phonemic position problem. However, this might also eliminate final-position phonemes, which might have the possibility of being highly frequent phonemes in the language.

The shortlisted vowels are vowel nuclei that form words with most of the consonants i.e. the most frequently used vowel nuclei (in CV form words). Top 10 vowel nuclei were selected, plus 2 next frequently used vowels as 'reserves'. One vowel nucleus, diphthong /oey/ were excluded even though it has the same frequency as the vowels on the 9th and 10th places, as the developers only wanted to include diphthongs that were made of pure vowels included in the top 10 list.

The consonants were chosen on the same basis as the vowels, which were the top 10 most frequently used consonants in CV-form words. Allophones in daily speech, even though they have the same frequency, were also excluded, e.g. /l/ and /n/, by which only /n/ was included in the short list and /l/ was not. Consonant /k/ was not included due to its acoustical variety, even though it was one of the most frequently-used consonants. Consonant /ts/ was also found to have high frequency but was put into the reserve list as its acoustic properties are similar to /t/ and /s/, both chosen vowels, put together.

The final list of phonemes comprise 10 vowels (plus 2 reserves) and 10 consonant (plus 1 reserve). The list was crosschecked with a previous list of most frequent Cantonese phoneme, and it was found that the Lau's and So's (1988) list did exclude some of the most frequently used phonemes.

In applying the above method of phoneme selection, Lau and So (1988) had succeeded in producing 100 individual words to be used in their word lists. Four lists contained all phonemes and consonants shortlisted, other 4 contained one word made up of one of the reserved phonemes (consonant or vowel) and the remaining 2 contained 2 words each made up of 1 of the reserved phoneme. The lists were deemed interchangeable due to the almost equivalent occurrence of phonemes in each list (Lau and So, ibid.).

However, due to the method of choosing the vowels and consonants that could make the most CV-structured words, Lau and So had missed to include several of the most frequently-occurring vowels and initial consonants of that language, which meant that the list was not strictly phonemically-equivalent and representative of the speech sounds of the Cantonese language. This might have affected the assessment of the true ability of the listener to discriminate speech sounds, and therefore, words.

Ashoor and Prochazka (1982) looked into the development of speech audiometry material that suits the Arabic speakers of Saudi Arabia. Two of their main concerns are the educated and uneducated groups of Saudi population, and the two different forms of language used in Saudi Arabia, colloquial Arabic and modern standard Arabic. To ensure

that the speech audiometry material addresses these concerns, word familiarity, word homogeneity with respect to intelligibility, morphological similarity, language representation and phonetic balance were considered in the word selection. Technical and scientific jargons as well as words deemed as difficult are excluded from the list. All chosen words are in the same lexical group, which are nouns. To address the difference between colloquial and modern standard (also known as classical) Arabic, words which have similar forms in both are chosen whenever possible.

Word selection was done by sourcing monosyllabic words from primary school books, daily newspapers and children's story books. The words were then evaluated by a selection of medical students from the local university and their family members through familiarity rating, with words rated as 'not familiar' and 'somewhat familiar' by a certain extent being excluded. The remaining words were divided into groups of 20, each containing even representation of phonemes and syllable types with each other (Ashoor and Prochazka, 1982).

3.2.7 Construction of word list in Malay

There are two previous studies on the development of speech audiometry materials in Malay language (Hong, 1984; Mukari and Said, 1991). Hong (ibid.) aimed to develop a short word lists that reflect the natural usage of Malay-speaking population in Singapore, Malaysia and Brunei for the purpose of speech recognition audiometry. The set consisted of 10 different lists with 10 bisyllabic words. The words were chosen based on familiarity, equal average difficulty, equal range of difficulty, representation of Malay language, and common usage. Hong (ibid.) decided not to have his word lists phonetically balanced. However, he claimed that each list had 20 phonemes and contained a 'rough approximation' of phonetic balance found in everyday spoken Malay (Hong, 1984). There was no indication of the source or materials used to build the list. The method of measuring familiarity, and language usage and representation were not indicated as well. Upon reading the lists, there is a possibility that Hong (1984) meant to have 20 syllables instead of 20 phonemes in each list. The lists contain bisyllabic words which are CV-CV, CV-CVC, V-CV and V-CVC in nature. However, not all structures were included in each list. The phonemes included in the set were vowels [a], [i], [u], [ə] and [o], and consonants [b], [č], [d], [g], [h], [j̆], [k], [l], [m], [n], [p], [r], [s], [t], and [ʔ]. All lists failed to incorporate each of the phonemes; therefore, the representation of phonemes between each list is not the same.

The equal word list difficulty was tested through analysing the mean scores heard at a constant, near threshold intensity of each list. The scored were compared using the analysis of variance, and showed no significant differences between the lists.

Mukari and Said (1991) developed another bisyllabic word lists for the use of speech recognition audiometry. The word lists, unlike Hong's (1984), are phonemically balanced. There are 25 lists with 10 words each. All of the words are of CVCV structure, as it was found to be the most common syllable structure in Malay language. Four hundred and fifty commonly-used words were pre-selected from the Dewan Bahasa dictionary. The familiarity of these words was assessed by 150 Malaysian adults of different racial backgrounds, resulting on 196 being shortlisted for further use.

To determine phonemic balance, Mukari and Said (1991) considered the Malay phonemic system to have 25 consonants (actually 26, counted based on the phonemes listed in the article), 6 short vowels and 6 diphthongs. The term 'diphthong' may have been misrepresented here as diphthong is defined as 'a vowel sound, occupying a single syllable, during the articulation of which the tongue moves from one position to another, causing continual change is vowel quality' (Collins English Dictionary, n.d.), whereas [ia], [ua], and [io] are actually vowel sequences (Teoh, 1994). Due to low usage in Malay, [f], [ʃ], [z], [oi] and [io] were excluded. Mukari and Said (1991) had also excluded the phonemes [ð], [ɣ], [x] and [θ] on the basis that they are actually pronounced as [d], [g], [k] (also [h]) and [s] respectively. This indicated that the 38 'phonemes' that was listed were actually grapheme, which is the unit of writing that represents one phoneme.

Contrary to the word lists developed by Hong (1984), the developers of these word lists had included the 28 selected phonemes in each list. Twenty-five lists were compiled, each with 10 CVCV words selected from 196 words shortlisted earlier. Here, the phonemic balance was defined by the occurrence of all phonemes in almost the entire list. The frequency of occurrence, however, does not reflect the real-life usage, indicating that the word lists could be phonemically balanced or isophonemic instead of phonetically balanced.

The interlist intelligibility difference between the lists was also tested. The prerecorded word lists were presented to 25 normal-hearing participants at a constant intensity of 10 dB above their average hearing thresholds at 500, 1000 and 2000 Hz. The mean scores between lists were then compared and no significant differences were detected.

These two sets of word lists by Hong (1984) and Mukari and Said (1991) indicated that the development of speech audiometry in Malay was possible, and that the lists were able to demonstrate the hearing level for speech. However, the weaknesses of these two sets of word lists were that they are not actually phonetically balanced, and that the claims of phonetic balance were not verified.

The current study aims to deal with the weaknesses posed by the word lists developed by Hong (1984) and Mukari and Said (1991). The phonetic balance of the word lists is carefully determined based on the analysis of phonemic content of Malay words. The process of determining the phonetic distribution is discussed further in Section 3.3. The verification of the phonetic balance is discussed in Section 3.5.5.

3.3 Analysis of the phonemic content of Malay words with CVCV structure

The phonemic content of Malay words with consonant-vowel-consonant-vowel (CVCV) structure was studied using the analysis of phoneme distribution in the selected language sample. The method used in the phonemic analysis was a commonly used method in previous studies of phonemically-balanced word list development.

Language sample is taken from an online corpus with words sourced from daily newspapers. The selected words were then analysed for their phoneme distribution. The following sections describe the procedures for the study.

3.3.1 Purpose of the study

The purpose of this part of the study was to obtain the distribution of phonemes in Malay CVCV word sample. The distribution would be used towards the selection of words and their phonemic content in the final word lists.

3.3.2 Research design

The following section outlines the research design of the phoneme distribution study of Malay CVCV words. The research design is observational, starting with compiling Malay consonant-vowel-consonant-vowel (CVCV) words sourced from daily newspapers to form a CVCV word corpus and followed with quantitative analysis on the distribution of phonemes. The analysis of phoneme distribution of Malay CVCV words determines the

distribution of phonemes in the speech audiometry word lists in order to achieve phonetic balance.

3.3.2.1 Word Sources

Words were sourced from an online corpus built by Dewan Bahasa dan Pustaka (DBP), the national council for languages and libraries for Malaysia. The online corpus compiles words from selected newspapers, books, school textbooks, journals and articles published in Malaysia. The DBP corpus was selected as it was the only Malaysian Malay word corpus available in publication. It is accessible at http://sbmb.dbp.gov.my/korpusdbp .

3.3.2.2 Word collection

A word analysis was done on the online corpus. The online corpus allows users to do several types of word analysis – word frequency, word length and uppercase and lowercase letters analyses, among others.

To ensure that the words used as test items are words that are being used in daily life, they were sourced from two main Malay-language daily newspapers in Malaysia, Utusan Malaysia and Berita Harian, and their respective Sunday papers, Mingguan Malaysia and Berita Minggu. These newspapers are two of the highest selling daily newspapers in Malaysia (Audit Bureau of Circulation, 2011).

The period of publication of these newspapers was also taken into consideration. To ensure that the words used are contemporary, the experimenter has limited publication search to a period of 5 years, starting from 2006 and ending in 2010.

The output of the exercise resulted in a list of all words that had appeared in the newspapers for the stated period, their frequency of usage and their percentage. The output was then transferred into Microsoft Excel 2010 for further filtering.

Bisyllabic, consonant-vowel-consonant-vowel (CVCV) words were collected from the corpus.The collated words would be termed as CVCV-1. To limit the words to those used contemporarily, the experimenter had limited the word sources according to type of publication and date of publication.

The corpus filtering output was given in text documents, an output each for each of the publication (daily and weekend editions). Frequency of use and percentage of use of the

words were included in the output. These data were then transferred into Microsoft Excel 2010 for further analysis.

3.3.2.3 Phonemic analysis

A phonemic analysis of Malay words has to be done to in order to determine the phonemic content of the prospective word lists. Literature search produced no previously published studies on Malaysian Malay phonemic analysis.

For the purpose of this research, the phonemic analysis was done on the CVCV-1 collated from the word corpus. The analysis was done using Microsoft Excel 2010. CVCV-1 were filtered according to their (written) vowel contents (e.g. '?a?a', '?a?e', '?a?i' and so on), resulting in 25 lists. The words ending with the written vowel 'a' were further filtered down according to its phoneme, 'a' or 'ə'. Words containing the written vowel 'e' were also filtered manually according to their correct phoneme, 'e' or 'ə'. Kamus Dewan Edisi Keempat (2010), the dictionary published by the Dewan Bahasa dan Pustaka and used throughout the study, was referred to ascertain the proper pronunciation of the words.

The frequency of occurrence and the percentage of occurrence for each phoneme were then calculated using Microsoft Excel. The results were then tabulated according to individual phonemes.

In order to be able to reflect the proportions of phonemes in the prospective word lists, the proposed number of appearance for each phoneme was calculated. The number of appearance for vowels and consonants were calculated separately with two denominators – over 20 phonemes (for each vowels and consonants) for the 10 familiar words in each list and over 30 for the complete list (15 words, 10 familiar words plus 5 nonwords).

3.4 Analysis of the frequency of occurrence of Malay words with CVCV structure

In order to ensure that the words that would be included in the set of word lists were familiar words, an analysis of frequency for words was done on the CVCV-1. The collated words from both sources were ranked according to its frequency and percentage of appearance in the texts. The top 350 words from each source were extracted and further filtered to remove words that are dialectal, slang, proper nouns or sensitive. The results

from the two sources were then combined and any duplication between the two sources was deleted. The words from the final extraction formed the bank of words from where the final 15 bisyllabic, CVCV word lists were built.

3.4.1 Purpose of the study

The purpose of this part of the study is to establish the frequency of occurrence of CVCV words in Malay. The frequency of occurrence is used as the mark of familiarity of the words. The data would be used in determine which of the words to be included in the word lists.

3.4.2 Research design

3.4.2.1 Analysis of frequency of occurrence

To obtain the consonant-vowel-consonant-vowel words from the corpus, an analysis on word frequency was first done on the corpus. The frequency of occurrence facility in the DBP online corpus was utilised in this part of study.

The word sources and method of word collection were the same as section 3.3.2.1.

Two criteria were imposed on the analysis; the type of publication and the period of publication from which the words were originated. The type of publication was set to those that use formal or standard language as opposed to colloquial language. The use of standard language ensured that the words were familiar across the population and that dialectal words were not used. Publication that fell into this category include newspapers, textbooks and selected magazines. To further ensure that the data was contemporary, the period of publication was set to the most recent relative to the study.

The corpus database of Dewan Bahasa dan Pustaka (http://sbmb.dbp.gov.my/korpusdbp/SelectUserCat.aspx) was accessed. The Researcher→Word Analysis function was accessed to initiate the filtering options. Under the option of source material, newspapers 'Utusan Malaysia' and 'Berita Harian', including their Sunday editions, were selected. Due to the magnitude of the data to be analysed, analyses for each newspaper were done separately.

The database was filtered the same way as in section 3.4.1. The period of publication was filtered at 5 years, beginning from 2006 and ending on 2010. Options for dates and months, however, were not given in the filter at the time of analysis.

The resulting output was delivered in the form of text and subsequently transferred to Excel for further analysis.

3.4.3 Selection of words and nonwords for the word lists and building the word lists

The selection of words to be included in the list was subjected to the number of appearance of each phoneme in each list as well as its frequency. The words from each list were selected from the word bank (created after the analysis of frequency) and contained the number of phonemes equal to the proportion that was calculated in the phonemic analysis. However, there were a few restrictions:

Due to the small number of words in each word list (10 words, 5 nonwords, 15 test items in total), it was impossible to include all phonemes. Only phonemes with percentages of appearance >1.0% were included in the word lists

The small number of test items in each list also limits the number of appearance of each phoneme that has percentages >1.0%. Phonemes that were calculated to appear less than once in each list would not appear in every list.

3.4.4 Evaluation of phonetic balance

To ensure that the word lists followed the phoneme distribution found in the earlier language sample, the phonemes of the words in the list were tabulated and compared to the language sample. Paired sample T-Test was used to compare the phoneme distribution of the word lists with that of the corpus, as well as between the two versions of the word lists.

3.5 Results

3.5.1 Development of the speech material

The development of the speech material involves selecting the word structure for the words to be included in the word lists, determining the phoneme content for the word

structure, selecting the meaningful words to be included in the lists and producing the nonsense word to be included in the word lists.

Unavailability of any form word corpus in Malay meant that the researcher had to build a compilation of word in order to study the components of linguistics needed to build the word lists. As the word lists are designed to be phonemically-balanced, an analysis of percentage of phoneme content in Malay had to be done. Analysis of familiar words in Malay was also done in order to provide a selection of words suitable to be included in the word lists. The percentage of phoneme content was then utilised to determine the words, both meaningful and nonsense, that was to be incorporated in the final lists.

3.5.2 Development of corpus

There is a lack of published corpus in Malay; which prompted the building of a word corpus in this study. Utilising the Dewan Bahasa dan Pustaka (The Institute of Language and Literature), a government body responsible for regulating the Malay language in Malaysia, website, a CVCV word corpus was built using the in situ word analysis program. Two major daily newspapers were chosen as the word source – Utusan Malaysia and its Sunday publication, Mingguan Malaysia (UM/MM), and Berita Harian, with its Sunday publication Berita Minggu (BH/BM). Period of word retrieval was set between the years 2006 to 2010.

This process of developing the word corpus was designed based on several previous studies on speech audiometry material development (Ashoor and Prochazka, 1982; Harris et al, 2007; Nissen et al, 2011). Often, the word corpus in the studied language already exist, which allowed the researchers to just extract the words according to criteria (e.g. syllable, familiarity, frequency of occurrence). In this study, it was assumed that the online software by Dewan Bahasa dan Pustaka (DBP) would provide the quickest way of developing a corpus, as opposed to sourcing the words from printed matter. Further evaluation on the process of assembling the word corpus will be discussed in the next chapter.

There is no published protocol regarding the source, size of the corpora or the number of words from which the word stimuli can be drawn from. Ashoor and Prochazka (1982) started with a corpus of 168 monosyllabic words sourced from books and newspapers to produce a 120-word stimulus material. On the other hand, CNC word lists were developed based on a 1263-word corpus derived from a volume of 30 000 words (Lehiste

and Peterson, 1959). The corpus was also used as the basis of phonetic analysis (and, therefore, phonetic balance) for the CNC word lists.

The current corpus yielded 518 CVCV words after excluding nonwords and proper nouns. The existence of the words was confirmed using Kamus Dewan, a dictionary published by the Dewan Bahasa dan Pustaka. This corpus would be used in the analysis of phoneme distribution. It would also be the source of words to be used as the stimuli in the speech audiometry word lists. Close inspection of the corpus showed that the collection of words included those which are adapted from other languages, e.g. ‘gala’, ‘diva’ (both English) and ‘qada’ (Arabic). As these adapted words are regularly used in the Malay language, they will be included in the analyses. The list of words included in the corpus are shown in Table 3.1.

3.5.3 Analysis of phonemes

An analysis of the distribution of phonemes in the corpus and the projected distribution of phonemes in the word lists were made to form the basis of phonetic balance of the developed word lists.

The phonemes were analysed by calculating the frequency of appearance of each phoneme in the CVCV word corpus. Information on which phonemes that are most used were needed to build the word lists and therefore were drawn by ranking the distribution of the phonemes. Rank and distribution of vowels and consonants are summarised in Table 3.2 and Table 3.3.

Due to the limitation posed by the number of words in each list which affects the ability to include a similar distribution of phonemes to the findings in Tables 3.2 and 3.3, only phonemes with a distribution percentage of >1.0% were included in the word lists. All vowels (diphthongs and vowel sequences were not considered in the analysis) and 17 consonants fit the criterion. Consonants included in the development of the word lists are given in bold in Table 3.3. Consonants /v/ and /z/ are loan consonants into Malay (Teoh, 1994) while a search of words with /q/ in Kamus Dewan showed loan words mostly from the Arabic language.

Table 3.1 List of CVCV words extracted from UM/MM and BH/BM

baba	beku	badi	bora	babi	bida	bahu	bucu
baca	beli	bani	boya	bagi	biji	baju	budi
baja	beri	bari	buka	baki	bini	balu	budu
baka	besi	baru	buta	bali	biro	bayu	buku
bala	bila	basi	coli	bayi	biru	beca	buli
bapa	bina	batu	cuba	caca	bisa	bela	bulu
bara	ceti	dadi	cuka	cari	bisu	beta	bumi
bata	cina	dahi	cuma	dari	bola	beza	buru
bawa	cita	dani	duga	fana	cili	cela	busu
cara	debu	dasi	duka	gaji	ciri	dagu	cuci
dada	demi	dato	duta	gala	dini	daku	cucu
dana	demo	gali	foto	gama	diri	datu	curi
dapa	dewi	gani	gula	haba	diti	dera	cuti
dara	felo	gari	guna	haji	diva	desa	duda
data	feri	gasi	hobi	hama	dosa	dewa	dulu
daya	gebu	hasi	hoki	hara	fisi	gema	duri
fasa	geli	jali	huda	hari	gigi	hela	duti
gaya	geri	jani	joli	hati	giro	jamu	guni
hala	gila	kadi	juga	jadi	gitu	kaku	guru
hawa	hero	kamu	juta	jama	hiba	kayu	gusi
jaga	jeli	kari	koko	jame	hidu	kena	hulu
jala	jemu	kasi	koma	Jari	hina	kera	huni
jana	jeti	kawi	kopi	Jati	hipi	labu	huru
jasa	jika	laci	kosa	jawi	jitu	ladu	judi
jawa	jiwa	lagu	kuda	kaba	kilo	laju	Juri
jaya	kedi	lali	lobi	kafe	kima	laku	Juru
kaca	keji	lalu	logo	kaji	kini	layu	Kubu
kala	keju	madi	loji	kaka	kiri	lega	Kudu
kama	keli	mahu	lori	kaki	kiwi	leka	Kufu
kana	kelu	maju	loya	kali	kota	lena	Kuku
kata	kenu	maki	luka	kami	lidi	lewa	Kula
kaya	kira	mali	lupa	kara	liku	madu	Kuno
lada	kita	mami	moda	kasa	limo	malu	Kura
lama	lesi	mani	moga	kawa	miki	mega	Kutu
lara	lesu	mari	mono	laba	mini	meja	Kuyu
maha	levi	nadi	muda	laga	misi	nahu	Lucu
maka	liga	nafi	muka	lagi	nila	pacu	Lulu
mama	lima	nahi	mula	Lala	nini	padu	Lusa
mana	menu	pari	musa	Lari	nona	paku	Mudi
mara	pedu	qari	noda	Lava	nota	pasu	Mutu
masa	peha	raji	nusa	Mala	pili	payu	Nuri
maya	peti	rani	polo	Nabi	pipi	pena	Pura

Table 3.1 List of CVCV words extracted from UM/MM and BH/BM (cont.)

mata	peri	rali	pola	Mati	pilu	peka	Puji
nada	regu	rasi	popi	Nara	pita	peta	Puri
naga	reti	rawi	pori	Nasi	piza	rabu	Putu
nama	riba	safi	pula	Padi	pupu	ragu	Rudi
nana	sedu	saki	puma	Pagi	ribu	ramu	rugi
pada	segi	sami	roma	Paha	rima	ratu	Rumi
papa	seli	sani	rona	Paka	roda	rayu	suci
para	seni	sapi	roti	Pati	sifu	reda	sudi
paya	sepi	sari	rupa	Pusu	silu	reka	sudu
raga	seri	satu	rusa	Puyu	sini	rela	sufi
raja	seru	sawi	sofa	Qada	sira	sagu	suhu
rasa	sesi	sawo	solo	Rapi	siri	saku	suku
rata	sila	tahi	soto	Saga	sisi	sapu	sula
raya	sisa	tahu	soya	Saka	situ	sayu	sumo
saja	tebu	taji	suka	Sasa	tika	sena	suri
sama	teki	tari	tofu	Sate	tipu	sewa	susu
sana	teko	wahi	toko	Sawa	tiri	tamu	tubi
sapa	telo	wali	tona	Tadi	tiru	tatu	tugu
sara	temu	wari	topi	Taja	tisu	tega	tuju
saya	tepi	veto	toya	Tali	titi	teka	tuli
tata	tepu	wira	tuna	Tani	vila	tema	wifi
tawa	tiba	yeti	yoyo	Tapa	visa	tera	zina
waja	tiga	wana	tara	Tapi	visi		

Table 3.2 Ranking of vowels in Malay CVCV words based on distribution and the projected frequency in each list according to the number if CVCV words per list

		Distribution		Frequency of occurrence per list		
Rank	Phoneme	frequency	Percentage (%)	5-word list	10-word list	15-word list
1	a	413	27.83%	2.78	5.60	8.35
2	i	356	23.99%	2.40	4.80	7.20
3	e (/ə/)	289	19.47%	1.95	3.89	5.84
4	u	284	19.14%	1.91	3.83	5.74
5	o	85	5.73%	0.57	1.15	1.72
6	e (/e/)	57	3.84%	0.38	0.77	1.15

Table 3.3 Ranking of consonants in Malay CVCV words based on distribution and the projected frequency in each list according to the number of CVCV words per list

		Distribution		Frequency of occurrence per list		
Rank	Phoneme	Frequency	Percentage (%)	5-word list	10-word list	15-word list
1	**r**	151	9.792%	1.0	2.0	3.0
2	**l**	145	9.403%	1.0	2.0	3.0
3	**s**	143	9.274%	0.9	1.9	2.8
4	**t**	135	8.755%	0.9	1.8	2.7
5	**k**	124	8.042%	0.8	1.7	2.5
6	**b**	116	7.523%	0.7	1.5	2.2
7	**m**	105	6.809%	0.7	1.4	2.1
8	**d**	104	6.744%	0.7	1.3	2.0
9	**n**	86	5.577%	0.6	1.2	1.7
10	**p**	85	5.512%	0.5	1.0	1.5
11	**j**	78	5.058%	0.5	1.0	1.5
12	**g**	72	4.669%	0.5	0.9	1.4
13	**h**	48	3.113%	0.3	0.6	0.9
14	**c**	43	2.789%	0.3	0.5	0.8
15	**y**	35	2.270%	0.2	0.5	0.7
16	**w**	31	2.010%	0.2	0.4	0.5
17	**f**	17	1.102%	0.1	0.2	0.3
18	v	10	0.649%	0.1	0.1	0.2
19	z	9	0.584%	0.0	0.1	0.1
20	q	5	0.324%	0.0	0.1	0.1

There are two interpretations regarding phonetic balance. The majority describe phonetically balanced material as having the same distribution of phonemes as the language (Lehiste and Peterson, 1959; Ashoor and Prochazka, 1982; Wang et al, 2007). This view can be further broken down to the set of language the distribution is referred to, that is, whether it is the language in general, or a selected sample of the language. Usually, for the convenience of research, only a selected sample of the language is used as the basis of the distribution.

Another interpretation of phonetic balance was given by Mukari and Said (1991). Here, phonetic balance was shown to be representation of the phonemes of the language, irrespective of their distribution. Boothroyd (1968) had given a more appropriate term for this type of word lists; lists representing the phonemes of the language without representing their frequencies of occurrence is termed isophonemic word lists. This study has elected to use the former interpretation of phonetic balance in order to be

comparable with most of the previous phonetically balanced word lists. Comparative discussion on the findings of phonetic distribution in this study and previous studies can be found in the next chapter.

3.5.4 Development of word lists

Based on the CVCV word corpus yielded in 3.5.2, a list of top 350 most frequent words were made in order to ensure that the words used in the speech audiometry lists were familiar. The list was filtered for dialectal, slang or sensitive words and proper nouns. The remaining words were kept as a word bank to build the speech audiometry lists.

Selection of words were made based on the frequency of appearance of phonemes shown in Tables 3.2 and 3.3. For each list, 10 meaningful words were chosen first, followed by 5 nonsense words. The nonsense words were devised following the phoneme distribution and the phonetic rules discussed in the Methods chapter. The nonsense words were checked against Kamus Dewan to ensure that they do not occur in the Malay language. Lists containing all words and only meaningful words are shown in Tables 3.4a and 3.4b respectively.

Although the projected frequency of occurrence has been given as per Tables 3.2 and 3.3, variability of phoneme content between lists is expected following the selection of words to be included in the lists. Maximum care was taken to ensure that the variability is kept to the minimum, therefore, several words in the lists were not included in the corpus. However, these words are familiar words in spoken Malay and can be found in the dictionary, Kamus Dewan.

3.5.5 Phonetic balance

To evaluate the phoneme distribution of the word lists, each of lists from both the AWL and MWL sets were calculated. The distribution of phonemes for both the AWL and MWL are shown in Tables 3.5a and 3.5b respectively. Larger variability could be seen in the vowels, with /a/, /i/ and /u/ having the greatest difference between the lowest and highest number of occurrences in a list (4 each). However, as presumably several lists are used and presented to the listener in each session, the variability would not be as large.

To evaluate the phonetic balance of the corpus, the total phoneme distribution of each set of the 15 lists in the AWL and MWL are calculated and compared with the phoneme distribution of the corpus, which was used as the guide to develop the word lists. It is important to note that the phonetic balance is based on the distribution of phonemes described in section 3.5.3. Calculation is made on the distribution of the phonemes. The distribution of phonemes based on the corpus are summarised in Table 3.6. It can be

Table 3.4a Malay bisyllabic word lists – All words lists (AWL)

LIST 1	**LIST 2**	**LIST 3**	**LIST 4**	**LIST 5**	**LIST 6**	**LIST 7**	**LIST 8**
LAGI	JANI	HATI	DUSI	SANA	RUDA	HAWA	BOMA
BAHU	TALI	BINA	REDA	SEMA	RELA	PATI	JIWA
RATU	CUBA	TUJU	GEPA	SURI	MOJE	BENI	KASA
DABI	BEKU	RAHI	HOBI	BERI	KOPI	DEPU	SUHU
SIGU	GULA	FERI	GURU	KALI	DAYA	LALU	BACA
NASI	SATU	WIRA	LALI	MERI	FASA	MUDA	DUPI
MEJA	KIRA	SEPU	RULI	BATU	BELI	BIRO	PETI
KACA	HILA	BEKI	PENA	HOKI	BULA	SAYA	MUTU
TEPI	DARI	KAMU	KETI	JADI	SERI	KAJI	KARI
MONI	BOGA	CELA	CABA	KATU	KAMI	BUKA	GURA
SUDI	SEDI	SAPI	SITU	RADU	HAJI	SURA	LIGA
BARU	SEPI	DOSA	MASA	TEMU	TAGI	TISU	CABU
RUGA	TURU	LAKU	JATI	PAYA	CATU	KIMI	NADA
KELI	MANA	LUMA	CURI	BUTI	KOSI	TUMA	SUKI
KETU	LOJI	DOTA	KALA	LAGU	GUNA	RAGA	LOBI
LIST 9	**LIST 10**	**LIST 11**	**LIST 12**	**LIST 13**	**LIST 14**	**LIST 15**	
PADI	TANI	LORI	BAJA	KENA	KADI	TIBA	
BESI	TEGA	SIFU	PAJI	GARI	SUPA	NILA	
KEJU	KUBU	SETI	TIPU	LUKA	RUJI	PADA	
CUTI	MAKA	RIDA	LUCU	SOYA	SILA	TERI	
LITA	LOYA	DULU	TUKI	WARI	KERA	KOLE	
TARI	SEGI	MOKE	MIGA	MAHA	MALU	WAJA	
JATU	PILU	SATE	DESA	SEBA	DANA	MIKU	
SEGA	BIRU	BILU	GUNI	TEPU	BIMI	MAHU	
BOYA	SUMI	BUMI	KOMA	DUTA	TORI	SARA	
GALA	ROKI	PECU	SAWI	KUMU	TOPI	SIKU	
LESU	DATA	NAGA	ROBA	NOJI	RAJA	CUTI	
NERU	LUBA	JARI	LARI	JELI	CETI	SONA	
MUKA	PEDU	KATA	KITA	PAGA	DOBA	RUGI	
HARI	JASA	BAPA	BARA	DURI	SAGU	LIDA	
KOBI	CARI	KAYA	SENU	MISI	WALU	KABA	

seen that the difference between all phonemes in the MWL and the corpus is less than 1%. For AWL, all phonemes showed less than 1% difference in the distribution, except the phoneme /ə/ (difference in percentage: 1.06%) when compared to the corpus. This proves that the phoneme distribution of both MWL and AWL follows that of the corpus, and therefore, suggests that they have the same phonetic balance as the corpus. Evaluation on the phonetic balance of the word lists will be discussed further in the discussion chapter.

Table 3.4b Malay bisyllabic word lists – Meaningful-words lists (MWL)

LIST 1	**LIST 2**	**LIST 3**	**LIST 4**	**LIST 5**	**LIST 6**	**LIST 7**	**LIST 8**
LAGI	TALI	HATI	REDA	SANA	RELA	HAWA	JIWA
BAHU	CUBA	BINA	HOBI	SURI	KOPI	PATI	KASA
RATU	BEKU	TUJU	GURU	BERI	DAYA	LALU	SUHU
NASI	GULA	FERI	LALI	KALI	FASA	MUDA	BACA
MEJA	SATU	WIRA	PENA	BATU	BELI	BIRO	PETI
KACA	KIRA	KAMU	SITU	HOKI	SERI	SAYA	MUTU
TEPI	DARI	CELA	MASA	JADI	KAMI	KAJI	KARI
SUDI	SEPI	SAPI	JATI	TEMU	HAJI	BUKA	LIGA
BARU	MANA	DOSA	CURI	PAYA	CATU	TISU	NADA
KELI	LOJI	LAKU	KALA	LAGU	GUNA	RAGA	LOBI
LIST 9	**LIST 10**	**LIST 11**	**LIST 12**	**LIST 13**	**LIST 14**	**LIST 15**	
PADI	TANI	LORI	BAJA	KENA	KADI	TIBA	
BESI	KUBU	SIFU	TIPU	GARI	RUJI	NILA	
KEJU	MAKA	DULU	LUCU	LUKA	SILA	PADA	
CUTI	LOYA	MOKE	DESA	SOYA	KERA	KOLE	
TARI	SEGI	BUMI	GUNI	MAHA	MALU	WAJA	
BOYA	PILU	NAGA	KOMA	TEPU	DANA	MAHU	
GALA	BIRU	JARI	SAWI	DUTA	TOPI	SARA	
LESU	DATA	KATA	LARI	JELI	RAJA	SIKU	
MUKA	JASA	BAPA	KITA	DURI	CETI	CUTI	
HARI	CARI	KAYA	BARA	MISI	SAGU	RUGI	

Table 3.5a Number of occurrence for phonemes in AWL

	List/Number of occurence																
Phoneme	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Total**	**%**
a	8	7	8	7	9	9	8	10	9	9	7	8	9	9	9	126	14.00
b	3	3	2	2	3	2	3	4	3	3	3	3	1	2	2	39	4.33
c	1	1	1	2	0	1	0	2	1	1	1	1	0	1	1	14	1.56
d	2	2	2	2	2	2	2	2	1	2	2	1	2	3	2	29	3.22
ə	5	7	6	7	5	6	6	3	4	5	6	5	4	4	5	78	8.67
e	1	0	1	1	1	1	1	0	2	0	2	1	2	1	1	15	1.67
f	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	3	0.33
g	3	2	0	2	1	2	1	2	2	2	1	2	2	1	1	24	2.67
h	1	1	2	1	1	1	1	1	1	0	0	0	1	0	1	12	1.33
i	8	8	7	8	7	7	7	7	7	7	7	8	7	8	8	111	12.33
j	1	2	1	1	1	2	1	1	2	1	1	2	2	2	1	21	2.33
k	3	2	3	2	3	3	3	3	3	3	3	3	3	2	4	43	4.78
l	2	4	3	4	2	3	2	2	3	3	3	2	2	3	3	41	4.56
m	2	1	2	1	3	2	3	2	1	2	2	2	3	2	2	30	3.33
n	2	2	1	1	1	1	1	1	1	1	1	2	2	1	2	20	2.22
o	1	2	2	1	1	3	1	2	2	2	2	2	2	3	2	28	3.11
p	1	1	2	2	1	1	2	2	1	2	2	2	2	2	1	24	2.67
r	3	3	3	4	4	3	3	2	3	3	3	3	3	4	3	47	5.22
s	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	45	5.00
t	3	3	3	3	4	2	3	2	4	3	3	3	2	3	3	44	4.89
u	7	6	6	6	7	4	7	8	6	7	6	6	6	5	5	92	10.22
w	0	0	1	0	0	0	1	1	0	0	0	1	1	1	1	7	0.78
y	0	0	0	0	1	1	1	0	1	1	1	0	1	0	0	7	0.78

Table 3.5b Number of occurrence for phonemes in MWL

	List/Number of occurrence																
phoneme	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Total**	**%**
a	6	4	5	5	6	6	6	7	6	7	6	5	5	6	5	85	14.17
b	2	2	1	1	2	1	2	2	2	2	2	2	0	0	1	22	3.67
c	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	11	1.83
d	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	17	2.83
ə	4	6	4	4	4	5	5	3	3	3	4	4	4	4	4	61	10.17
e	1	0	1	1	0	1	0	0	1	0	1	1	1	1	1	10	1.67
f	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	3	0.50
g	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	14	2.33
h	1	0	1	1	1	1	1	1	1	0	0	0	1	0	1	10	1.67
i	5	5	5	5	5	5	4	5	5	5	4	5	5	5	5	73	12.17
j	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	16	2.67
k	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	29	4.83
l	2	3	2	3	2	2	2	2	2	2	2	2	2	2	2	32	5.33
m	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	16	2.67
n	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	13	2.17
o	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14	2.33
p	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	2.50
r	2	2	2	3	2	2	2	1	2	2	2	2	2	3	2	31	5.17
s	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	30	5.00
t	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	29	4.83
u	4	4	4	4	4	2	4	4	4	4	4	4	4	3	4	57	9.50
w	0	0	1	0	0	0	1	1	0	0	0	1	0	0	1	5	0.83
y	0	0	0	0	1	1	1	0	1	1	1	0	1	0	0	7	1.17

Table 3.6 Comparison between phoneme distribution of the word corpus and phoneme distribution of the word lists

	Proportion (%)			Difference to corpus (%)	
Phoneme	**Corpus**	**MWL**	**AWL**	**MWL**	**AWL**
a	13.91	14.17	14.00	0.26	0.09
b	3.70	3.67	4.33	-0.04	0.63
c	1.45	1.83	1.56	0.39	0.11
d	3.27	2.83	3.22	-0.43	-0.04
ə	9.73	10.17	8.67	0.44	-1.06
e	1.92	1.67	1.67	-0.25	-0.25
f	0.57	0.50	0.33	-0.07	-0.24
g	2.36	2.33	2.67	-0.02	0.31
h	1.35	1.67	1.33	0.32	-0.01
i	11.99	12.17	12.33	0.18	0.35
j	2.56	2.67	2.33	0.11	-0.23
k	4.18	4.83	4.78	0.66	0.60
l	4.95	5.33	4.56	0.38	-0.39
m	3.43	2.67	3.33	-0.77	-0.10
n	2.49	2.17	2.22	-0.32	-0.27
o	2.86	2.33	3.11	-0.53	0.25
p	2.90	2.50	2.67	-0.40	-0.23
q	0.13	0.00	0.00	-0.13	-0.13
r	4.98	5.17	5.22	0.18	0.24
s	4.68	5.00	5.00	0.32	0.32
t	4.48	4.83	4.89	0.36	0.41
u	9.56	9.50	10.22	-0.06	0.66
v	0.30	0.00	0.00	-0.30	-0.30
w	0.91	0.83	0.78	-0.08	-0.13
x	0.00	0.00	0.00	0.00	0.00
y	1.14	1.17	0.78	0.02	-0.37
z	0.20	0.00	0.00	-0.20	-0.20

3.6 Discussion

This section will evaluate and provide argument for the development process of the speech material, starting from sourcing and developing the corpus to the analysis of phonemes and, lastly, assembling the set of word lists itself.

3.6.1 Development of word corpus

Largely, word corpora were used as a source of words in building the wordlists. There is no consensus on the size of the initial word compilation that would be used in developing the word lists.

Words to be included in a set of speech audiometry material can be extracted from many sources, or none at all. Common sources used in developing the preliminary word lists include previously published corpora (Harris et al, 2007; Nissen et al, 2007; Wang et al, 2007), printed material (Ashoor and Prochazka, 1982) and even previously published word lists (Hirsh et al, 1952; Boothroyd, 1968). Nissen et al (2011) had used established electronic corpora as the source for their word list. Cantonese phonetically balanced monosyllabic audiometry word list by Lau and So (1988) did not involve any corpus at all as the words were selected from a matrix of Cantonese phonemes.

Due to time constraints and unavailability of published corpus in Malay, it was decided that sourcing the words online was the best option for this research. The choice was justified by the fact that the sources for the words were actually printed materials, ie. daily newspapers, which were able to be accessed online through the software offered by Dewan Bahasa dan Pustaka (DBP). It is worth noting that Utusan Malaysia and Berita Harian, together with their Sunday newspapers, were Malaysia's largest selling non-tabloid newspapers, according to Audit Bureau of Circulation (2011). Based on this, it is assumed that the corpus extracted from these two material would contain words that are familiar to readers and, therefore, Malay speakers.

Although there is no consensus regarding the capacity of the corpus, the size of the corpus in this study is comparable to previous studies. Corpora as little as less than 200 words (Ashoor and Prochazka, 1982; Nissen et al, 2007) and as large as 1263 words (Lehiste and Peterson, 1958; Wang et al, 2007) have been recorded. Nissen et al (2007) and Wang et al (2007) have each used corpora of 120 and 1263 words to produce word lists the size of 28 and 500 words, respectively. This gives a ratio of corpus-to-word list of between 2.5 and 4.6. The proportion of corpus-to-word of the current study gives a ratio of 3.45 (taking into account the meaningful words only), and, therefore, should allow a good size of corpus from which the words in the lists could be chosen.

It is important to note that the method of developing a preliminary word list from online or electronic sources are not exactly precedent. Nissen et al (2011) had described a

similar procedure in their study. An electronic interface, similar to the software in DBP, was used in selecting the words from the electronic word corpora. It is worth to note that, due to the ease of use and potential breadth of coverage, electronic corpora could be the preferred method of word selection in developing speech audiometry material.

The downside of using printed matter as the source of words is that, looking at the selection of words, there is a difference between the vocabulary of the written language and that of the spoken language. Spoken Malay has a different set of grammar and vocabulary as compared written Malay, although a big portion of the vocabulary does overlap. However, there are some words, such as 'keju' and 'boya', both of which appear in both the corpus and the word lists, that are considered as Standard Malay, and predominantly used in written Malay and less in the spoken version. This also puts on the question whether or not the developed material would be fit to be used in children, considering the level of sophistication of some of the words which can be unfamiliar or even mistaken as a non-word. Nevertheless, these words are not unfamiliar to an adult Malay speaker; as they are used widely in print. An ideal state would be to have a corpus of spoken Malay words as the source for the speech audiometry material. As for paediatric speech audiometry material, word sources, either printed or spoken, that are originally targeted for children should produce a more suitable and age-appropriate sets.

3.6.2 Analysis of phonemes

Phonemes are abstract units that form the basis of speech. It is defined as "…a minimal unit of sound capable of distinguishing words of different meaning" (Hyman, 1975). Each language has its own phonetic system. Malay has 6 underlying vowels (not including diphthongs and vowel sequences) (Table 3.7) and 26 consonants (Teoh, 1994), as compared to 11 vowels (excluding diphthongs) and 24 consonants in English (Roach,1998) (Table 3.8). These differences may result in differences in the frequency spectrum of sounds between Malay and English. An analysis of Malay frequency spectrum and its comparison with English would be discussed further in a later section.

Table 3.7 A comparison between Malay and English vowels (Teoh, 1994; Roach; 2000)

Malay	**English**
a	ʌ
e	e
ə	æ
i	ɪ
o	ɒ
u	ʊ
	iː
	ɜː
	ɑː
	ɔː
	uː

In order to achieve phonetic balance in the word lists, an assessment of the phoneme distribution was done using the words accumulated for the corpus. Twenty two consonants and 6 vowels were recorded from the analysis of phonemes. Compared to the list of phonemes listed by Teoh (1995), all vowels can be found in the current CVCV corpus. However, five consonants, / θ/, / ð/, /x/, / ɣ/ and / ʔ/ were absent from the corpus. It is interesting to see that while /q/ was counted as a phoneme in the current study, it was not described as a Malay consonant by Teoh, possibly due to its common pronunciation as /k/ in Malay phonetics. Phonemes / θ/, / ð/, /x/ and / ɣ/ did not appear in the CVCV phoneme analysis except as phonemes for proper nouns (which were excluded in the analysis). The glottal stop, /ʔ/ were purposely excluded from analysis as it is a manifest of the letter ‘k’ that are situated at the end of certain syllables or words, e.g. ‘budak’ - /budaʔ/. Inclusion of the phoneme may pose a scoring problem for testers, especially those who are unfamiliar with the language. Therefore, it is decided that the glottal stop would be excluded from the word lists.

Table 3.8 A comparison between Malay and English consonants (Teoh, 1994; Roach, 2000)

Malay	**English**
p	p
b	b
t	t
d	d
k	k
g	g
f	f
v	v
θ*	θ
ð *	ð
s	s
z	z
x*	ʃ
ɣ*	ʒ
h	h
m	m
n	n
ŋ	ŋ
l	l
w	w
r	r
ǰ	j
č	tʃ
ɲ	dʒ
y	
ʔ*	

An analysis of phoneme distribution using a text corpus sourced from local Malay news web pages was done by Tan et al (2009) (Figure 3.1). A comparison between the distribution of phoneme described by Tan et al. (ibid.) and the phonemes in the current study (Figure 3.2) showed that there are several similarities and differences found

between the two. It is important to note that the range of phonemes for both studies was almost identical. While /a/ was still the highest occurring phoneme in both studies, other vowels showed different proportions in the distribution between the two studies. The same can also be said for the consonants. The differences can be attributed to several factors – 1) phonemes being analysed, 2) word structure and 3) word sources. Tan et al. (ibid.) has included more phonemes compared to the current study; diphthongs and glottal stop were included in their study while the current study opted to exclude diphthongs and glottal stop in the analysis. Inclusion of diphthongs would have affected the proportions of other vowels. There were also several unrecognised phonemes in the study done by Tan et al which were not or defined in the original article. Secondly, the current study focused on bisyllabic, CVCV words in the phoneme analysis. This would have also contributed to the difference in percentages in the phoneme distribution. Inclusion of only CVCV words, as compared to general and unfiltered text, would cause vowels to form half of the phonemes being analysed which in turn would make the proportions higher for vowels. Thirdly, although the word sources for both studies are from news-based material, the current study derived its words from printed daily newspapers, Utusan Malaysia and Berita Harian (and their Sunday editions), whereas Tan et al. (ibid.) sourced its words from web-based news. There is a possibility that the collection of word differs, depending on the nature of articles featured in the newspapers.

The distribution of phonemes is important in the current study as it determines how phonetic balance is achieved.

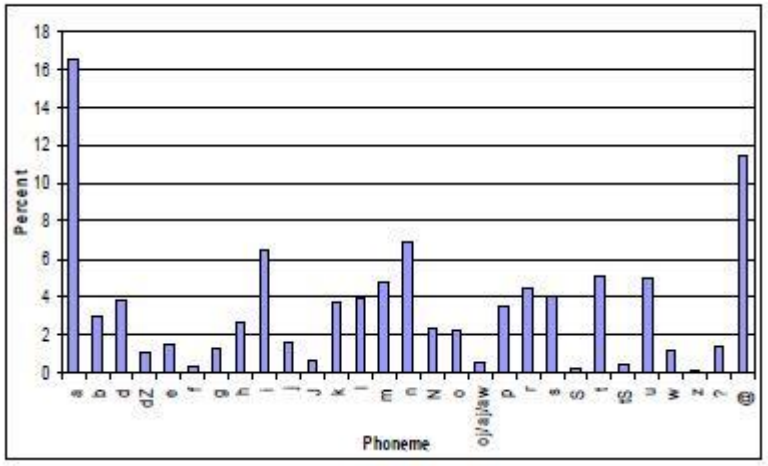


Figure 3.1 Phoneme distribution of Malay by Tan et al (2009)

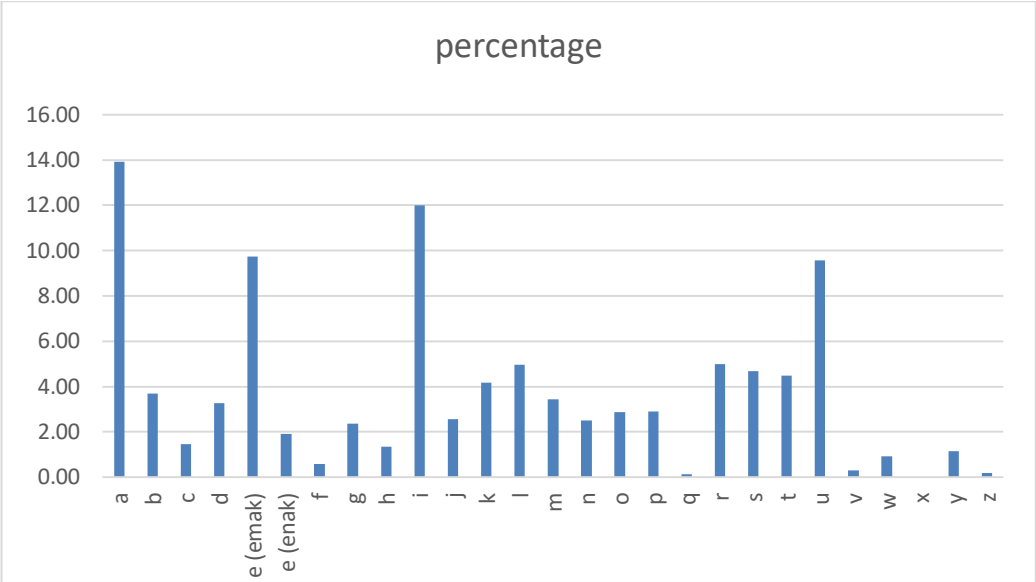


Figure 3.2 Phoneme distribution in the current study

3.6.3 Development of word list

Among the qualities preferred in a speech audiometric test are ease of administration and one that is quick to conduct. The test may be designed to fulfil these qualities through the size of stimuli, mode of delivery, mode of response and scoring style (Lawson and Peterson, 2011).

The current study aimed to produce a speech audiometry test that can be carried out using relatively basic audiometry equipment and can be performed as part of a routine hearing test session. In order to enable the test to be carried out with a regular audiometer, the test material can be designed to allow presentation using an external player, uploaded to audiometer, or even using live-voice. It was determined that the test would be an open-set test with verbal response from the listeners and therefore would not require any extra components to the test kit. Short lists allow for the test to be carried out as part of the routine hearing checks.

One of the objectives of the study was to investigate whether there is any difference in having nonsense words on top of familiar words as stimuli for speech audiometry as compared to having words that are all familiar. In order to achieve the objective, a set of lists containing both familiar words and nonsense words was constructed. The lists were made as such that they can be used as a 15-word meaningful+nonsense word lists or 10-word meaningful words lists interchangeably. Data collected in the analysis of

phonemes allowed for the calculation of the expected number of occurrence for the phonemes in each list.

This method for developing a phonetically balanced word lists has been described in previous studies. Hirsh et al (1952) applied a similar method using monosyllables, including phonemes with consonant-vowel (C-V), vowel-consonant (V-C) and consonant-vowel-consonant (C-V-C) structures, for their phonetically balanced W-22 word lists. It is interesting to point that, having sourced their words from printed materials, Hirsh et al included phonemes that are written as digraphs (e.g. /ph/, /th/). Having used similar sources, this method was also used in the current study for phonemes /ny/ and /ng/. Percentages occurrence of phonemes in each list was calculated, however, Hirsh has taken it further by assigning number of occurrences for initial and final consonant distributions separately. In contrast, this approach was not applied in the current study as the word structure that was applied, which is C-V-C-V, is distinctly different. It is justified by the reason that all of the syllable in the current word lists has the consonant-vowel structure instead of C-V-C or V-C, and therefore poses no difference whether the consonants are placed in the first or the third phoneme for the word. However, the placement of vowels in the current study does follow the phonetic rules described by Teoh (1994). Certain vowels, /e/ and /o/ only occur in the initial syllable but not in the final syllable.

Another similar method to the current study was described by Ashoor and Prochazka (1982). The same limits on words structure were imposed on the words included in their lists. Ashoor and Prochazka (ibid.) chose consonant-vowel-consonant (CVC) and consonant-vowel complex-consonant (CVVC) words in their lists. Consonant complex, for example $C_1VC_2C_3$, were not included due to the difference in their acoustic features as compared to the former two structures. In contrast with the current study, although other bisyllabic word structures, for example CVCVC or CVCCVC, were also excluded from the word lists, the rationale for it was based on the uniformity of scoring.

The developed lists in the current study is interchangeable, which means they can be used as full lists (all-word lists or AWL) (15 words) or as meaningful-words lists (MWL) consisting of 10 words. The rationale for having an interchangeable list was to fulfil an objective of the study, that is to observe whether or not speech audiometry material that mixes familiar and nonsense words is comparable to that made of only familiar words. An additional justification for having an interchangeable was that, in cases which the listener's ability to discriminate unfamiliar words is not an aim for an assessment or a

shorter test is preferred, the MWL can be used instead of the AWL to assess the listener's speech discrimination abilities and thresholds.

Having a set of interchangeable lists is not a novel concept in speech audiometry. Concept of full lists and partial-lists/half-lists have been utilized in speech audiometry material for some time. The reasons for having a shorter list were to reduce test time and to reduce patient fatigue which may be seen in a longer list (Mendel and Danhauer, 1997). Two well-known lists that has been investigated for and administered in half-lists were Phonetically Balance Kindergarten 50 Word lists (PBK-50) and Central Institute for the Deaf Wordlist 22 (CID-W22) (Runge and Hosford-Dunn, 1985).

3.6.4 Phonetic balance

The idea of having phonetic balance in a set of speech audiometry material comes from the aim of having a set of material that reflects the conversational speech. One of the objectives of speech audiometry is to assess the listener's ability to discriminate speech, therefore it is rational to mimic the attributes of conversational speech, particularly its sounds.

The phonetic balance of the word lists was confirmed by statistical analyses which showed no significant differences between the phoneme distribution of the AWL, MWL and the corpus.

Martin et al (2000) studied the importance of phonetic balance in speech audiometry. They compared the word recognition scores (WRS) when tested using the phonetically balanced NU-6 lists and the WRS obtained using random words selected from dictionary. They found no substantial difference between the WRS of those two. However, it is important to note that phonetic balance is a common characteristic of a word list; therefore, to allow better comparison with previously developed word list, it was decided that the current study employ a similar approach to its set of word lists.

The approach taken in the current study was to match the sounds in the speech audiometry material to the phonemes of Malay language. Several considerations were made in the methods; 1) a particular group of words, instead of conversational speech, were used as the basis of the study, 2) only single consonants and single vowels were analysed and included in the lists instead of consonant complex and/or vowel nucleus

(e.g. diphthongs) and 3) limitation on the phonemes included in the lists as compared to the phonemes available in the conversational Malay.

Phoneme analysis done on the word corpus earlier in the study provided a basis for the phoneme distribution for Malay speech sounds. The phoneme analysis provided the expected number of occurrences of the selected phonemes for the 15-word and 10-word versions of the developed lists. Although care had been taken in keeping the actual number of the phonemes in the lists close to the expected one, some variation could be seen in the phoneme distribution of the lists. This was to be expected as each list has a unique compilation of words. An assessment of the difference between the expected number of phonemes and the average of actual phonemes that occur showed that no phoneme in any of the lists showed a difference of equal to or more than 1. This means that although the difference was calculated based on the average of phoneme occurrence in the whole set, it is important to remember that clinically each test session would employ at least three lists.

Utilising a much larger word corpus as a basis of phonetic balance is a common method used in developing a phonetically-balanced word list (Hirsh et al, 1952; Lehiste and Peterson, 1959, Lau and So, 1988). The basis for the calculation of the expected number of occurrence per phoneme in the lists may come from connected text (Ashoor and Prochazka, 1982), conversational speech samples (Hirsh et al, 1952), previous study on phoneme distribution (Hirsh et al, 1952) or a published word corpus (Lehiste and Peterson, 1959). Differences in the source of phonetic distribution may affect the phonetic composition in the lists themselves; for example, sourcing the distribution from connected text may have included all sorts of word structures and not limited to the particular structure selected to be employed in the word list. It also prohibits direct comparison of phoneme distribution between individual sets of lists and between lists of different languages. Furthermore, using connected speech and/or word may not reflect the actual distribution of phonemes for the word structure selected for the lists, i.e. certain phonemes may only occur at the final position in a word, and therefore would not be reflective on a consonant-vowel-consonant-vowel (C-V-C-V) word lists. Taking this into consideration, it was decided that the phonetic balance of the lists in the current study would be based on a word corpus with the same structure. However, it is important to note that the method of determining phonetic balance in the current study is an adaptation of the methods used in the past studies.

The material for current study was developed to have two versions, a set of full lists, also known as all-word lists containing both meaningful and nonsense words (AWL), and a set of partial lists containing only meaningful words (MWL). It is intended that both sets of lists are phonetically balanced. The question of phonetic balance for partial lists was raised by Grubb (1963). A study on W-22 and a selection of PB-50 lists showed a large range of correlations between half-lists with the highest at .90 and lowest at .19. A closer analysis of the phonetic balance showed further imbalance in phoneme groups – voiced vs voiceless sounds, consonant clusters – within the split lists. This matter of phonetic balance between AWL and MWL has been taken into consideration in the current study. Preservation of phonetic balance were intended for both AWL and MWL, therefore, steps were taken by calculating the anticipated phoneme distribution in both groups of lists. It is important to note that all pairs of AWL and MWL showed no significant difference in phonetic balance between each other. All of the lists, both versions, also showed good agreement in terms of phoneme distribution with the distribution found in the corpus. These findings confirm that the phonetic balance is preserved in both versions of the list.

3.7 Summary

The research question that needed to be answered in this part of the study was 'What Malay phonological and phonetic features should be included in the Malay speech recognition test word lists?' with the aim of producing a phonetically balanced bisyllabic Malay word lists using combination of meaningful and nonsense words. The question was answered by building a CVCV word corpus in order to study the distribution of speech sounds in Malay language. The word structure included in the corpus was limited to CVCV as it was thought that distribution of phonemes in the word lists should reflect the distribution of the same word structure in the selected word source. The phonological structure was set based on previous studies in Malay phonology.

Based on the findings, a set of 15 phonetically balanced word lists consisting of 10 meaningful words and 5 nonsense words, AWL, was produced. The set of word lists was designed to have a shorter, phonetically balanced version consisting of only meaningful words, MWL.

The phonetic balance of both versions of word lists was verified by comparing the phoneme distribution of the word lists with the phoneme distribution of the corpus. To further verify the construct of the word lists in measuring speech intelligibility, the word lists were put through consistency and homogeneity analyses.

CHAPTER 4 VERIFICATION OF WORD LISTS

4.1 Introduction

This chapter discusses the second level in the development of the bisyllabic speech audiometry test in Malay. This part of the study involves the verification of the test material, a part of the speech audiometry material development process that deals with the equivalence and homogeneity of the test items. Homogeneity and consistency in the set of word lists are crucial, homogeneity ensures that the lists measure the same construct, which is speech intelligibility (Tavakol and Dennick, 2011). Consistency ensures that the results obtained are comparable no matter which list is used, as consistent results demonstrate reliability of the set (Wilson & Margolis, 1983). The aim of this part of the study was to find the answers to the question: Are the word lists equal to each other and can they be interchangeable? This question is important as it signifies the trustworthiness of the word lists in assessing speech intelligibility.

In the previous chapter, the processes that were carried out in constructing the word lists were discussed. It showed that the items are phonetically balanced, which met the intended definition of the test material and signifies that each list are equal in terms of phonetic content. The next step was to ensure that all the word lists are homogenous; that is, they are equal in terms of difficulty when listened to by normal hearing participants. A further assessment of acoustic content validity was carried out assessing the frequency spectra of the lists, and comparing the frequency spectra of the word lists to the long term average speech spectrum (LTASS) of Malay language.

The layout of this chapter contains the methodology and methods applied in the study, the results and the discussion on the findings.

4.2 Methodology

This section discusses the methodology of testing the validity of SRT test material. Two methods of validity testing are included in the discussion, validity testing through the assessment of homogeneity of the material, and validity testing through the assessment of the acoustic content of the material.

To determine the worthiness of the speech audiometry material, evaluative measures has to be performed. There are two major methods to evaluation – verification and validation. Verification is done on the speech audiometry material in order to ensure that the words lists are equal and interchangeable with one another. There are several verification methods that have been described in previous studies.

The most used method in verifying the lists is by measuring interlist intelligibility difference, which is, comparing the scores of normal hearing listeners at a set level (Ashoor and Prochazka, 1982; Causey et al, 1984; Hosford-Dunn & Runge, 1985; Magnusson, 1995). Scores from different lists are then statistically tested to see whether or not they are equal to each other. Any list that shows significant differences from the rest is taken out of the set.

Another verification method, although uncommon, is to test for homogeneity and filter the word stimuli before grouping them into lists. Russian monosyllabic speech audiometry material (Harris et al, 2007) is among the lists that were verified using this method. Due to the nature of grouping the words after verifying them, this method is more suitable in materials that do not employ phonetic balance.

Nissen et al (2007) introduced a new method to homogenise their word stimulus further. After excluding words that show significant variability, the set of words was equalised further by adjusting the intensity of the words digitally. More difficult words were increased in intensity to make it easier to hear while easier words were decreased in intensity to match the overall difficulty in intelligibility.

The frequency spectrum analysis and comparison with long term average speech spectrum (LTASS), however, have never before described in literature. Previously, spectral analysis of the speech stimuli was only done to produce speech-shaped noise for masking and/or filtering in speech audiometry studies (Nilsson et al, 1994; Peters et al, 1998). Each phoneme has its own unique acoustic properties and, in itself as well as when combined (to produce a word, sentence etc.), these acoustic properties (frequency and intensity, in particular) can be illustrated as a speech spectrum. Based on the premise that each list is phonetically balanced, i.e. containing the same distribution of phoneme, it is hypothesized that the speech spectrum of the lists is equal to each other as well as to the LTASS in terms of content.

4.2.1 Literature review of test material verification methods

Early developers of speech audiometry have proposed the importance of homogeneity in speech audiometry materials. Hudgins et al. (1947) and Egan (1948) emphasised on the importance of having tests items with equal difficulty in order to acquire higher probability of having the performance-intensity (P-I) function to be limited within a narrow range of intensity, thus increasing their sensitivity; as well as to allow subdivision of and interchange between test items. P-I function is defined as the recognition probability as a function of average speech amplitude (Boothroyd, 2008). Clinically, it is presented as an S-shaped curve denoting the correct score as a function of stimulus intensity. Hudgins et al. (ibid.) proposed choosing words that reaches the listener's threshold at the same intensity or adjusting their individual amplification to gain the effect.

The principle of having equal difficulty in terms of the P-I function for the test items still holds in later studies (Harris et al., 2007; Nissen et al., 2007; Han et al., 2009; Nissen et al. 2011). Nissen et al. (2011) selected only the top 28 words with the steepest psychometric function slopes, therefore having P-I functions with the narrowest range of intensity, to be included in their trisyllabic word list. For their bisyllabic word lists, Nissen et al. (ibid.) ranked the words according to the number of correct identification and then divided the words into four lists "equally" according to the ranks using S-curve distribution, i.e. four highest ranking words were put into list 1 through 4 respectively, then the fifth went to list 4, sixth list 3 and so on. The objective of implementing this method was to have a set of lists with equal difficulty.

A slightly different method was applied by Wang et al. (2007). Here the P-I function of the tests items were not used in the equivalence analysis. Instead, the constructed lists were presented to the listeners at one particular intensity. The scores were then statistically analysed using a test of normality. The lists with equivalent difficulty were selected to be used further.

The analysis of equivalence study was aimed to ensure that each test item (in this case, the lists) in the set was homogenous and consistent in terms of difficulty. Homogeneity ensures the reliability of the lists, whereby any one of the lists would generate the same result and therefore allow the lists to be interchangeable.

4.2.2 Review of methods

In general, there are three methods of analysis of equivalence used in testing speech audiometry word lists. The method chosen depends on the design of the material, that is whether the items are grouped in a fixed list (Ashoor and Prochazka, 1982; Comstock and Martin, 1984; Lau and So, 1988; Wang et al, 2007) or not (Nissen et al, 2005a; Harris et al, 2007). Two methods involve having the initial long list of words test items divided into shorter lists, and the equivalence study is done by comparing the lists. The content of the shorter lists may be preserved and used in the actual testing, i.e. 'fixed list', or interchangeable, i.e. 'non-fixed'. The third method involves testing the equivalence between the words instead of the lists. The selection of method is dependent on the final form the material and how the material will be utilised in the testing. Below are the descriptions of the methods:

Method I: Fixed list

This method analyse the equivalence of each list. The lists are being compared to each other for their homogeneity and difficulty in audibility. This method is suitable for test material that employs word lists in its design (as opposed to individual words). It is especially suitable for lists that are phonetically- or phonemically-balanced. Phonetically balanced lists contain the same distribution of phonemes as found in daily speech, while phonemically balanced lists have similar phoneme distribution for each list. Examples of phonetically and phonemically balanced lists are given in Table 4.1 and Table 4.2 respectively.

In this method, the prepared lists are presented to the participants at pre-selected intensity level or levels. The lists may be presented at just one intensity level (Wang et al, 2007) or several (Ashoor and Prochazka, 1982; Comstock and Martin, 1984; Lau and So, 1988). All lists are presented at the chosen intensities. The idea is to compare the score of the lists at each of the presentation levels. There is no consensus for the levels of presentation; Lau and So (ibid.) chose three arbitrary levels (10, 20 and 30 dB SL) for his study, while others opted to present each word lists at several levels to produce psychometric functions for the lists (Harris et al., 2007; Nissen, Harris and Slade, 2007; Nissen et al., 2011).

The results are then analysed statistically. Previous studies used t-test and ANOVA to determine the equivalence of the lists.

Table 4.1 Words in Auditory Test W-22 in alphabetical order, example of phonetically balanced word lists (Hirsh et al., 1952)

List 1		List 2		List 3		List 4	
Ace	Me	Ail	Move	Add	May	Aid	My
Ache	Mew	Air	New	Aim	Nest	All	Near
An	None	And	Now	Are	No	Am	Net
As	Not	Been	Oak	Ate	Oil	Arm	Nuts
Bathe	Or	By	Odd	Bill	On	Art	Of
Bells	Owl	Cap	Off	Book	Out	At	Ought
Carve	Poor	Cars	One	Camp	Owes	Bee	Our
Chew	Ran	Chest	Own	Chair	Pie	Bread	Pale
Could	See	Die	Pew	Cute	Raw	Can	Save
Dad	she	Does	Rooms	Do	Say	Chin	Shoe
Day	Skin	Dumb	Send	Done	Shove	Clothes	So
Deaf	Stove	East	Show	Dull	Smooth	Cook	Stiff
Earn	Them	Eat	Smart	Ears	Start	Darn	Tea
East	There	Else	Star	End	Tan	Dolls	Tin
Felt	Thing	flat	Tear	Farm	Ten	Dust	Than
Give	Toe	Gave	That	Glove	This	Ear	They
High	True	Ham	Then	Hand	Three	Eyes	Through
Him	twins	Hit	Thin	Have	Though	Few	Toy
Hunt	up	Hurt	Too	He	Tie	Go	Where
Isle	Us	Ice	Tree	If	Use	Hang	Who
It	Wet	Ill	Way	Is	We	His	Why
Jam	What	Jaw	Well	Jar	West	In	Will
Knees	Wire	Key	With	King	When	Jump	Wood
Law	Yard	Knee	Young	Knit	Wool	Leave	Yes
Low	You	Live	Your	Lie	Year	Men	Yet

Table 4.2 A sample of AB Word Lists, example of phonemically balanced word lists (Boothroyd, 1968)

List 1	List 2	List 3	List 4
Ship	Fish	Thud	Fun
Rug	Duck	witch	Will
Fan	Gap	Wrap	Vat
Cheek	Cheese	Jail	Shape
Haze	Rail	Keys	Wreath
dice	Hive	Vice	Hide
Both	Bone	Get	Guess
Well	Wedge	Shown	Comb
jot	moss	Hoof	Choose
move	tooth	bomb	job

Method II: Non-fixed list A

This method also employs the analysis of equivalence using lists. The difference between this method and Method I is that in this method, the lists may not be preserved in the final test material design. This method can be used in material designs with the test items grouped in lists or presented individually. However, this method is only suitable for lists that are not phonetically- or phonemically-balanced.

With this method, the selected words are divided randomly and equally (in quantity) into several lists. The lists are then presented to the listener at different levels. Nissen et al (2005a) and Harris et al (2007) both presented one randomly selected list at each specified level. In both studies, the words were then regrouped again and presented again in the same manner to another group of listeners. Equal number of presentations at each level for each word was determined. The results were then analysed for normal distribution and logistic regression.

Method III: Non-fixed list B

This method refrains from using any list. Instead, all selected words are presented to the listener at each chosen intensity level (Nissen et al, 2005b). A designated number of words that shows the most statistical equivalence (using ANOVA, for example) is chosen for the final speech audiometry material. This method is suitable for material designs that

employ just one, usually long, list of words as the test items, instead of a set of different word lists.

The difference between the fixed and non-fixed lists lies in the actual method of clinical testing and, ultimately, the information the tester intends to obtain. Non-fixed list type B, due to its single list, allows only speech reception threshold (SRT) investigation. This type of list is suitable for either descending threshold seeking method or methods similar to the Hughson-Westlake method (5-up, 10-down), where the initial presentation level is set close to the estimated SRT and testing is halted when the listener misses a percentage of the words presented (Lawson & Peterson, 2011). These methods, however, do not allow for suprathreshold assessment, particularly on speech discrimination. On the other hand, non-fixed list type A and fixed list allow for the formation of P-I function, which, in turn, allow for both SRT and suprathreshold speech discrimination assessments anticipated in the current study. However, as the current study intends to produce phonetically-balanced word lists, the fixed list method is the method of choice.

Statistical analysis

There are several methods of statistical analysis for homogeneity and equivalence testing used in the previous studies of speech audiometry material development. Nissen et al. (2011) devised a statistical method using two-way chi-square testing on the various slopes of the speech curves obtained using the speech material, with intensity and list as their independent variables. Other developers had used analysis of variance (ANOVA) alone or together with Student-Newman-Keuls-Q (SNK-Q) test to test the consistency of their word lists (Lau and So, 1988; Wang et al. 2007).

On a slightly different procedure, logistic regression was used to get the psychometric functions and regression slopes of the word lists. No particular statistical analysis was used to measure the homogeneity; instead, lists with the steepest slopes were chosen to be included in the set (Harris et al., 2007; Nissen, Harris and Slade, 2007; Nissen et al., 2011).

It was thought that the method devised by Nissen et al. (2011) was unsuitable for the current study as it was formulated to allow comparison between male and female talkers. On the other hand, repeated measures ANOVA would be able to provide certain level of information on the equivalence of the items (in this case, lists) in the group, but it might

not be able to pinpoint which of the list or lists that differ significantly from the rest. Post-hoc tests, for example SNK-Q, would have been done to identify said list/lists.

In addition to repeated measures ANOVA, the measure of equivalence between the lists in the current study was also carried out in terms of homogeneity or internal consistency. Homogeneity can be measured using Cronbach's alpha, an intraclass correlation coefficient (ICC) (McGraw and Wong, 1996). The use of Cronbach's alpha is supported by Boyle (1991) who stated that it is a "more adequate" test of homogeneity.

4.2.3 Methods used in this study

In the current study, the word lists were tested for their homogeneity in terms of difficulty and analysed for validity of their acoustic content. The evaluations included both measurements through audiometric testing and objective measurement through LTASS analysis.

The LTASS analysis was a new technique of determining validity for speech test materials. In previous studies of speech audiometry material, validity testing only involved audibility of test items and/or linguistic content (Lehiste and Peterson, 1959; Ashoor and Prochazka, 1982; Harris et al., 2007; Nissen et al., 2007). Furthermore, the homogeneity and consistency analyses confirm the lists comparability, but they do not validate the lists representativeness of the language.

LTASS, on the other hand, represent the acoustic content of the language, particularly its frequency components. Validation that the speech audiometry material actually represents the acoustic content of the language can only done through LTASS.

4.2.3.1 Analyses of consistency and homogeneity: Research design

Because the prepared set of word lists are phonemically-balanced, the 'fixed list method (Method I in Section 4.2.2), similar to the method described by Lau and So (1988) was applied. The lists were presented to the listeners at two intensities, 15 dB dial and 40 dB dial. It was found that this method is suitable to be used for homogeneity test using internal consistency analysis through intraclass correlation coefficient. This experimental design is also thought to be more straightforward, simpler and less demanding to the

listeners as compared to the 'psychometric function' method. The following paragraphs outline the details of the research design in the analyses of consistency and homogeneity. The research design is summarised in Figure 4.1.

Ethics approval to perform the current study was gained from the Research Ethics Committee, Faculty of Health and Life Sciences, De Montfort University.

Normal hearing participants were selected to take part in the study. Participants were recruited among the Malaysian students living in Leicester and their family members. The inclusion criteria were hearing thresholds better than 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000Hz, and uneventful otological and hearing histories. Exclusion criteria were history of noise exposure, significant tinnitus and recent illness that were possibly related to hearing problems. The better ear was selected for further testing. Participants were native speakers of Malay to ensure familiarity to the test items.

Prior to audiometric testing, the AWL was recorded in a professional recording studio using male voice. The recording was done in a recording studio using Digidesign Pro Tools LE7 audio recording software. The talkers were given the word lists before the recording in order to familiarise themselves with the text. A large font print-out of the word lists were placed at least 1 meter in front of the talker to enable them to read the text while facing the microphone without having to move their head. The microphone was placed at roughly 30 cm in front of the mouth at 45° azimuth relative to the axis of the mouth to lessen wind effect. For this recording, the talkers were asked to read the text with natural stress, speed and level of intensity. The talkers were asked to continue reading even though they made mistakes. To avoid any variations in stress, intonation and voice levels, each list was read uninterruptedly. The recordings were repeated for any lists that were deemed unsatisfactory by the researcher, talker and/or sound engineer. The recorded voice samples were written into a CD for further use.

The editing was made in the studio using Digidesign Pro Tools LE7 audio recording software. The recordings were edited for noise and amplitude. As the recording for the words in each list was made in one continuation, the words were segmented to eliminate any noise and to allow further editing. The words were then spliced with an allowance of 5-second interval in between each item. The recording of each list formed a track. Tracks recorded from each talker were saved into a file. A 30-second 1000 Hz track was added into each file to allow for the purpose of calibration of the VU meter on the audiometer before the hearing test. The RMS amplitude of the list tracks were then adjusted to be within ± 1 dB of the amplitude of the 1000 Hz track.

The recordings were then given to a panel of three judges to be selected. The judges consisted of two audiologists and a linguist. The recording with the best quality in terms of clarity, voice quality, intonation, stress and speed were selected and used for the following investigations.

Routine audiometric tests such as otoscopy, tympanometry and pure tone audiometry were conducted in a sound-treated room on each participant as a preliminary assessment to determine the hearing thresholds and otological condition. Pure tone audiometry comprised of frequencies 250, 500, 1000, 2000, 4000 and 8000Hz.

Speech audiometry was carried out following the preliminary assessment. The selected recording of AWL was used for the speech reception threshold test. The participants were first briefed on the test procedure. The participants were instructed to listen to the bisyllabic words and repeat the words they hear. Guessing was encouraged. They were then given a sheet containing the instructions to the test and were asked if they need any clarification. The instructions were in Malay and carried the translation, "*You will be presented with several words. The words may be loud or soft. Please repeat the words you hear, no matter if it carries any meaning or not. You will be given time to repeat the word after it is presented. You are also encouraged to make a guess*."

All lists were presented at two intensity levels, 15 dB dial and 40 dB dial. The lists were first presented at 15 dB HL, followed by 40 dB HL. Familiarisation to the test materials was given at the beginning of the test by presenting the first three words from a list that was chosen randomly at the level of 40 dB HL. A randomly selected list was chosen to be the practice list so that the participants would become familiar with the test. The sequence of list presentation was randomised.

Before any testing was begun, the audiometer was calibrated for the speech audiometry material CD. The recorded 1000 Hz calibration tone was played on the CD and the audiometer intensity dial was adjusted so that the VU display was set at 0.

The responses from participants were noted on the response sheet and were later scored. Scoring was done according to the common speech discrimination assessment procedure which is phonemic scoring. Each word item carried a score of 4 (1 for each phoneme). A score was given for every correct phoneme. Maximum score for each list was 60. The scores from each participant were counted, noted and saved into an Excel file.

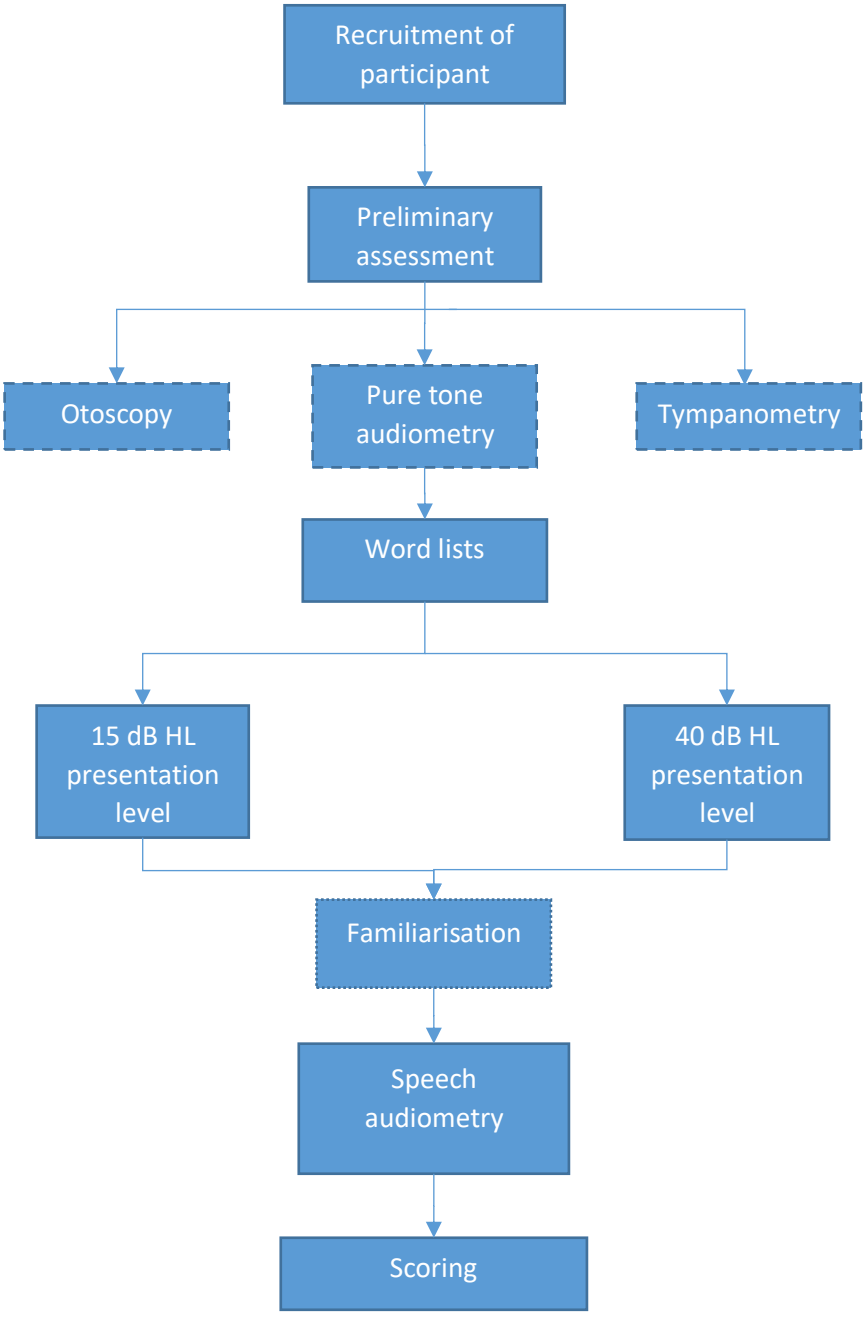


Figure 4.1 Data collection process for the analyses of consistency and homogeneity

The results for each list were then analysed using Friedman Test for consistency assessment and Cronbach's alpha for homogeneity testing. Statistical Package for Social Sciences (SPSS) was used to execute the statistical tests.

4.2.3.2 Validity of acoustic content of the word lists

The analysis of the acoustic content of the word lists was done to determine that the frequency content of the wordlists reflects the frequency content of the Malay language. It was also done to verify that the frequency content of the word lists is similar to each other.

To compare the phonetic balance between the word lists and the Malay language in general, long term average speech spectrum (LTASS) (Byrne et al., 1994) was used as the reference spectrum. There is no published study in the LTASS of Malay or a language similar to it (e.g. Indonesian, Brunei Malay), therefore an analysis of Malay LTASS was carried out first.

The method of establishing LTASS as described by Byrne et al. (1994) was used as a guideline in constructing the Malay LTASS. Byrne et al.'s (ibid.) study was done to create a 'universal' LTASS, taking into account several languages from different parts of the world – North America, Europe, East Asia, Egypt and Australia. It involved the compilation of vocal samples and examining their dynamic range across the frequencies. The findings from the 12 selected languages were combined to produce a universal LTASS.

The purpose of this section of the study was to establish the long term average speech spectrum (LTASS) in Malay. Recordings of Malay speech were used to form the LTASS. The Malay LTASS was then used as a reference in the homogeneity of speech spectrum study of the word lists. As the word lists were supposed to represent the real-life speech spectra, they were expected to have similar frequency content and intensity range with the LTASS. Comparisons between the frequency spectrum of Malay LTASS and the frequency spectra of the word lists were made. How Malay LTASS compares to the universal LTASS would also be discussed.

The method for measuring LTASS is outlined below.

Participants

Ten participants were selected to read the text and record their voice for the purpose of LTASS measurement. The participants were staff and students of the Faculty of Allied Health Sciences, International Islamic University Malaysia who volunteered for the study. Five participants were male and 5 female. All participants were native speakers of Malay

and had no obvious speech defects. No other selection criteria were employed. The range of age is 22 to 36 years old with an average of 29 years.

Speech materials

Two passages in Malay were selected to be read in the collection of voice samples. The passages were titled ‘Kampung’ (The Village) and ‘Datuk’ (Grandfather) as shown in Table 4.3. Both are narrative essays. The passages contained all phonemes used in Malay, although the distribution of phonemes in passaged did not reflect the distribution of phonemes in general Malay discourse. When read at normal speed, the combination of both phonemes provided more than 100 seconds of continuous discourse.

Recording procedure

The method of recording was adapted from the method described by Byrne et al. (1994). Byrne et al. (ibid.) studied the LTASS of several languages around the world, based on collected language samples. Adapting the method would allow better comparisons between Malay and other languages.

The recording was done in a sound treated room (an audiology booth) using a Zoom H1 recorder in order to minimise noise. The recorder was placed approximately 15cm away from the speaker, at approximately 0° azimuth to the speaker’s mouth. The procedure of placing the microphone at 45° azimuth relative to the speaker’s mouth as per Byrne et al. (1994) was not followed as the Zoom H1 recorder employs an X/Y stereo microphone system, which has 2 microphones at 90° to each other. The method of microphone placement at 0° azimuth relative to the speakers mouth has been described in previous LTASS studies (Cornelisse et al., 1991a; Cornelisse et al., 1991b; Noh and Lee, 2012). The reading material was placed on a desk in front of the speakers, taking care that the direction of the microphone was preserved and there was no obstruction between the recorder and the speaker.

The participants were instructed to read the passages at normal pace, loudness and intonation. The participants were allowed some time to familiarise themselves with the reading material. The participants were instructed to keep on reading even when mistakes were made. Participants who made significant amount of mistakes (e.g. more than 5 reading errors; or substitutions like chuckling, coughing etc.) were asked to repeat the recording.

Table 4.3 'Kampung' and 'Datuk' passages

'Datuk' (Grandfather)

Minggu lepas, saya bersama ayah dan ibu pulang ke kampung, kerana menziarahi datuk yang sedang sakit. Datuk telah tiga hari sakit tetapi dia tidak mahu dihantar ke hospital. Datuk lebih suka dirawat di rumah sahaja.

Walaupun nenek, ayah dan ibu puas memujuknya untuk berjumpa doktor, datuk tetap berdegil. Akhirnya ayah meminta Cikgu Syarif, cucu saudara datuk untuk memujuknya. Barulah datuk mengalah dan mahu dibawa ke hospital. Syukurlah kini datuk telah sembuh dari sakitnya.

Kami pun berkemas untuk pulang semula ke rumah. Malangnya kereta ayah rosak pula. Pakcik Johan mengajak kami pulang menaiki van barunya. Kami sangat gembira dan pulang ke Johor keesokkan harinya bersama-sama.

'Kampung' (The Village)

Semasa saya masih kecil, saya tinggal di kampung bersama nenek. Ayah dan emak tinggal di bandar. Abang dan adik saya tinggal bersama ayah dan emak. Setiap minggu saya dan nenek pergi ke rumah ayah.

Jika kami tidak ke sana, ayah, emak, abang dan adik datang menziarahi kami di kampung. Saya gembira apabila mereka datang kerana saya boleh bermain bersama-sama abang dan adik. Di belakang rumah nenek ada sebatang sungai. Abang sangat suka mandi di sungai, begitu juga saya tetapi adik tidak boleh bermain di sungai. Dia masih kecil dan belum boleh berenang.

Setiap pagi, abang dan saya pergi ke sungai. Kadang-kadang kami lupa makan dan minum. Di tebing sungai itu ada sebatang pokok rambai. Abang suka memanjat ke dahannya yang rendah lalu terjun ke dalam air. Semasa dia di dalam air, saya bimbang dia akan lemas. Mujurlah kepalanya sentiasa berada di atas permukaan air. Apabila melihat abang demikian, saya ketawa kegembiraan.

The recording was set at 44.1 kHz sampling frequency and 16-bit resolution, similar to a previous LTASS study using digital recording (Noh and Lee, 2012). The speech samples were saved in a digital memory card in the form of WAV files and then transferred to a portable hard drive.

Spectrum analysis procedures

The spectrum analysis procedures consisted of two parts – spectrum analysis of the speech samples and spectrum analysis of the word lists.

The speech samples were analysed using audio editing software Audacity (Cornelisse et al., 1991a; Byrne et al., 1994), an open source software available and could be downloaded free from the internet (audacityteam.org). A high pass filter at 100-Hz with 6-dB per octave roll-off was applied to the speech sample to remove any background noise. The individual speech samples were then concatenated in to come up with a normalised combination of all voice samples.

The frequency spectra of the concatenated speech were then plotted using Audacity. The data were then exported to and saved in the form of Excel files for further analysis.

The recorded AWL word lists were also analysed in a similar way. The WAV files of AWL were uploaded in Audacity. No filtering was applied to any of the files. A new file was also created, containing concatenation of all 15 word lists. All 16 files (15 word lists and 1 concatenate file) were then analysed for their frequency spectrum. The plotted graph were then exported and saved in the form of Excel files.

Statistical analysis

To apply statistical analyses to the data, the absolute intensity levels at the centre of each 1/3 octave frequencies between 100 to 10000 Hz were derived from the raw data of both the voice samples and word lists. This derivation generated 24 points along the frequency spectra whose values were used for comparison. Table 4.4 shows the list of selected frequencies.

Repeated measures ANOVA was intended for the 24 selected frequencies (Table 4.4) comparing the 15 word lists and the Malay LTASS. The current study chose to employ its non-parametric equivalent, Friedman Test, as the data fulfilled the required assumptions (Statistics.Laerd.com, 2015).

Significant differences between any of the variables were noted.

4.3 Results

Two methods of verification were carried out on the word lists. The first method determined that the word lists were equally difficult while the second ensured that the frequency spectrum of the word lists reflected that of the general spectrum of speech in Malay.

The method that assesses equal difficulty in the word lists are commonly applied in the verification of speech audiometry word lists. Several approaches are available in determining equal difficulty, as discussed in section 4.2.2. As the current study utilised phonetically balanced word lists, the verification of equal difficulty was done by the fixed list method, whereby the words within a list is preserved, the lists are presented at a fixed intensity and the scores are compared to determine equal difficulty.

Table 4.4 Frequencies selected for the word lists-LTASS comparison

Frequency (Hz)
43.07
86.13
129.20
172.27
215.33
258.40
301.46
387.60
516.80
646.00
818.26
990.53
1248.93
1593.46
1981.05
2497.85
3143.85
4005.18
4995.70
6287.70
8010.35
9991.41
12489.26
16020.70

4.3.1 Participants' audiological assessment

Six adults were recruited to participate in the homogeneity and consistency assessment and given identification codes C1 to C6. Three of the participants were male and three were female. The participants' age ranged between 27 to 39 years old with an average of 32. Prior to the homogeneity and consistency assessment, audiological assessment consisting of otoscopy, tympanometry and pure tone audiometry was done on the participants. Only one ear, which was the better ear, was chosen to be the test ear. One male participant, C5, did not meet the criterion for pure tone thresholds as he showed thresholds of more than 20 dB HL in several test frequencies. Therefore, he was

excluded from the study and was advised to seek further audiological assessment and consultation.

The remaining five participants were tested on the right ear as they showed to be the better ear. All showed pure tone thresholds of 15 dB HL or less across the frequencies. Details on the pure tone thresholds of each participant are tabulated in Table 4.5.

Table 4.5 Pure tone thresholds of participants in the homogeneity and consistency study

		Pure tone thresholds (dBHL)					
ID	**Test ear**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**
C1	right	15	15	10	5	0	15
C2	right	15	15	5	5	15	10
C3	right	10	0	0	0	-5	0
C4	right	10	0	5	0	15	10
C6	right	5	0	10	5	5	-5
Average		11	6	6	3	6	6

4.3.2 Consistency of the word lists

The AWL word lists were administered in random orders to each participant at 15 dB dial. Scores of phoneme were presented in Table 4.6. The phoneme scores were then converted to percentage scores and presented in Table 4.7. The highest score achievable is 60. The scores at 15 dB dial vary both list-wise and participant-wise. On average, List 12 produced the lowest correct scores at 45.4 (75.7%) while List 15 produced the highest at 53.6 (89.3%). Participant C2 scored relatively lower than the other participants across the lists, with an average of 39.2 (65.3%) correct scores. Meanwhile, participant C6 showed the highest average correct scores at 56.0 (93.3%). Participant C6 also showed the most consistent performance across the list as seen in the narrow range of correct scores. C6 correct scores ranged between 54 (90%) and 59 (98.3%), while the other participants showed differences of >10 between their highest and lowest scores.

The AWL were also administered in random orders to each participant at 40 dB dial. Scores of phonemes were presented in Table 4.8. The phoneme scores which were converted to percentage scores were presented in Table 4.9. The correct scores of all

participants for all lists were equal to or exceeded 57 (95%) except for one instance where participant C4 scored 56 (93.3%) for List 7.

In terms of difficulty, there was no obvious pattern of the 'easiest' or the 'hardest' lists, judging from the highest and lowest scores achieved by each participants. There was also the possibility of reaching the plateau of correct phoneme score at 40 dB dial judging by the consistently high scores showed by all participants. However, looking at the results for 15 dB dial presentation level, List 15 did give the highest scores in three of the participants (C2, C3 and C6) and List 12 the lowest scores in three of the participants as well (C3, C4 and C6). A statistical analysis was carried out on the scores of 15 dB dial for AWL presentation level to see whether these lists are significantly more difficult (or easier) than the others.

The correct phoneme scores for MWL at 15 dB HL were also calculated. Maximum score achievable for MWL was 40. Similar patterns in the performance of the participants can be seen in the MWL; C6 showed the most consistent performance and the highest average score at 37 (92.5%). The lowest average score was 25.73 (64.3%) showed by C2, similar to the AWL finding. List 15 produced the highest average correct score at 35.2 (88%) and List 12 the lowest at 30 (75%). Correct phoneme scores at 15 dB dial for MWL and the scores in percentage are presented in Tables 4.10 and 4.11 respectively.

The consistencies of speech discrimination using AWL and MWL word lists were analysed using Friedman test due to the non-normal distribution of some of the word list scores. There was statistically no significant difference in correct scores achieved using any of the AWL word lists, X^2 (14) = 19.584, p= 0.144 (Table 4.12). There was also statistically no significant difference in correct scores achieved using any of the MWL word lists, X^2 (14) = 17.554, p= 0.228 (Table 4.13).These results showed that the choice of word list used in testing had no effect on the speech audiometry scores in normal hearing participants. This marked the consistency of both AWL and MWL word lists, signifying that the performance of the listener would not be affected by the selection of word lists for either versions, and that the results should be comparable irrespective of which word list was used.

Table 4.6 Correct phoneme score scores at 15 dB dial for AWL

Patient ID	LIST 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average (range)
C1	52	49	46	40	47	52	52	50	51	49	44	43	52	55	47	48.6 (40 - 52)
C2	37	42	32	38	41	40	40	42	39	39	39	41	37	34	47	39.2 (32 - 47)
C3	54	50	52	45	48	53	53	49	52	45	52	44	51	53	58	50.6 (44 - 58)
C4	54	52	55	54	56	48	51	54	56	55	54	45	55	58	57	50.3 (45 - 58)
C6	56	54	58	55	56	56	56	56	54	57	56	54	57	56	59	56.0 (54 - 59)
Average (range)	50.6 (37-56)	49.4 (42-54)	48.6 (32-58)	46.4 (38-55)	49.6 (41-56)	49.8 (40-56)	50.4 (40-56)	50.2 (42-56)	50.4 (39-56)	49.0 (39-57)	49.0 (39-56)	45.4 (41-54)	50.4 (37-57)	51.2 (34-58)	53.6 (47-59)	

Table 4.7 Correct phoneme score scores at 15 dB dial for AWL in percentage

	LIST															
Patient ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Average (range)**
C1	86.7	81.7	76.7	66.7	78.3	86.7	86.7	83.3	85.0	81.7	73.3	71.7	86.7	91.7	78.3	81.0 (66.7 - 91.7)
C2	61.7	70.0	53.3	63.3	68.3	66.7	66.7	70.0	65.0	65.0	65.0	68.3	61.7	56.7	78.3	65.3 (53.3 - 78.3)
C3	90.0	83.3	86.7	75.0	80.0	88.3	88.3	81.7	86.7	75.0	86.7	73.3	85.0	88.3	96.7	84.3 (73.3 - 96.7)
C4	90.0	86.7	91.7	90.0	93.3	80.0	85.0	90.0	93.3	91.7	90.0	75.0	91.7	96.7	95.0	89.3 (75 - 96.7)
C6	93.3	90.0	96.7	91.7	93.3	93.3	93.3	93.3	90.0	95.0	93.3	90.0	95.0	93.3	98.3	93.3 (90 - 98.3)
Average (range)	84.3 (61.7-93.3)	82.3 (70 -90)	81.0 (53.3-96.7)	77.3 (63.3-91.7)	82.6 (68.3-93.3)	83.0 (66.7-93.3)	84.0 (66.7-93.3)	83.7 (70 - 93.3)	84.0 (65 - 93.3)	81.7 (65 - 95)	81.7 (65 - 93.3)	75.7 (68.3-90)	84.0 (61.7 - 95)	85.3 (56.7-96.7)	89.3 (78.3 - 98.3)	

Table 4.8 Correct phoneme scores at 40 dB dial for AWL

Patient ID	LIST 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average (Range)
C1	57	59	60	58	60	59	60	60	58	57	60	58	60	60	59	59.00 (57 - 60)
C2	60	58	60	58	60	58	59	58	59	60	60	60	60	59	60	59.27 (58 - 60)
C3	58	59	59	60	58	60	58	57	59	60	59	59	60	59	57	58.80 (57 - 60)
C4	60	59	60	59	59	59	56	60	59	60	59	60	59	58	60	59.13 (56 - 60)
C6	60	60	60	59	60	59	60	60	60	60	60	60	59	57	60	59.60 (57 - 60)
Average (Range)	59 (57 – 58)	59 (58 – 60)	59.8 (59 – 60)	58.8 (58 – 60)	59.4 (58 – 60)	59 (58 – 60)	58.6 (56 – 60)	59 (57 – 60)	59 (58 – 60)	59.4 (57 – 60)	59.6 (59 – 60)	59.4 (58 – 60)	59.6 (59 – 60)	58.6 (57 – 60)	59.2 (57 – 60)	

Table 4.9 Correct phoneme scores at 40 dB dial for AWL in percentage

	LIST															
Patient ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Average (range)**
C1	95	98.3	100	96.7	100	98.3	100	100	96.7	95	100	96.7	100	100	98.3	98.3 (95-100)
C2	100	96.7	100	96.7	100	96.7	98.3	96.7	98.3	100	100	100	100	98.3	100	98.8 (96.7-100)
C3	96.7	98.3	98.3	100	96.7	100	96. 7	95	98. 3	100	98.3	98.3	100	98.3	95	98 (95-100)
C4	100	98.3	100	98. 3	98.3	98. 3	93. 3	100	98. 3	100	98. 3	100	98. 3	96.7	100	98.6 (93.33-100)
C6	100	100	100	98. 3	100	98. 3	100	100	100	100	100	10	98. 3	9	100	99.3 (95-100)
Average (range)	98.3 (95 - 100)	98.3 (96.7 - 100)	99.7 (98.3 - 100)	98 (96.7 - 100)	99 (96.7 - 100)	98.3 (96.7 - 100)	97.7 (93.3 - 100)	98.3 (95 - 100)	98.3 (96.7 - 100)	99 (95 - 100)	99.3 (98.3 – 100)	99 (96.7 - 100)	99.3 (98.3 - 100)	97.7 (95 - 100)	98. 7 (95-100)	

Table 4.10 Correct phoneme scores at 15 dB dial for MWL

	LIST															
Patient ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Average (range)**
C1	33	33	30	26	30	34	34	33	34	32	29	28	34	36	31	31.80 (26-36)
C2	24	28	21	25	27	26	26	28	26	26	26	27	24	21	31	25.73 (21-31)
C3	36	33	33	30	32	35	34	31	34	30	34	29	34	34	37	33.07 (29-37)
C4	36	35	36	36	37	32	34	36	38	36	36	30	37	39	38	35.73 (30-39)
C6	37	36	38	37	37	37	37	37	36	38	36	36	38	36	39	37.00 (36-39)
Average (range)	33.2 (24-37)	33 (28-36)	31.6 (21-38)	30.8 (25-37)	32.6 (27-37)	32.8 (26-37)	33 (26-37)	33 (28-37)	33.6 (26-38)	32.4 (26-38)	32.2 (26-36)	30 (27-36)	33.4 (24-38)	33.2 (21-39)	35.2 (31-39)	

Table 4.11 Correct phoneme scores at 15 dB dial for MWL in percentage

	LIST															
Patient ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Average (range)**
C1	82.5	82.5	75	65	75	85	85	82.5	85	80	72.5	70	85	90	77.5	79.50 (65-90)
C2	60	70	52.5	62.5	67.5	65	65	70	65	65	65	67.5	60	52.5	77.5	64.33 (52.5-77.5)
C3	90	82.5	82.5	75	80	87.5	85	77.5	85	75	85	72.5	85	85	92.5	82.67 (72.5-92.5)
C4	90	87.5	90	90	92.5	80	85	90	95	90	90	75	92.5	97.5	95	89.33 (75-97.5)
C6	92.5	90	95	92.5	92.5	92.5	92.5	92.5	90	95	90	90	95	90	97.5	92.5 (90-97.5)
Average (range)	83 (60-92)	82.5 (70-90)	79 (52.5-95)	77 (62.5-92.5)	81.5 (67.5-92.5)	82 (65-92.5)	82.5 (65-92.5)	82.5 (70-92.5)	84 (65-95)	81 (65-95)	80.5 (65-90)	75 (67.5-90)	83.5 (60-95)	83 (52.5-97.5)	88 (77.5-97.5)	

Table 4.12 Friedman test on speech audiometry scores at 15 dB dial for AWL

Test Statistics[a]

N	5
Chi-Square	19.584
df	14
Asymp. Sig.	.144

a. Friedman Test

Table 4.13 Friedman test on speech audiometry scores at 15 dB dial for MWL

Test Statistics[a]

N	5
Chi-Square	17.554
df	14
Asymp. Sig.	.228

a. Friedman Test

4.3.3 Homogeneity of word lists

Inter-item correlation analysis using the SPSS software package was done to measure the homogeneity and internal consistency of the lists. The single-measures intraclass correlation coefficient (ICC) in the form of Cronbach's alpha was taken as the indication for the internal consistency among the lists. High alpha corresponds to strong agreement and, therefore, high internal consistency, between the items tested. In this case, a strong internal consistency reflects the homogeneity among the lists.

The scores at 15dB HL using AWL were chosen to be the data used for the equivalence study as this level was estimated to produce scores between 40-60%, which should be located on the steepest slope of the speech audiometry curve, based on previous studies (Hirsh et al., 1952; Nissen et al., 2005; Harris et al., 2007). The steep curve denotes the

rapid increase of speech intelligibility relative to the increase of hearing level (Bench et al., 1979). The score range is commonly used in previous studies of speech audiometry, possibly due to its proximity to the score corresponding to speech reception threshold, which is 50% correct (Wang et al., 2007; Han et al., 2009). The Cronbach's alpha for an analysis of one on all 15 lists showed a value of 0.78, indicating a strong agreement among the lists (Table 4.14). The ICC value suggested high internal consistency among the lists, and therefore, indicated good homogeneity among the lists.

An exercise to try to achieve better Cronbach's alpha was done by eliminating lists that showed lower correlation in the inter-item correlation matrix. Five lists with inter-item correlation ≤0.8 (Lists 4, 6, 7, 12 and 15) were excluded from the internal consistency analysis. The resulting alpha value for the remaining 10 lists was 0.88 (Table 4.15).

Table 4.14 Cronbach's alpha (Intraclass Correlation Coefficient) for all 15 AWL lists

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.78	.537	.968	64.089	4	56	.000
Average Measures	.98	.946	.998	64.089	4	56	.000

Table 4.15 Intraclass Correlation Coefficient (Cronbach's alpha) for lists with inter-item correlation ≥ 0.8 for AWL

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.88	.702	.985	68.747	4	36	.000
Average Measures	.99	.959	.998	68.747	4	36	.000

As both Cronbach's alpha measures indicate strong agreement between the lists, and therefore strong internal consistency among the lists, it was decided that all 15 lists would

be included in the clinical validation study. The strong internal consistency and good homogeneity signified that the 15 lists were interchangeable and that the results gathered from any one of the lists were comparable with the results from the other lists.

The intraclass correlation coefficient for MWL was also calculated. The Cronbach's alpha for MWL measured using all 15 lists showed a value of 0.81, indicating strong agreement between the lists (Table 4.16). The strong internal consistency among the MWL lists indicated good homogeneity among the lists. Therefore, the 15 lists using the MWL version was also interchangeable and the results gathered from any list were comparable to those gathered from other lists.

Table 4.16 Cronbach's alpha (Intraclass Correlation Coefficient) for all 15 MWL lists

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.81	.573	.973	63.965	4	56	.000
Average Measures	.98	.953	.998	63.965	4	56	.000

4.3.4 Validity of acoustic content of the word lists

To ensure that the word lists are not just equally difficult but equal in content with the general Malay language, a comparison between the frequency speech spectrum and (LTASS) in Malay was made. The justification for using LTASS as a comparison is that it provides an estimate of the speech dynamics, particularly intensity as a function of frequency, of the language.

Tracks of the voice samples were normalised and then mixed to produce a composite. The frequency spectrum of the composite was plotted to produce the LTASS (Figure 4.2).

Figure 4.3 demonstrates the comparison between Malay LTASS and published LTASS by Cox and Moore (1988) and Byrne et al (1994). Large gap can be seen at frequencies

more than 800 Hz, with notable gaps at frequencies 1000Hz and 2500Hz, between the Malay LTASS and curves proposed by Cox and Moore (ibid.) and Byrne et al. (ibid.). The Malay LTASS is generally lower intensity-wise. No previous literature on Malay LTASS has been published to compare with the current results. However, similar disparities were demonstrated in the LTASS of several Asian languages as studied by Byrne et al (ibid.). These discrepancies will be discussed further in the next chapter.

To compare the LTASS with the word lists, frequency spectra of each of the word lists were analysed using Audacity and then plotted. Mean intensity at 24 frequency points between 43Hz and 16020Hz were extracted to produce a mean frequency-intensity contour of the word lists (Figure 4.4). Visual inspection showed that the frequency spectra of the lists were similar to each other. A superimposed LTASS curve also showed that the spectra of the word list closely match that of the LTASS, except at high frequencies (6000Hz – 10000Hz) where up to approximately 10dB gap between the average word lists and LTASS is observed. This suggests that the acoustic content of the lists mimics the acoustic content of the LTASS at least for the frequencies that are tested by pure tone audiometry, and therefore, to an extent, provides an example of the general Malay speech. To test this hypothesis, a statistical analysis on the closeness between the frequency spectra of the lists and the LTASS were made. The outcome of this test would objectively determine the closeness between the frequency spectra of the word lists and that of the LTASS. The closer the frequency spectra of the word lists in simulating the frequency spectrum of the Malay LTASS would imply a better representation of the Malay language in terms of its acoustic properties.

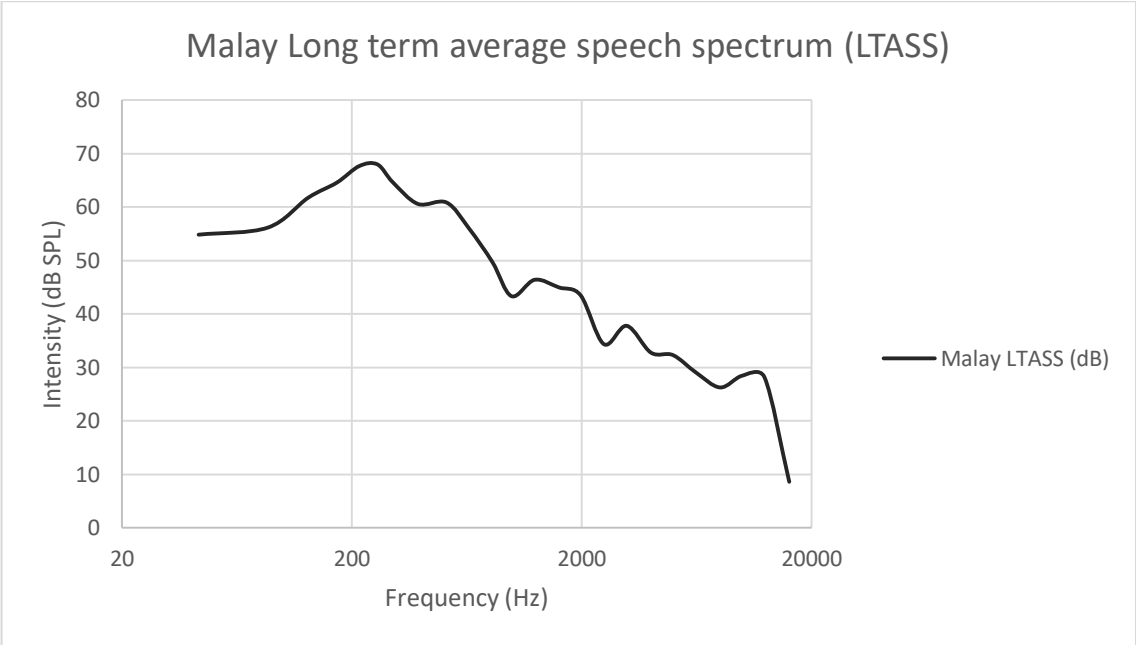


Figure 4.2 Malay long-term average speech spectrum (LTASS)

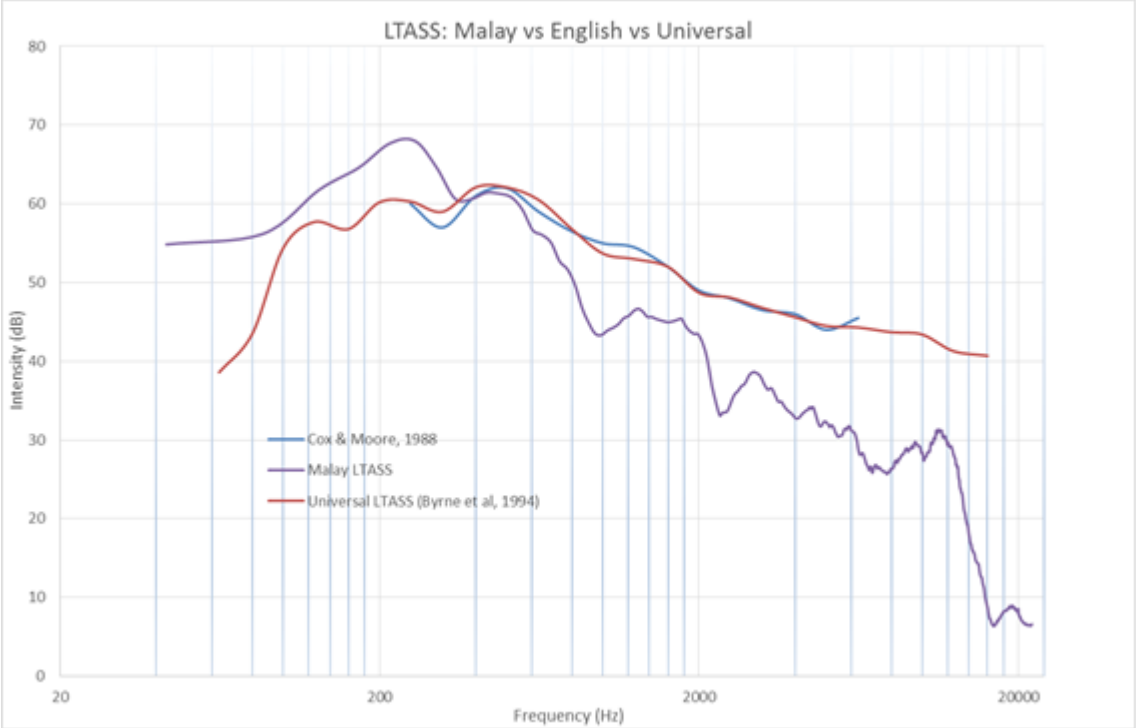


Figure 4.3 Comparison between Malay LTASS and published LTASS

To study the proximity between the LTASS and the frequency spectra of the stimuli in each list, a comparison between them was made. The frequency spectrum of each list was extracted and plotted using Audacity. Twenty-four frequency points between 43 Hz to 16020 Hz, as described in the experimental design, were selected as points of comparison. Repeated measures ANOVA was done between the frequency spectra of the word lists and the LTASS. The lists and Malay LTASS were labelled as 'tracks' in SPSS. Mauchly's Test of Spherecity indicated that the assumption of spherecity had been violated, χ^2=307.356,p=0.00, therefore, a Greenhouse-Geisser correction was used (Table 4.17). There was no significant effect of tracks on the frequency spectrum, F=1.229, p=0.302. This validated the consistency of the acoustic content throughout the lists. It also suggested that the acoustic content of the lists were similar to that of the Malay LTASS. It can also be deduced that the word lists contain the same acoustic content as the speech sample and, therefore, can be used as a representative of Malay language in the assessment of hearing.

Table 4.17 Repeated measures ANOVA on the LTASS and the frequency spectra of the word lists

Mauchly's Test of Sphericity[a]

Measure: Level

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Track	.000	307.356	119	.000	.303	.386	.067

Tests of Within-Subjects Effects

Measure: Level

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Track	Greenhouse-Geisser	102.113	4.538	22.504	1.229	.302	.051

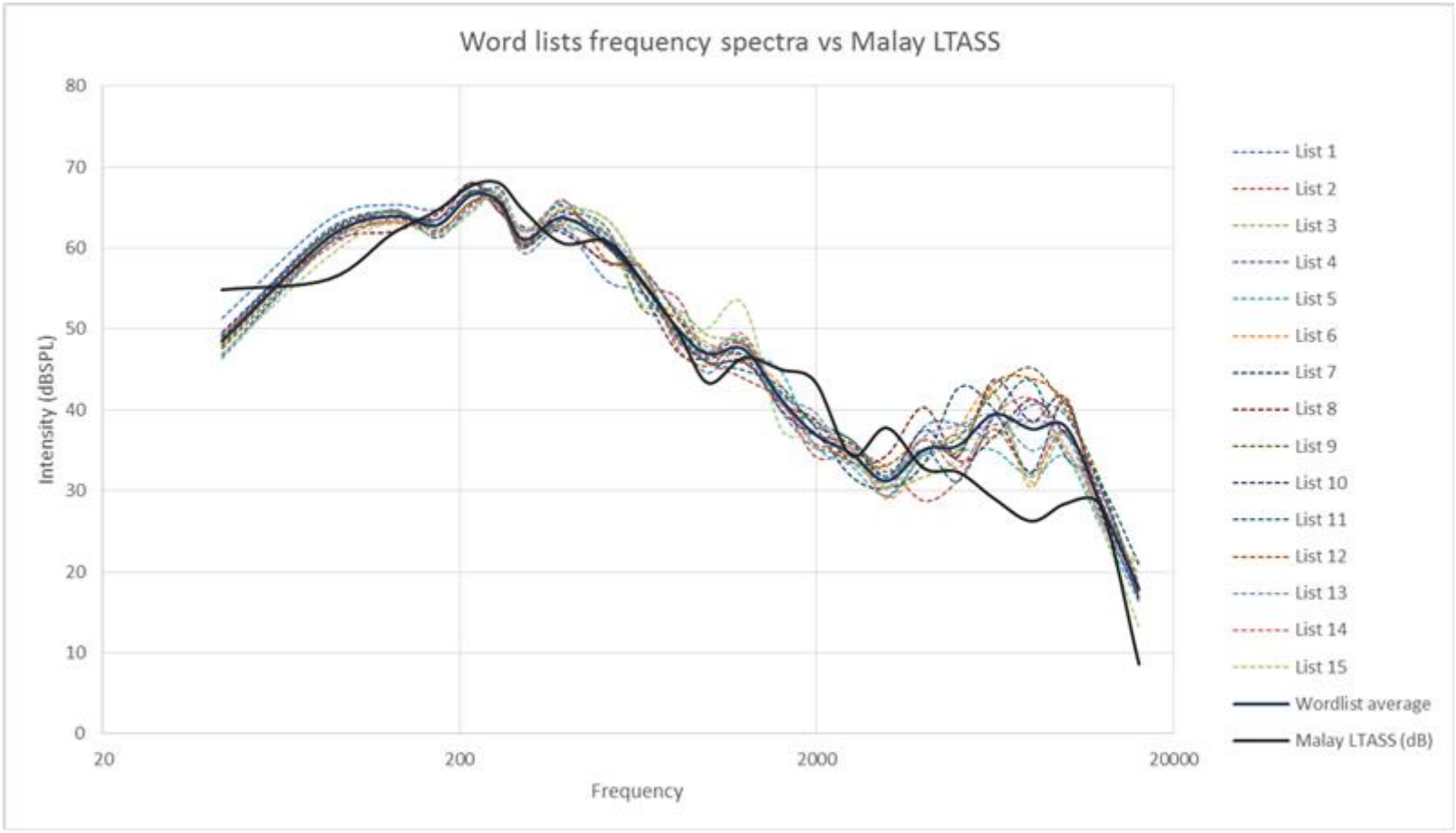


Figure 4.4 Comparison between the frequency spectrum of each list and Malay LTASS

For comparison, the phoneme distribution of the passages used for LTASS is presented in Table 4.18. The distribution of phonemes in the passages varies greatly with the distribution of phonemes in the corpus and in the word lists. This may be due to the variety of syllabic length in passages, which is not restricted to bisyllables of the CVCV form. The word lists and the corpus, which contain 50% vowels and 50% consonants due to the CVCV nature, show higher distribution of some of the vowels compared to the passages, namely for phonemes /ə/, /i/, /o/ and /u/. The passages, on the other hand, showed higher distribution of several phonemes, particularly /a/, /k, /m/, /n/ and /y/. This can be due to the affixes commonly found in Malay language, such as prefixes 'men-' and 'meny-', and suffix '-kan'.

It is worth noting that, from the psychoacoustics perspective, our hearing system analyses sounds, including speech sounds, by detecting their frequencies and intensities. This information will be carried out to the auditory cortex for processing and to be 'translated' into what we actually comprehend as speech. Based on this perspective and bearing in mind that LTASS is like a 'snapshot' of the physical characteristics of a language, it could be said that a set of word lists that is similar to the LTASS may represent the language, at least in terms of its physical characteristics. However, it is important to bear in mind that a language is not made of only frequencies and intensities; there are other elements that make a language, such as context and structure. But then again, from the view of audiology, the focus is given on the frequency and intensity of the speech sounds that can be recognised by the hearing system.

4.4 Discussion

The verification process in this study looks into the homogeneity and consistency of the word lists and tries to answer the question of whether the lists are equivalent to each other. Here, two aspects of the word lists were evaluated, the difficulty in audibility and the acoustic content. The study tries to see whether or not the lists can be used interchangeably, that is, if using any of the 15 lists under the same condition can generate the same results, as well as to see whether the lists have the same content acoustically.

The following sections will discuss the findings of homogeneity and internal consistency measurements done on the word lists, comparison between the acoustic properties of the word lists and comparison between the acoustic of the word lists and the Malay speech sample in the form of long-term average speech spectrum.

Table 4.18 Comparison between phoneme distribution of the word, word lists and the passages used in LTASS

	Proportion (%)			
Phoneme	**Corpus**	**MWL**	**AWL**	**Passages**
a	13.91	14.17	14.00	22.18
b	3.70	3.67	4.33	3.33
c	1.45	1.83	1.56	0.45
d	3.27	2.83	3.22	4.76
ə	9.73	10.17	8.67	6.28
e	1.92	1.67	1.67	0.76
f	0.57	0.50	0.33	0.08
g	2.36	2.33	2.67	4.39
h	1.35	1.67	1.33	3.10
i	11.99	12.17	12.33	7.49
j	2.56	2.67	2.33	0.91
k	4.18	4.83	4.78	6.36
l	4.95	5.33	4.56	2.95
m	3.43	2.67	3.33	5.22
n	2.49	2.17	2.22	7.95
o	2.86	2.33	3.11	1.06
p	2.90	2.50	2.67	2.27
q	0.13	0.00	0.00	0.00
r	4.98	5.17	5.22	3.79
s	4.68	5.00	5.00	4.54
t	4.48	4.83	4.89	3.93
u	9.56	9.50	10.22	5.22
v	0.30	0.00	0.00	0.08
w	0.91	0.83	0.78	0.30
x	0.00	0.00	0.00	0.00
y	1.14	1.17	0.78	2.42
z	0.20	0.00	0.00	0.15

4.4.1 Homogeneity and consistency of the word lists

The verification of the difficulty in audibility of the word lists was done through homogeneity and consistency assessments. Two statistical tests were employed; Friedman Test was used to evaluate the consistency between lists, while intraclass correlation coefficient was used to test the homogeneity of the word lists. The consistency analysis using Friedman's Test indicated no significant difference in the performance across the 15 lists for both AWL and MWL. Cronbach's alpha measures showed strong internal consistency for both AWL and MWL.

Prior to the homogeneity and consistency assessments, the participants were presented with AWL lists at two separate levels – 15dB dial and 40dB dial in order to obtain the correct phoneme scores. The lower presentation level was aimed to produce a score that ideally should be situated along the steepest line on the psychometric, or Performance-Intensity (P-I), function curve. Previous studies had set the range of scores at anywhere between 20% to 80% as having the steepest slope in a P-I function curve (Harris et al, 2007; Wang et al. 2007; Nissen et al., 2011). The higher 40 dB dial level was estimated to produce the maximum score the participants could achieve.

The speech audiometry scores at 15 dB dial presentation level ranged between 53.3% to 98.3% correct, with an average of 82.2% correct (SD=11.15). At 40 dB dial, the correct score ranged from 93.3% to 100%, with an average of 98.6% (SD=1.67). For the purpose of observing the similarity (or dissimilarity) between the current and previous studies, a sample of average correct scores vs level obtained from normal hearing participants in other studies is given in Table 4.19. Average correct scores in reference to the presentation level can vary widely and the differences in the correct scores in reference to the presentation level can be attributed to the recording intensity, which is the intensity level of the input signal digitally recorded into the compact disc or hard drive. The level of the input signal has direct effect on the output intensity, that is, the level of sound that is transmitted into the listener's ear. High level of input signal generates high output level and vice versa. This effect is manipulated by some developers of speech audiometry material in order to make test items equal in terms of difficulty in audibility, whereby the developers digitally adjusted the input levels of the tests items so that the output level of individual test item better matches the average output level (Nissen, et al., 2005; Nissen, Harris and Slade, 2007; Nissen et al., 2011)

In the current study, consistency and homogeneity of the word lists were used to verify both AWL and MWL word lists. Consistency analysis was also used to verify the reliability of the test items in previous studies, although they mainly employ repeated measures analysis of variance (ANOVA) (Wang et al., 2007; Han et al., 2009.; Nielsen and Dau, 2009). The current study chose to employ its non-parametric equivalent, Friedman Test, as the data fulfilled the required assumptions (Statistics.Laerd.com, 2015). In the previous studies, lists that were significantly different from the rest were identified through post-hoc analyses and taken out of the set (Wang et al., ibid.; Han et al., ibid.). The level selected for consistency, by way of difficulty in audibility, was 15 dB dial and the scores reflect the difficulty of each list. A higher score indicated that the words in the list used were relatively easy to recognise while a lower score indicated a more challenging list. Although there were no obvious patterns of the easiest or hardest list could be seen, Lists 15 and 12 gave the highest and lowest scores for 3 out of the 5 participants respectively. Consistency analysis using Friedman test showed no significant difference in the correct phoneme scores using any of the lists for both AWL and MWL versions, indicating that the choice of word lists had no effect on the phoneme scores and, therefore, post-hoc analysis was not required. This finding verified the word lists as having equal difficulty in audibility and, therefore, allowed the lists to be used interchangeably. The equal difficulty among the lists implied that, at a particular presentation level, using any list would generate similar result. This also signified that all lists could be included in the following clinical validation study. This also indicated that both AWL and MWL versions were fit for the clinical validation study.

Homogeneity or internal consistency was assessed using Cronbach's alpha, a type of ICC that can be used to measure homogeneity (McGraw and Wong, 1996). In this part of the study, Cronbach's alpha was selected as the measure to test the homogeneity in terms of difficulty in audibility of the word list. The finding showed that the Cronbach's alpha value was high. This indicates that the word lists had strong internal consistency, and therefore, had good homogeneity in terms of difficulty in audibility. This finding signified that the set of word lists were valid to be used interchangeably in a speech recognition test.

Table 4.19 Comparisons of average correct scores vs presentation level between previous studies and current study

Study	Correct score	Level	Test items
Current study	82.19% [a]	15 dB dial	Bisyllabic Malay words
Wang et al. (2007)	92% [a]	15 dB HL	Disyllabic (bisyllabic) Mandarin words
Nissen, Harris and Slade (2007)	50% [b]	0 dB HL [c] (female) 4.4 dB HL [c] (male)	Trisyllabic Taiwan Mandarin words
Lau and So (1988)	65.4% [a]	20 dB dial	Monosyllabic Cantonese words
Ashoor and Prochazka (1982)	50% [b]	22.2 dB	Monosyllabic Saudi Arabic words
Boothroyd (1968)	50% [b]	20 dB [d]	Monosyllabic English words
Mukari and Said (1991)	34.4% [a]	15 dB	Bisyllabic Malay words

[a] Average correct score
[b] Speech reception threshold
[c] Average intensity
[d] Modal value

The internal method differs from measuring the homogeneity through the P-I function slope, which would require more presentation levels to form the psychometric curve. According to Tavakol and Dennick (2011), Cronbach's alpha could be used to confirm the unidimensionality of the test items, and therefore, their homogeneity, provided that the test items are of the same concept or construct. This value quantify the reliability of the test items and decide whether or not the test items were able to measure consistently (Tavakol and Dennick, ibid.). Consistency in the results obtained using the word lists would increase the validity of the word lists

Two ICC measurements were done on the lists. One perused the whole set of lists while the other analysed only lists that showed inter-item correlation of ≥ 0.8. The second analysis was done to determine whether or not having lists that showed lower inter-item correlation of less than 0.8, in this case five lists, eliminated from the set would provide higher ICC. It is important to note that Cronbach's alpha values of 0.7 and above demonstrate good internal consistency (Bland and Altman, 1997; Tavakol and Dennick, 2011). Both full-set and high inter-item correlation set showed ICCs of 0.781 and 0.883

respectively. This means that removing the lower inter-item correlation word lists improved ICC by 0.102.

Although the high inter-item correlation set showed higher ICC indicating stronger internal consistency, and therefore higher level of homogeneity, compared to the full set, there is a question of whether having a smaller set of lists with higher ICC is better than having a set with more lists but with slightly lower ICC. One factor that need to be considered is item or stimulus familiarity. Stimulus familiarity affects the listeners' ability to recognise the items, hence, influences the scores. Mendel and Danhauer (1997) outlined several factors that affect stimulus familiarity, among them number and variety of stimulus items in the test. In the current study, the test items i.e. words are divided into fixed lists. The permanence of content in each list is needed to preserve the phonetic balance of each list. This feature limits the flexibility of the lists in terms of the ability to have alternative combinations of words within each list. This means when a certain list is presented to the listener, for example List 2, the items in that list remains as they are, no matter how many times the list is presented. To avoid stimulus familiarity from occurring, it is important to have a wide selection of test items, or in the case of this study, lists. This allows the tester to have adequate reserve of lists and avoid using the same list for multiple presentations. Taking this factor into consideration, it is thought that the benefit of having a wider choice of lists outweighs the higher level of consistency and homogeneity.

Homogeneity testing using ICC in this study measures the proportion of variance in the scores obtained using different lists (McGraw and Wong, 1996). Assessment was done based on the internal consistency, that is, how consistent are the results that are obtained between two different test items (Bland and Altman, 1997). Previous studies have used other statistical methods in testing the equivalence of their speech audiometry material, for example, two-way chi-square (Nissen et al., 2011) and ANOVA and SNK-Q (Wang et al., 2007). No previous study has been found to utilise ICC in their homogeneity study, therefore, no direct comparisons can be made. However, it is interesting to compare several of the methods that have been used in the equivalence analysis and the steps taken to produce a set of lists or words that are homogenous. Wang et al. (2007) used ANOVA and SNK-Q to indicate and identify if there was any of their 10 lists that might not be equivalent to the others in the set. The identified list was taken out leaving 9 lists to remain in the set. The same principle of excluding or omission was also applied in several other speech material development (Magnusson, 1995; Han et al.,

2009; Nielsen and Dau, 2009; Nissen et al., 2011). Another method is to adjust the level of intensity of the test items that are notably different from the others (Harris et al., 2007; Nissen et al., 2007; Nissen et al, 2011). This method is based on the psychometric function of normal hearing for speech, that is, by adjusting the intensity level the ease of word recognition can be adjusted as well. Adjustment to the test items means that the test items are preserved in the set and need not to be omitted. However, this method relies on observation, i.e. visual inspection on the dispersion of the P-I functions of the test items, rather than statistical analysis. The current method of homogeneity evaluation through intraclass correlation should be considered as an alternative to the earlier methods as it allows preservation of the test items supported by statistical data.

4.4.2 Validity of acoustic content of the word lists

In addition to ICC, the current study also looks into the acoustic content of the word lists. A comparison was made among the frequency spectra of the word lists as well as between the frequency spectrum of the word lists and the long term average speech spectrum of the Malay language. The rationale for this investigation is that the speech spectrum of each language is unique (Byrne et al., 1994), therefore it is important to have the current word lists to emulate as much as possible the Malay speech spectrum. Although language is made up by many components – phonology, syntax and semantics, among others – this part of the study attempts to study the extent of representation of the physical properties of Malay phonetics in the word lists. By having similar acoustic content as continuous speech, it can be assumed that the word lists may represent the sounds of the language.

4.4.2.1 Malay long term average speech spectrum (LTASS)

An attempt to establish Malay long term average speech spectrum (LTASS) was made using speech samples collected from male and female adult native speakers of Malay. The speech sample was based on two short compositions containing the phonemes of Malay language. The frequency spectra of the speech samples were analysed and plotted in an intensity-frequency function graph.

Initial evaluation showed a function generally similar to the LTASS of other languages (Byrne et al., 1994; Noh and Lee, 2012). The intensity increased at very low frequencies and peaked at between 200 and 300 Hz. The intensity gradually decreased towards the higher frequencies before declining steeply after 11000 Hz. The shape of the LTASS may be influenced by the gender of the talker, the language, the structure of the speech used in the sample (running speech, monosyllables etc.) as well as the style of speech, e.g. conversational speech vs speech by professional speakers/announcers (Benson and Hirsh, 1953; Byrne, 1986; Noh and Lee, 2012a; Noh and Lee, 2012b).

Malay language, as in other languages, has its own set of speech sounds, therefore, it was expected that the Malay LTASS has a unique contour. Visual inspection of the Malay LTASS in comparison with American English LTASS (Cox and Moore, 1988) and Universal LTASS (Byrne et al., 1994) showed marked differences at frequency lower than and higher than ≈300 Hz. Much higher intensities can be observed in Malay LTASS at frequencies lower than 300 Hz as compared to the universal LTASS and Cox's and Moore's (ibid.) LTASS. Investigations on phoneme distribution by Tan et al. (2009) and the current study showed that the Malay language contains around 40-50% vowels (/a/, /e/, /ə/, /i/, /o/ and /u/). This could have contributed to the higher intensities at the lower frequencies as compared to the English and Universal LTASSs. The effect of having different proportions of vowels and consonants, or even phonemes, to LTASS has not been studied before. However, the fact that the central frequencies for the first and second formants, F_1 and F_2, of most of the vowels are located upwards of 370Hz (Ling, 2002) may demonstrate that phoneme distribution may not be the cause of this difference. Several studies on the effect of gender to LTASS found that male speakers demonstrate significantly higher intensity levels for frequencies 160Hz and lower as compared to women (Cox and Moore, ibid.; Byrne et al., 1994; Noh and Lee, 2012a). The difference is attributed to the difference in voice pitch, also known as fundamental frequency or F_0, ranges between male and female voices. This raises the question of possible differences in the voice pitch of Malay speakers as compared to speakers of other languages. There is also a possibility, due to the formal setting of the voice sample recording, that the speakers unknowingly changed the tone or pitch of their voice. It has been found that voice and pronunciation exercises have an effect to the F_0 of both professional and non-professional speakers (Varosanec-Skaric, 2003). Repeated reading prior to the actual recording of voice sample might have affected the voice quality of the speakers in the current study. Slight differences can also be seen at lower frequencies in the comparison of LTASS between conversational speech made by

ordinary speakers and clear speech made by professional announcers (Noh and Lee, 2012b). These differences were attributed to the speech style of professional speakers – to achieve clearer speech, the vowels, which relatively have higher energies in the lower frequencies, are longer in duration when compared to non-professional speakers. Again, how this affects the LTASS were not discussed. Nevertheless, this could explain the difference between the Malay LTASS and the universal LTASS and the American English LTASS by Cox and Moore (ibid.). Speakers in the Malay LTASS were recruited among students and staff of International Islamic University Malaysia (IIUM) and were not known to have any experience as professional speakers. There was no indication of the experience as professional speakers in the universal LTASS and American English studies as well, however, the possibility of difference in speech styles between the speakers who participated in three studies could not be discounted.

The study on the LTASS of several languages from around the world had clearly demonstrated the effect of language and dialect on the LTASS (Byrne et al., 1994). The LTASS of three languages – Singhalese, Vietnamese and Arabic - showed similar variations at higher frequencies from the universal LTASS as the Malay LTASS, with the LTASS of the three languages being lower than the universal LTASS at higher frequencies. The reason for the difference could be due to the difference in phonemes used, as well as the difference in the frequency of occurrence of the phonemes. It is not known whether any of the three languages are similar to Malay, but the effect of language could be one of the main reasons for the dip in intensity at high frequencies and the rise at low frequencies. It is interesting to point out that only Russian language showed significantly higher intensities at frequencies lower than 300Hz compared to the universal LTASS (Byrne et al., ibid.). Byrne (1986) found that running speech presented considerably higher intensities at lower frequencies (≤ 500 Hz) as compared to nonsense syllables taken from the CUNY Nonsense Syllable Test. However, the difference was due to the decreased intensities at the lower frequencies for the nonsense syllables. What caused the differences was not discussed in the study.

4.4.2.2 Comparisons of acoustic content among the word lists and between the word lists and Malay LTASS

Comparisons among the word lists and between the word lists and Malay LTASS were made to verify that consistency of acoustic content of the word lists within the set as well

as with the Malay speech spectrum. The verification of the acoustic content within the set served to affirm that the word lists are not only homogenous in terms of difficulty in audibility but also in terms of their frequency spectra. Comparison between the word lists and the Malay LTASS served to verify that acoustic content the word lists are comparable with that of Malay general speech sample. Agreement between the frequency spectra of the word lists and the LTASS signifies that the acoustic content of the word lists are representative of the acoustic content of Malay language. Comparisons are made through visual inspection and repeated measures ANOVA.

A comparison between the Malay LTASS and the frequency spectra of the word lists showed basically identical curves between 150 Hz to 4000 Hz. At frequencies lower than 150 Hz, the word lists seemed to show higher intensities compared to the LTASS. This was consistent with the gender of the speaker effect; all of the test items were recorded using male voice while the LTASS were established using both male and female voices. As what that has been discussed in previous paragraph, male speakers demonstrate higher intensity levels at lower frequencies, particularly ≤ 160 Hz (Cox and Moore, 1988; Byrne et al., 1994; Noh and Lee, 2012a).

Between 150 Hz and 4000 Hz, the frequency spectra of individual lists as well as the average spectra of the list followed the shape of the Malay LTASS closely. This suggests the acoustic content of the word lists reflected the acoustic content of Malay running speech sample within the range of 150-4000 Hz, even though there are variations found in the phoneme distribution for the passages used for the LTASS.

Above 4000 Hz, the LTASS and the frequency spectra of the word lists showed a difference in intensity levels, with the word lists having higher intensities compared to the LTASS. The cause for this difference is unknown; however, there are three major differences in the process of recording of the LTASS voice samples and the test items that have potentially affect the curves. Firstly, the speech style employed in the recording of the test items varied considerably. The speakers for the LTASS were encouraged to read the texts with natural intonation, speed and stress. This differs from the style employed by the speaker for the test items; to have a uniform intensity for both of the syllables in each of the words in the word lists, the speaker was encouraged to pronounce the words with equal stress. This might have intensified the level of energy of the phonemes in the syllables not normally stressed in normal conversation or normal speech and therefore increases the intensity levels of some of the frequencies in the word lists' spectra. Secondly, the word structure used in the word lists was limited to

bisyllabic CVCV form. This contrasts greatly with the words used in the texts for LTASS, which is a mix of several word structures occurring in Malay. As suggested by Byrne (1986), different word structures produced different LTASS curves. Although Byrne's (ibid.) finding was more pronounced at the lower frequencies, the difference in language may have caused a slightly different effect in Malay. Thirdly, the recording set up was different between the word lists and the voice samples for the LTASS. To ensure quality recording and sound, the test items were recorded in a professional recording studio and underwent a mixing process in order to have a better, clearer and uniform sound. The voice samples for the LTASS, on the other hand, were recorded in a sound treated room (audiology booth) using a digital voice recorder. Normalisation and mixing was basic compared to those done on the word lists. The frequency response of the equipment, e.g. microphones, might also have an effect on the frequency spectra of both LTASS and the word lists.

Despite the differences, repeated measures ANOVA showed no statistically significant difference between the word lists and the Malay LTASS. This indicates the consistency in the acoustic content of the word lists and between the lists and the running speech sample. Although the bisyllabic words in the test material are not true representations of the Malay language as a whole, it can be said that the word lists provide a 'snapshot' of the acoustic content in terms of frequency and intensity of the Malay language. This is particularly important in hearing assessment as hearing loss configuration varies between patients with hearing impairment. Findings from speech audiometry using the word lists may provide information on the limitations in the ability of the listener to hear the speech sounds of the Malay language. However, it is important to note the limitations of the scope of representation by the test items, especially the differences at high frequencies as well as the elements of language, for example context and nonverbal cues, represented by the words as opposed to normal conversation.

4.4.3 Summary

Verification is a process to ensure that the product meets the specification. In this study, verification of the word lists was done through consistency and homogeneity analyses.

Two aspects of the word lists were studied; one, the difficulty in audibility, and two, the acoustic content. The aim was to see whether the word lists are equal in these two

aspects, which in turn, allow the lists to be interchangeable and still give reliable results. The findings show no significant difference found in the performance of listeners across the lists for both AWL and MWL versions, strong internal consistency and good homogeneity among the lists for both AWL and MWL, and no significant difference in the acoustic content across the lists.

Performance based on the correct recognition of phonemes of normal hearing listeners were used as the basis of the consistency and homogeneity studies on the difficulty of audibility of the word lists. Statistical analyses indicated that all of the word lists were both equal in difficulty and homogeneous. Consistency and homogeneity in the difficulty of audibility was expected as the familiarity (by way of frequency of occurrence) of words that were included in the lists was regulated during the development of the word lists. Dirks et al. (2001) had shown that frequency of occurrence of words in the language affects the performance in speech recognition, with higher occurring words correspond with better recognition by the listeners, and vice versa. Meaningful words that were used in the lists were standardised based on their frequency of occurrence in the word corpus. In addition, the use of nonsense words instead of unfamiliar but meaningful words provides better control on the familiarity factor. Statistical analysis had also shown that the acoustic content of the word lists were homogenous. This was anticipated as the word lists were phonetically-balanced, with equal distribution of phonemes across the lists. Each phoneme has its own unique frequency spectrum which allows recognition by the hearing system (Stelmachowicz et al., 2004; Liebenthal et al., 2005). Having phonetically—balanced word lists meant having similar number of occurrence of each phoneme in every list, resulting in homogenous composite frequency spectra across the lists. As the distribution of phonemes in the word lists reflect the distribution of phonemes in the word corpus, it was also anticipated that the acoustic content of the word lists would have some similarity to the acoustic content of the running speech sample that was used to produce the Malay LTASS. This was also proven statistically by the consistency shown between the word lists and the Malay LTASS.

CHAPTER 5 CLINICAL VALIDATION

5.1 Introduction

Clinical validity refers to the accuracy of a test or measure in predicting the condition it is supposed to detect (PHG Foundation, 2015). In other words, it refers to the ability of the test or product to 'do what it is expected to do'. A clinically valid test means that the clinicians can be confident with the outcome of the test; for example, if the test comes out as negative, the clinician can be certain that the condition it is testing is absent, and vice versa. In the current study, the speech audiometry using the developed material is expected to reflect the hearing level of the listeners.

There are five types of validity in a research project – content, predictive, concurrent, construct and face validity (Burns, 2000). Content validity describes the level of representativeness of a particular measure to the content of the aspect that is being measured, which had been discussed in Chapter 4. Prediction of a performance of a feature through assessments or techniques as the predictors requires predictive validity. Concurrent validity is similar to predictive validity except that, instead of having predictors to predict the future, the predictors predict the performance at present. Construct validity describes the validity of tests that measures aspects of human behaviour, also known as 'constructs'. Face validity reflects the validity of a technique when taken at face value, that is, whether the technique 'seems' to measure the aspect of concern.

In this study, there are two types of validity of interest, the construct validity and the concurrent validity. . The word lists are being used as an instrument to 'predict' the hearing acuity of the listeners, particularly their speech discrimination ability. Therefore, in order to establish the validity of the word lists in predicting hearing impairment, a study to establish the relationship between the speech audiometry results and the pure tone audiometry results is required. The aims of this validation study are to assess the construct validity using the performance-intensity function of normal hearing participants and the concurrent validity using participants with varying types and levels of hearing losses.

The main question to be answered in this part of the study was "How does bisyllabic Malay wordlists reflect the hearing in a speech audiometry?" The aim of this study is to clinically validate the bisyllabic Malay word lists in two main aspects; whether the word lists are able to reflect the different types of hearing conditions, and whether the word

lists are able to reflect the hearing level. Therefore, the objectives of this study was to recruit normal hearing volunteers and hearing impaired patients, perform hearing assessments including pure tone audiometry and speech discrimination test, establish the characteristics of speech audiometry curve, also known as the P-I function, of different types of hearing conditions, and establish the correlation between the speech audiometry threshold and the pure tone threshold. This chapter presents the methodology, research design, findings and discussions on the validation process of the development of the word lists.

5.2 Methodology and research design

5.2.1 Review of methods

Validation study is usually performed in two parts – one involving normal hearing participants and the other hearing-impaired participants, although some developers of speech audiometry material would only establish the norms for normal hearing participants (Ashoor and Prochazka, 1982; Nissen et al, 2007).

Methods of establishing the norms are similar throughout developers. However, two major patterns can be seen – the same (normal hearing) group of participants doing both verification and validation analyses (Boothroyd, 1968; Lau and So, 1988; Harris et al, 2007; Nissen et al, 2007), or two separate normal hearing groups for verification and validity studies (Ashoor and Prochazka, 1982; Wang et al, 2007). In the current study, three new groups of participants were recruited for the clinical validation. The clinical validation employs a different research design as compared to the verification study, therefore it was decided that a new group of normal hearing participants would be selected in establishing the norms. The other two groups, participants with conductive hearing losses and participants with sensorineural hearing losses, would also be selected in the clinical validation process. The inclusion of hearing impaired participants in the validation process will be discussed below.

Throughout the literature, speech reception threshold (SRT) has been used as the main parameter that is used to validate speech test material. SRT is defined as the level of intensity required to obtain 50% of the maximum correct score, and is the standard measure used to describe the result of a speech audiometry (Boothroyd, 1968;, 1988;

Bess and Humes, 2003; Boothroyd, 2008). The SRT is also used in comparison with the pure tone hearing thresholds, both in research as well as in the clinical setting. In keeping with the academic and clinical practices, the SRT would be used as the main parameter in establishing the norms for the bisyllabic Malay word list.

Participant criteria are also similar throughout the studies. In order to validate the word lists for clinical use, the participants should be native speakers of the language, in line with the intended target user of this speech audiometry material. Establishing the norms require normal hearing participants with pure tone hearing thresholds within normal limits. Previous studies have had different definitions of normal hearing thresholds; several studies defined normal hearing as having pure tone thresholds equal to 20 dBHL or less at frequencies 250 Hz to 8000 Hz (Boothroyd, 1968; Ashoor and Prochazka, 1982; Lau and So, 1988; Killion et al, 2004), while others set the criteria higher by requiring thresholds to be 15dB HL or less (Nissen et al, 2007; Harris et al, 2007; Wang et al, 2007; Nielsen and Dau, 2009). There was a wide range of the number of participants recruited, from as low as less than 10 to more than 100 (Lau & So, 1988; Han et al, 2009). Inclusion criteria may include normal tympanometry and normal acoustic reflex thresholds while the exclusion criteria may include otologic and hearing disorders (Harris et al, 2007; Wang et al, 2007; Han et al, 2009). In the current study, the stricter criteria for pure tone hearing thresholds in normal hearing participants were elected to allow for the ±5 dB variability in pure tone audiometry. To ensure that the study has sufficient statistical power, sample size determination was done using the G-Power software.

Although most of the studies in the literature only involve normal hearing participants, the current study decided to also employ hearing impaired participants. Two groups of hearing impaired participants were recruited, those with conductive hearing loss and those with sensorineural hearing loss, determined through pure tone audiometry. The justification of having a wider range of hearing configurations in the study was that speech perception could be altered with hearing loss, therefore altering the results of the speech tests. An example is the rollover effect at high presentation levels in the results obtained from participants with acoustic neuroma and retrocochlear hearing loss while being absent in normal hearing participants or participants with conductive hearing losses (Hannley and Jerger, 1981) . Several previous studies employed both normal hearing and hearing impaired participants, with the degree of loss included in these studies ranged from mild to severe (Palmer et al, 1991; Peters et al, 1998; Wang et al,

2007). Several of the studies had only involved sensorineural losses (Peters et al, ibid.; Wang et al, ibid.), although Tillman and Carhart (1966) had included participants with otosclerosis, a condition that is usually related to conductive hearing loss. It was decided that cases with sensorineural hearing loss and conductive hearing loss are included in this study as it would provide better insight to the patterns of speech audiometry in hearing impaired listeners.

The scoring in this clinical validation study consisted of two calculations, one for AWL and another for MWL. It is important to highlight that this scoring method is different from the commonly used calculation which usually only employ single calculation for each tested list and includes all the test items in the list. The reason for having an additional calculation to the score was to explore the effect of having additional nonsense words in the test material as opposed to the more commonly used 'meaningful words only' lists. The double marking scheme, although unconventional, has been utilised in a previous study of speech audiometry material. Lau and So (1988) had devised a similar method of double calculation with their set of Cantonese word lists by having a marking scheme for the vowels and consonants and a separate scheme that included the vowels, consonants and tone. This was especially important for the hearing impaired group, as recruitment of participants with similar levels, types and configurations of hearing loss might be difficult and time consuming. Testing the participants using the all word lists and meaningful words-only lists in two separate session was also considered, but it was thought that the memory effect that may influence the participants' responses. Separate calculations for AWL and MWL using a single set of responses should eliminate this factor.

5.2.2 Research design

The following sections outline the research protocol involved in the clinical validation study. The details include the participants and the assessment procedures as well as the sample size calculation based on the pilot study.

5.2.2.1 Participants

This study employed the three-group participant method, which consisted of normal hearing participants and participants with conductive hearing loss and participants with sensorineural hearing loss. Participation was voluntary; those who were interested were given an information sheet and a consent letter to sign, indicating that they agreed to participate in the study. Those who consented, also by writing, were given a questionnaire to complete. Each were given an identity code and their personal identification details were kept separately and confidential.

The ethics application for the selection of both normal hearing and hearing-impaired participants was approved by the Research Ethics Committee, Faculty of Health and Life Sciences, De Montfort University.

Normal hearing participants

Recruitment of normal hearing participants was based on convenient sampling. Advertisements to call for participants were put up in and around the Faculty of Allied Health Sciences (FAHS), International Islamic University Malaysia (IIUM), Kuantan, Malaysia. Participants were the staff and students of FAHS as well as acquaintances of the researcher.

Hearing impaired participants

Recruitment of hearing impaired participants was based on identification of potential candidates from two audiology clinics' databases. Clinics involved were Hearing and Speech Clinic, IIUM and Audiology Unit, Department of Otorhinolaryngology, Tengku Ampuan Afzan Hospital (HTAA), Kuantan, Malaysia, an associate hospital of IIUM. Identification of potential patients in the Hearing and Speech Clinic, IIUM was done by two means, first, by looking at the appointment list of follow-up patients for the period of four months (June to September 2013); and second, by identification of new case patients who came for hearing assessment within the same time period and fulfil the selection criteria. The identification of potential patients in the Audiology Unit, HTAA was done by looking at the appointment list of follow-up patients for the period of one month (September 2013). Other methods of recruitment in the form of advertisements through the social media (Twitter and Facebook) were also employed; however, there were no response received.

5.2.2.2 Preliminary assessment

Preliminary assessment was done on all hearing participants. It consisted of otoscopy, tympanometry and pure tone audiometry. Air-conduction and bone-conduction pure tone audiometry were performed on both ears using the modified Hughson-Westlake procedure and were done at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz. Bone-conduction pure tone audiometry at 250 Hz was done following the IIUM Audiology testing protocol (International Islamic University Malaysia, 2014) as opposed to the bone-conduction recommended procedure by British Society of Audiology (2011). The participants were also asked to fill in a short questionnaire on their hearing, otologic and general health.

a) Selection criteria for normal hearing participants

Inclusive criteria were normal otoscopy, tympanometry values of between +50 to -50 daPa for middle ear pressure (MEP) and between 0.3 to 1.4 mmho for static acoustic admittance. Pure tone hearing thresholds levels should be ≤15dBHL at all frequencies. Exclusion criteria were significant/diagnosed tinnitus, history of otologic and hearing disorders, significant exposure to noise, and recent (within the past 3 weeks of assessment) history of illness that may lead to hearing problems (e.g. flu, colds, cough). The better ear was selected for further testing.

b) Selection criteria for hearing impaired participants

The hearing impaired candidates were further divided into two groups. Candidates with sensorineural hearing loss with degrees ranging from mild to profound formed the first group and candidates with conductive hearing loss, also ranging from mild to severe, formed the second. Otoscopy and tympanometry should be consistent with the type of hearing loss experienced by each participant. Exclusion criteria include tinnitus, auditory neuropathy and central auditory processing disorder. The better ear was selected for further testing, except in cases with unilateral hearing loss.

5.2.2.3 Speech reception test

Speech reception test was done to determine the performance-intensity (PI) function for participants using the AWL. The PI function should be able to indicate the threshold of

speech reception as well as giving an indication of the level of speech discrimination of the listener. Speech reception threshold is defined as the intensity level at which 50% of the maximum score level was achieved. The level of speech discrimination can be estimated by looking at the maximum score.

a) Test instruction

The participants were first briefed on the test procedure. The participants were instructed to listen to the bisyllabic words and repeat the words they hear. Guessing was encouraged.

They were then given a sheet containing the instructions to the test and were asked if they need any clarification. The instructions were in Malay and carried the translation, "*You will be presented with several words. The words may be loud or soft. Please repeat the words you hear, no matter if it carries any meaning or not. You will be given time to repeat the word after it is presented. You are also encouraged to make a guess*."

b) Procedure

All assessments were done in certified audiometric booths at the Hearing and Speech Clinic, IIUM. Madsen Itera II audiometers were used in the pure tone audiometry and speech audiometry testing. All used audiometers were calibrated regularly by the clinic administration according to the manufacturer's instruction. Daily calibration was also done before the first session of testing every day. Due to clinic policy and lack of equipment, acoustic calibration in between assessment sessions (e.g. midway of data collection) could not be done.

The ethics approval for the test procedure was gained from the Research Ethics Committee, Faculty of Health and Life Sciences, De Montfort University.

Before any testing was begun, the audiometer was calibrated for the speech audiometry material CD. The recorded 1000 Hz calibration tone was played on the CD and the audiometer intensity dial was adjusted so that the VU display was set at 0. Two sets of different methods were designed for the participants, depending on whether they had normal hearing or hearing loss. Results were explained to the participants at the end of each session.

I. Normal hearing participants

To establish the P-I function, the normal hearing participants were assessed for speech recognition score at 10 different levels: -5, 0, 5, 10, 15, 20, 25, 30, 35 and 40 dB dial. Different lists were used in each level. The list that was presented at each level were chosen randomly. No list was used twice in each session. The presentation began with the highest intensity level (40 dB dial) and decreased by 5-dB steps for the following lists. No practice list was given. The participants were allowed a short break in between lists.

II. Participants with hearing loss, sensorineural or conductive

The method of speech audiometry was similar to that of the normal hearing group. The difference was on the levels of presentation. In this group, the level of initial presentation as well as the subsequent presentations was dependent on the average pure tone thresholds at 500, 1000 and 2000 Hz. These frequencies are commonly used combination in pure tone averages (PTA) (Humes, 2002; Nissen et al., 2007; Scollie, 2008; Nissen et al., 2011; The average was then added to each of the 10 different levels that were presented in the normal hearing group to obtain the presentation levels of each of the participants with hearing loss. A formula for the levels of presentation for these participants is as follows

(Average HTL at 500, 1000 and 2000 Hz) = PTA

Presentation levels:

PTA-5 dB dial, PTA+0 dB dial, PTA+5 dB dial….PTA+40 dB dial.

The list for each presentation was also chosen randomly and no list was used twice in each session. The presentation began with the highest intensity level (PTA+40 dB dial) and then decreased by 5-dB steps. However, to protect the participants hearing from very high level of sound, the level of presentation was capped at 95 dB dial. No practice list was given to the participants. The participants were allowed a short break in between lists.

c) Scoring

The response from participants were noted on the response sheet and scored afterwards. Scoring was based on phonemic scoring, with each phoneme scored individually. Each phoneme was given a score of 1, and each word item carries a maximum score of 4.

Two calculations were made for each participant, AWL scores and MWL scores. AWL scores were the total scores obtained from all words in each of the tested lists. In this scoring method, each list would carry the maximum score of 60 (15 words x 4). MWL scores, on the other hand, were the total scores obtained from only the meaningful words in each tested list and carry the maximum score of 40 (10 meaningful words x 4). The scores for each tested list and level (both AWL and MWL) were then compiled in an Excel file for further analysis.

d) Analysis

The clinical validity was determined through three assessments. First, the relationship between the PI function curve and the pure tone audiogram in all groups was analysed and described subjectively in terms of the marker points, such as half peak level (HPL) and maximum speech recognition score (MSRS), and shape. Next, the correlation between pure tone and speech reception thresholds were established. Several combinations of pure tone threshold (PT) average were tested for the correlation to establish the PT average that best correlate with the HPL. Thirdly, the normal range as well as the 95% confidence interval for the HPL was established and its sensitivity, specificity and predictive values measured.

5.2.3 Pilot study and sample size determination

A pilot study was carried out following the research design described in 5.2.2. The aim was to assess the feasibility of the research design as well as to collect the data required to determine sample size.

Twenty-two normal hearing participants were recruited for this purpose. All participants had normal hearing threshold and fulfilled the criteria outlined in the research design. The pilot study was carried out at the Hearing and Speech Clinic, IIUM, the same venue as the main study.

The research design for the clinical validation was designed based on routine hearing assessment protocol. It was found that each testing session lasted between 45 minutes to 1 hour. Each session started with the participant filling out the questionnaire on health history, followed by a briefing of the aims and objectives of the research as well as the flow of tests. The tests started with otoscopy, followed by tympanometry, pure tone audiometry and speech audiometry. Prior to otoscopy, the clinician conducting the tests

would verify the answers given in the questionnaire. Results were explained to the participants at the end of each session.

The research design was found to be feasible for a larger scale of study. However, based on the outcome of the pilot study, several modifications were made to the flow of each session:

i. The questionnaire was presented using the interview method instead of having the participants write down their responses. This modification was found to save the time taken for the participants to answer the questionnaire in writing and the tester to confirm the answers.
ii. The research briefing was given before the start of the interview and tests. It was found to be more effective in building rapport with the participants at the beginning of a session.

The data collected from the pilot study were also used to determine sample size. Sample size determination was calculated using the GPower 3.0.10 software. The correlation coefficient value between the pure tone threshold average of 500, 1000 and 2000 Hz and the speech reception threshold (SRT) were used to calculate the minimum sample size. A minimum sample size of 14 was suggested to achieve a statistical power of at least 0.95.

As there was no significant change to the research design and test protocol, it was decided that the data collected during the pilot study to be included in the analyses of results of the main study.

5.3 Results

5.3.1 Volunteers and patients

Twenty-six normal hearing adult participants, 10 males and 16 females, participated in this part of the study. Participants were recruited using convenience sampling method. Participants were drawn from a pool of staff and students of International Islamic University Malaysia (IIUM) as well as friends and acquaintances. Participants' age range was between 19 and 38 years old at the time of testing (average: 26 years, SD: 5).

From the questionnaire, only one of the normal hearing participants reported a suspicion of hearing loss. Sixteen participants out of the 26 had undergone previous hearing test. Five of the participants reported a history of occluded ear canal and one reported a history of glue ear, all of which had happened more than 3 months before testing and was already resolved by the time of testing. One of the participants reported of having tinnitus in both ears but was not experiencing any at the time of assessment.

The sensorineural hearing loss (SNHL) group was made of sixteen participants with sensorineural hearing loss. Participants were recruited through convenience sampling method from the patient list at the IIUM Hearing & Speech Clinic, Kuantan and Hospital Tengku Ampuan Afzan, Kuantan, Malaysia.

There were eleven males and 5 females in the SNHL group. Their age range was between 19 to 70 years old, with average age of 47 years (SD: 13). Fifteen considered themselves to have hearing loss while one participant, who was a new patient of the IIUM clinic, only suspected the loss. One participant reported having the loss for less than a year, 3 between 3-5 years and 4 each of between 1-3 years and more than 5 years, while the rest of the participants did not provide an answer. Six of the participants claimed to have rapid hearing loss and 4 gradual hearing loss, while 11 claimed to have postlingual onset of impairment and none reported prelingual deafness. Five of the participants were fitted with hearing aids on at least one ear.

The conductive hearing loss (CHL) group was made of fourteen participants. Participants were also recruited through convenience sampling method from the patient list at the IIUM Hearing & Speech Clinic, Kuantan and Hospital Tengku Ampuan Afzan, Kuantan, Malaysia.

The CHL group consisted of 8 males and 6 female participants. Their age range was between 21 to 54 years old, with the average age of 30 years (SD: 11). All of the CHL participants considered themselves to have hearing loss. Nine reported having the loss for less than a year, and two claimed to have had the loss for 1-3 years and 3-5 years respectively. Three participants reported gradual loss and another three rapid decline of hearing. All of the participants experienced the hearing loss postlingually.

Among reported medical history related to loss were occluded external ear canal (5 participants), ear discharge (5), glue ear (3), perforated ear drum (1) and chronic suppurative otitis media (CSOM) (1). Some of the participants reported more than one

medical symptoms. Three of the participants had undergone otological treatment and/or surgery.

The number of participants, average age and male to female ratio of each group are summarised in Table 5.1.

Table 5.1 Summary of baseline data of each group of participants

Group	Number of participants	Female	Male	Average age (SD)
Normal hearing	26	16	10	26 (5)
SNHL	16	5	11	47 (13)
CHL	14	6	8	30 (11)

Following the preliminary pure tone audiometry, one ear was chosen as the test ear for each of the participants. For the normal hearing participants, selection was based on the participants' ear preferences. For participants with hearing loss, selection was based on the pure tone thresholds obtained during the preliminary assessment:

i) for unilateral loss, the ear with the loss was the test ear,
ii) for assymetrical hearing loss, the ear with the better thresholds was chosen as the test ear to reduce the need for masking,
iii) for symmetrical hearing loss, selection was based on the participant's preference.

The following pure tone audiometry and speech audiometry results represent findings of the test ear.

All normal hearing participants showed pure tone hearing thresholds (HTLs) of 15 dB HL or lower at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz, indicating hearing thresholds within normal limits across the frequencies. Figure 5.1 shows the average pure tone HTLs in an audiogram form, accompanied by the minimum and maximum HTLs for each tested frequency. Table 5.2 shows the values mean pure tone HTLs and their respective standard deviations across the frequencies. All participants fulfilled the criterion for hearing thresholds set earlier. It is important to note that the average hearing threshold levels for the participants were not 0 dB HL for each of the tested frequencies. This could be due to the size of the group of participants, although several contemporary studies have reported similar findings of higher than 0 dB HL thresholds for their participants (Lau and So, 1988; Nissen et al., 2005; Nissen et al., 2007; Wang et al.,

2007; Nissen et al. 2011). Average HTLs according to frequency were the lowest at 4000 Hz (4 dB HL), which can be used to indicate that the participants were not affected by noise-induced hearing loss or presbyacusis at the time of testing. The highest average HTL was recorded at 500 Hz (9 dB HL). Although elevated thresholds at lower frequencies may indicate conductive hearing problems, all of the participants had also undergone and passed otoscopic examination and tympanometry, which would exclude any suggestion of conductive problem.

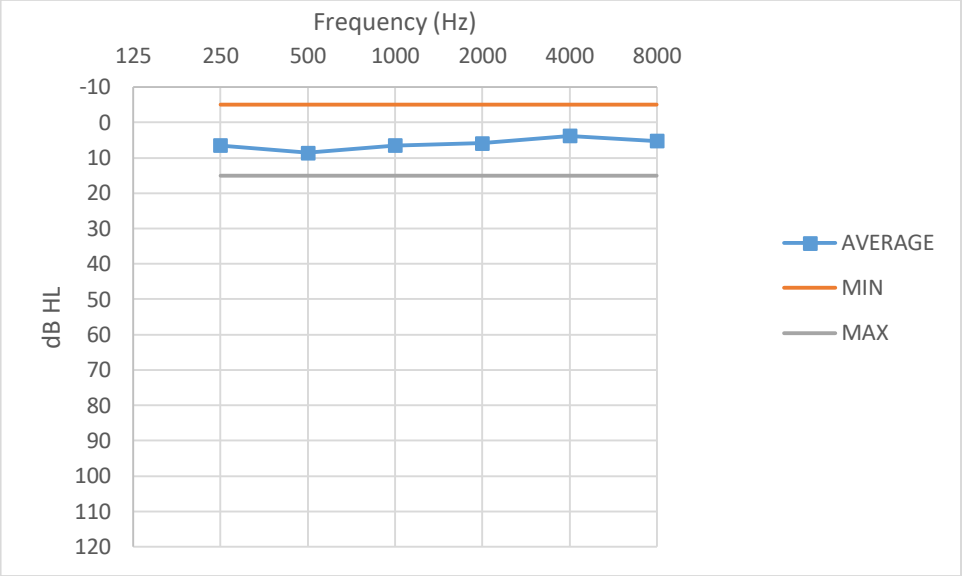


Figure 5.1 Average hearing threshold levels for normal hearing participants with the minimum and maximum levels

Table 5.2 Mean pure tone HTLs across the test frequencies for normal hearing participants

	Frequency (Hz)					
	250	500	1000	2000	4000	8000
Mean HTL (dB HL)	7	9	7	6	4	6
Standard Deviation	6	5	6	6	6	8

The average air conduction (AC) hearing threshold levels of SNHL group showed a slightly sloping configuration with mild-to-moderate level of hearing loss (Figure 5.2). The bone conduction thresholds showed similar configurations as the AC thresholds, with A-

B gaps less than 10 dB at 500, 1000 and 2000 Hz. However, the air conduction-bone conduction (A-B) gaps at 250 and 4000 Hz displayed gaps larger than 10 dB (Table 5.3). This can be justified by the audiometric limits at those frequencies; the audiometer limits the output at 250 and 4000 Hz at 25 dB HL and 75 dB HL, respectively. In addition, the vibrotactile limit at 250Hz can be reached at a level as low as 25 dB. These limits resulted in an apparent A-B gap in SNHL cases with higher AC thresholds.

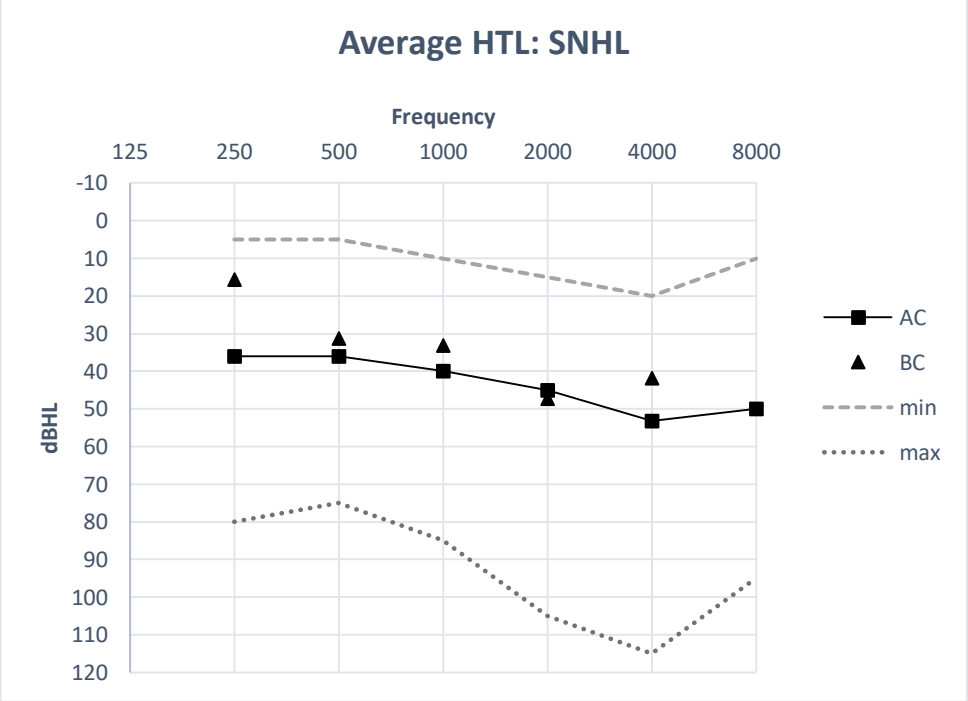


Figure 5.2 Average hearing threshold levels for participants with SNHL

The average CHL showed a flat, mild hearing loss (Figure 5.3). The average BC thresholds were within normal limits across the frequencies, with A-B gaps ranging between 21 to 35 dB for all tested frequencies, consistent with the expected findings for CHL. (Table 5.4).

Table 5.3 Summary of mean pure tone HTLs in participants with SNHL

		Frequency (Hz)					
		250	**500**	**1000**	**2000**	**4000**	**8000**
Air Conduction	Mean (dB HL)	36	36	40	45	53	50
	SD	21	20	22	26	26	27
Bone Conduction	Mean (dB HL)	16	31	33	47	42	
	SD	8	19	21	23	22	
A-B Gap	Mean	20	5	7	-2	11	
	SD	16	8	9	10	16	

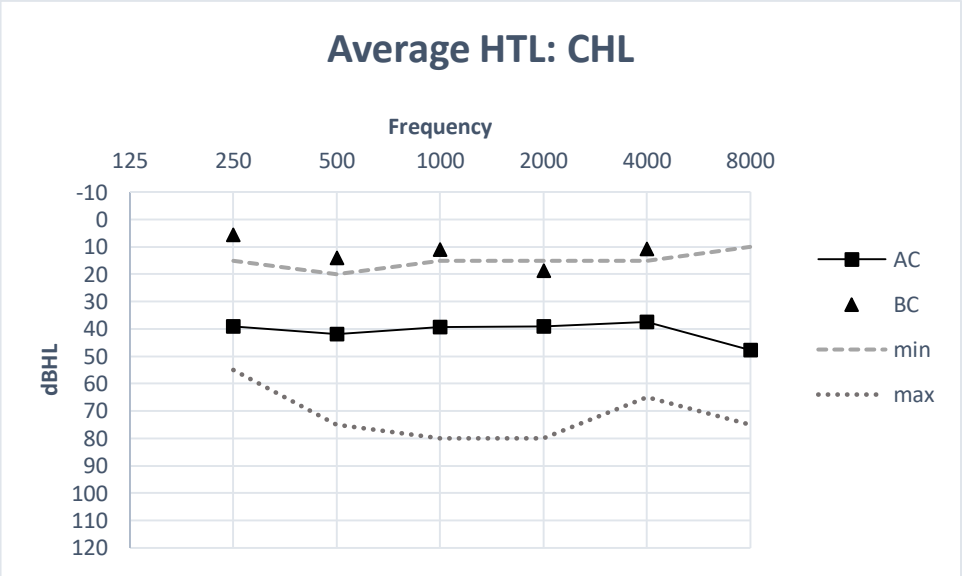


Figure 5.3 Average pure tone hearing thresholds in participants with CHL

5.3.2 Construct validity through construction of normative speech recognition score

The main objective for this part of the study was to establish the construct validity based on the normal range of the readings that can be extracted from the results, such as speech reception threshold (SRT) and maximum speech recognition score (MSRS). For

the current study, two versions of measures were analysed – one using AWL and another using the MWL.

Table 5.4 Summary of mean pure tone HTLs in participants with CHL

		Frequency (Hz)					
		250	**500**	**1000**	**2000**	**4000**	**8000**
Air Conduction	Mean (dB HL)	39	42	39	39	37	48
	SD	11	15	15	15	15	18
Bone Conduction	Mean (dB HL)	6	14	11	19	11	
	SD	9	9	11	12	9	
A-B Gap	Mean	35	30	30	21	28	
	SD	15	13	12	14	13	

5.3.2.1 All words lists (AWL)

A performance-intensity curve was constructed from the mean speech recognition score (SRS) at each presentation level. The curve followed the S-shape of typical normal discrimination curve as described by Boothroyd (1968) with three visible stages, a gradual increase at low presentation intensities followed by a steeper climb as the presentation intensity increases and ending with a plateau. Figures 5.4a and 5.4b show the performance/intensity function (P-I function) for all normal hearing participants using the AWL.

One participant, L05, showed a markedly lower correct score across the presentation levels (Figure 5.4a). The participant fulfilled all the inclusion criteria (otoscopy, tympanometry, pure tone audiometry and health history) for participant selection. Due to prior arrangement at the clinic, the assessment for participant L05 had to be done in an audiometric booth different from all the other participants (normal hearing or otherwise). To minimise bias in the data, it was decided that data collected from participant L05 would be excluded from subsequent analysis.

The reference range with 95% confidence interval for the P-I function using AWL was calculated using the results obtained from the normal hearing participants (Table 5.5). In order to establish the values for the reference range, the mean P-I function curve was plotted. Two additional curves were added, each one was built from mean correct scores

plus 2 standard deviations (mean+2SD) and mean correct scores minus 2 standard deviations (mean-2SD) respectively for each of the presentation levels (Figure 5.4b). The widest variation of scores can be seen at levels between 5 to 15 dB dial, with narrower range of scores seen at presentation levels of 0 dB dial and lower and 20 dB dial and higher. Although the level of initial audibility varies greatly, most of the participants had reached plateau at presentation level of 30 dB dial (Table 5.5). The lower and upper 2SD were limited to the actual minimum and maximum possible correct scores, which were 0% and 100% respectively, despite the calculation that resulted in lower or higher scores.

The half optimum speech reception threshold level or half peak level (HPL) is a measure used to find the speech reception threshold (SRT). These two terms are often used interchangeably; however, in this study, the term HPL would be used more as a reference point on a P-I function curve whereas SRT would be used in the correlation with the pure tone threshold.

Two methods of HPL calculation was done based on the findings. The first set was calculated based on the average P-I function curve as well as the upper and lower 2SD curves displayed in Figure 5.4b. The HPL of each of the curves were then calculated by finding the intensity level corresponding to the halfway point between the lowest and the highest scores of each curve. The levels were labelled as 'Curve HPL' in Figure 5.4b. It was found that the half peak levels of the mean curve is 10 dB dial, whereas the speech reception threshold levels for the lower 2SD curve and upper 2SD curve were 5 dB dial and 14dB dial respectively. It is important to note that the lower the level, the better the threshold, which means the hearing is better, and vice versa.

Another set of thresholds were calculated using the mean of individual participants' half peak levels. Each of the participants' HPLs were calculated and the total averaged. The mean HPL was 10 dB dial, equal to the HPL given by mean P-I function.

The reference range, with 95% reference interval, of the mean HPL was also calculated (Table 5.5). The lower limit of the normal range (mean-2SD) was 4 dB dial and the upper limit (mean+2SD) was at 17 dB dial. The normal range using the individual participants' HPLs were wider than the curve HPLs. The graphic representations of the HPLs in Figure 5.4b showed that the upper and lower limits of the HPL (not the curve HPLs) did not correspond to the half peak point of the curves. The difference would be discussed further under the Discussion section.

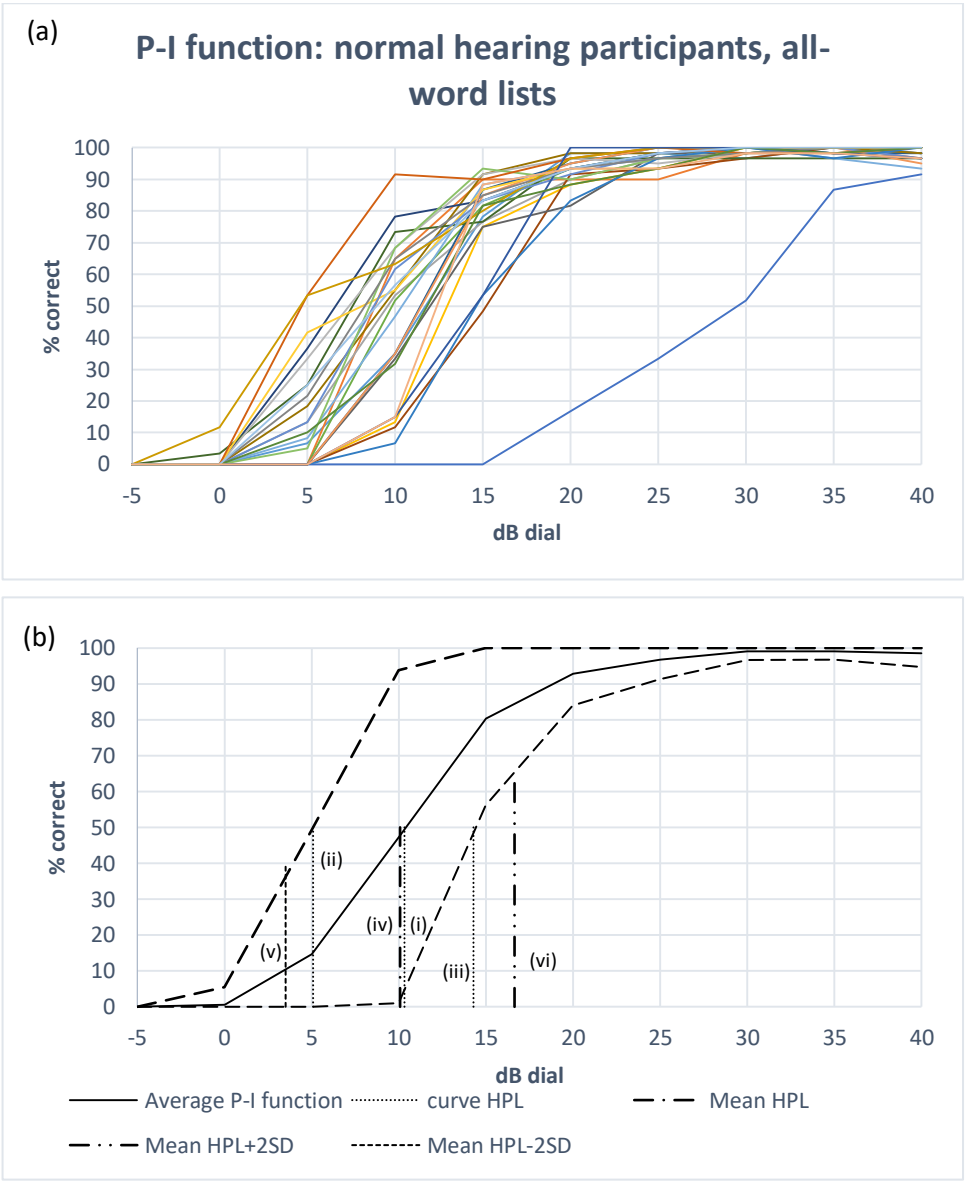


Figure 5.4 Performance/intensity (P-I) function for normal hearing participants – all-word lists: (a) shows the P-I function for all participants, (b) shows the mean P-I function curve (in solid line) and the curves of upper and lower 2 standard deviations (SDs) (in dashed lines). Six half peak levels are noted on the diagram: (i)-(iii) showed the half peak levels of the average P-I function curve and the HPLs of its upper and lower 2SDs; (iv)-(vi) showed the calculated mean of the participants' HPL and its lower and upper 2SDs.

Table 5.5 Mean correct scores for normal hearing participants using bisyllabic Malay speech audiometry, AWL

	Correct score (%)									
	-5	0	5	10	15	20	25	30	35	40
Mean	0	0.6	14.6	47.4	80.3	92.9	96.8	99.1	99.1	98.5
SD	0	2.4	17.3	23.2	12.	4.4	2.7	1.2	1.2	1.9
mean-2SD	0	0	0	1.	56.	84.1	91.4	96.7	96.8	94.6
mean+2SD	0	5.4	49.2	93.8	100	100	100	100	100	100

5.3.2.2 Meaningful words-only lists (MWL)

Similar analysis was done on the scores obtained from the normal hearing participant group using meaningful words-only list. The scores were not obtained from a repeat of speech audiometry, instead the scores were calculated by adding the scores of correct responses for the meaningful words in each tested list. This method was done to avoid any memory effect on the participants, which might happen if the lists were presented more than once within a short time.

For comparison purposes, participant L05 was included in the compilation of P-I function curves of the participants (Figure 5.5a). However, data from participant L05 was excluded from any other analyses due to its marked difference from the findings of other participants.

In general, the P-I function for the meaningful word lists (MWL) is similar to function seen in the all word lists. The P-I function using the MWL retained the S-shaped curve typical of speech audiograms, with three distinct stages – gradual rise at low intensities followed by a steep rise at higher levels and then plateau.

The data from the MWL set were also used to construct the 95% reference range for the P-I function (Figure 5.5b). The mean scores for the tested presentation levels together with the standard deviations (SDs) were used to calculate the upper limits (mean+2SD) and lower limits (mean-2 SD) for each presentation levels (Table 5.6). The general shape of curves are similar to the curves derived using AWL, with levels between 5 and 15 dB dial having the largest range of scores for each presentation level. All three curves have also reached plateau by 30 dB dial.

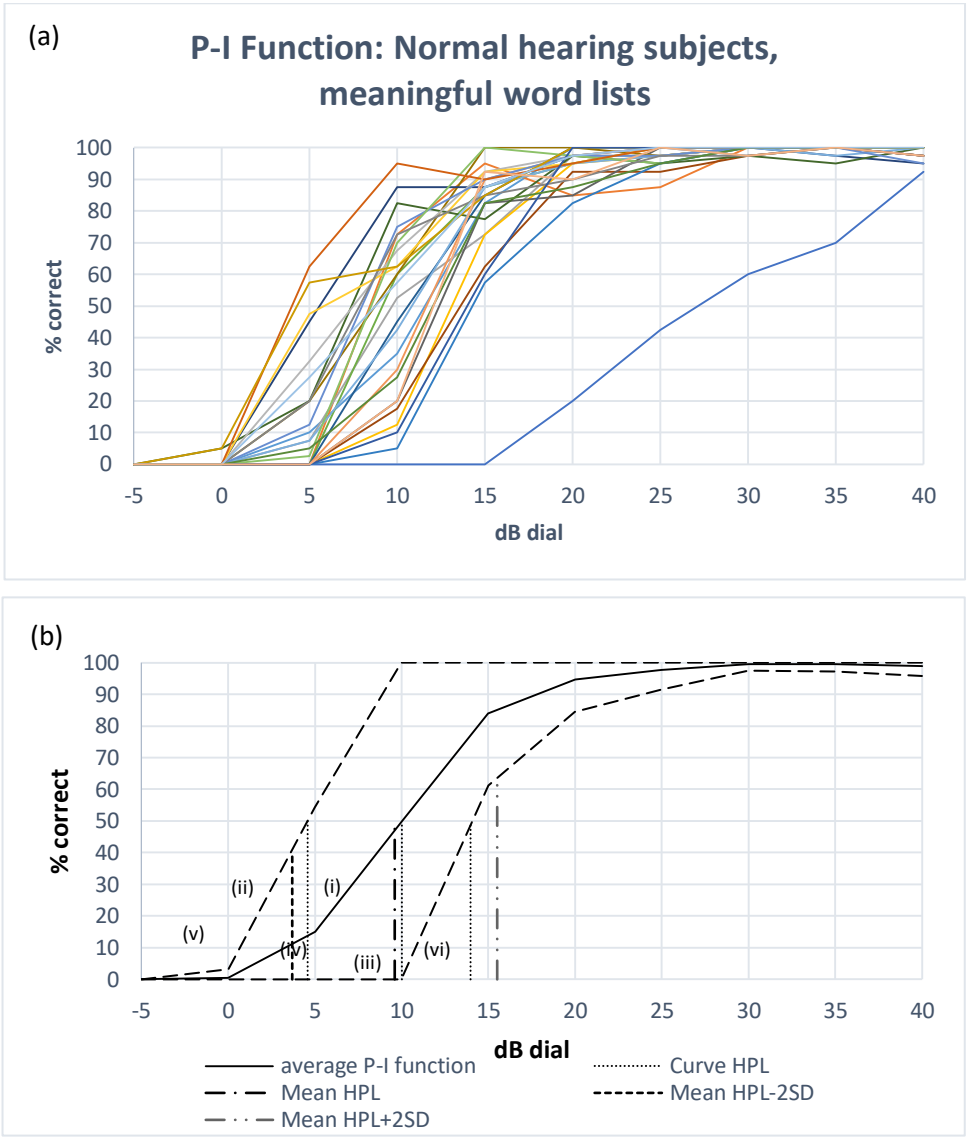


Figure 5.5 Performance/intensity (P-I) function for normal hearing participants – meaningful words-only lists: (a) shows the P-I function for all participants. The line rightmost (participant L05) is excluded from overall analysis (b) shows the mean P-I function curve (in solid line) and the upper and lower 2 standard deviations (in dashed lines). Six half peak levels were noted on the curves: (i)-(iii) are the HPL of the average P-I function curve and the HPLs of its lower and upper 2SDs ;(iv)-(vi) are the calculated mean of the participants' HPL and its lower and upper 2SDs

Table 5.6 Mean correct scores for normal hearing participants using bisyllabic Malay speech audiometry, MWL

	Correct score (%) / Presentation level (dB dial)									
	-5	**0**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
Mean	0	0.4	15.1	49.7	83.9	94.7	97.7	99.5	99.6	99
SD	0	1.4	19.6	26.2	11.3	5.2	3.1	1.0	1.2	1.6
mean-2SD	0	0	0	0	61.3	84.4	91.6	97.5	97.2	95.8
mean+2SD	0	3.2	54.3	100	100	100	100	100	100	100

A comparison between AWL and MWL findings in normal hearing participants showed that the P-I functions of both sets of stimuli are highly similar (Figure 5.6). There was a small difference at presentation levels between 10 and 20 dB dial, with MWL producing slightly higher percentage of correct scores compared to AWL. The slopes for the steepest part of the P-I function curve are 6.7% per dB for AWL and 6.9% per dB for MWL. The slightly steeper slope in MWL confirms the difference in the percentage of correct scores. This finding, although in line with the common findings that meaningful words produce better performance compared to nonsense words, did not agree with the extent of difference expected between AWL and MWL. The similarity of the P-I function constructed with AWL and MWL will be discussed further in the next section.

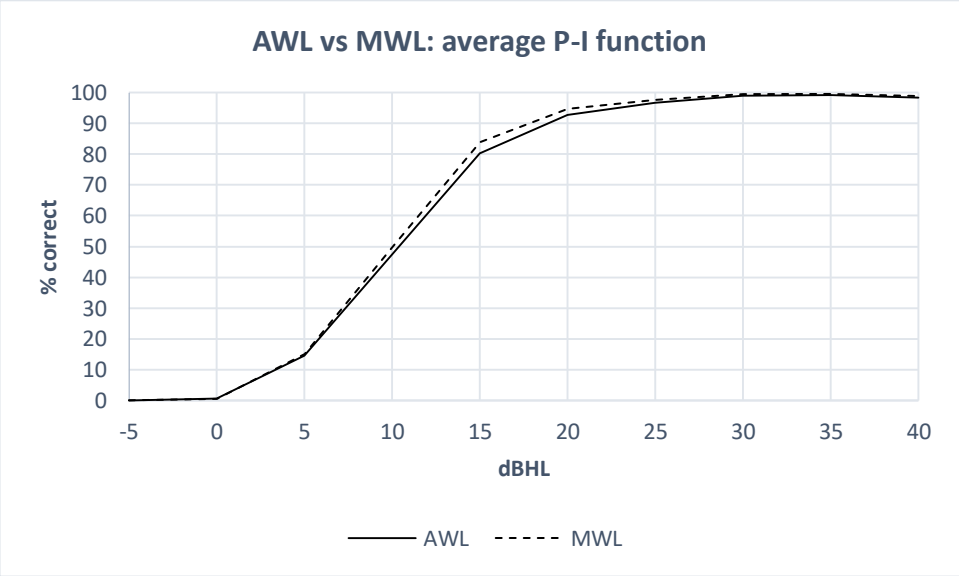


Figure 5.6 A comparison between average P-I function curve obtained using AWL and average P-I function curve obtained using MWL

5.3.3 Validity of bisyllabic Malay word lists on participants with sensorineural hearing loss

In addition to normal hearing participants, the study also looked at the P-I functions of participants with hearing loss. Two main types of hearing losses were included in the study, sensorineural hearing loss and conductive hearing loss.

Previous studies have shown that sensorineural hearing loss alters the speech audiometry P-I function (Wang et al, 2007; Boothroyd, 2008). Changes include elevated HPL, displacement of P-I function curve towards the higher presentation intensity level, lower maximum performance or correct scores and, in some cases, presence of rollover following the plateau. These changes reflect the characteristics of the loss, such as elevated hearing threshold, diminished frequency discrimination abilities and abnormal loudness growth, particularly loudness decay.

Speech audiometry using the bisyllabic Malay word lists were then performed on the test ear. Method of testing was as described in the research design. All of the participants completed the test in a single session, with each session taking an average of 30 minutes (including instructions). Most of the participants were able to complete the speech audiometry without any breaks in between lists.

A compilation of P-I function curves of the sensorineural hearing loss participants showed a variety of curve patterns for both AWL and MWL (Figure 5.7). Although the curves follow the general shape of speech audiogram (gradual start, steep rise and followed by plateau), most of the curves, in both AWL and MWL, had a very short gradual rise, or missing the part altogether. Several cases showed rollover, with the correct scores decreasing with increasing presentation level. It is interesting to note that, even for more severe hearing losses, most cases displayed highest scores or peaks at more than 90%, demonstrating that at suprathreshold levels, the participants were able to discriminate more than 90% of the presented phonemes. These characteristics were evident in both AWL and MWL.

Listeners with hearing loss, due to their elevated hearing thresholds, may experience decrease in dynamic range, that is the range between the threshold of hearing and the level at which the listeners experience discomfort. Reduced dynamic range means that the listener takes less change in intensity to advance from the level at which the sound is softest that he can hear to the level at which the sound is too loud to hear. In Figure 5.7, the reduced dynamic range may explain the short or missing gradual rise at the

lower presentation level, and the rapid rise to upper comfortable level is illustrated by the steep slope of the curve.

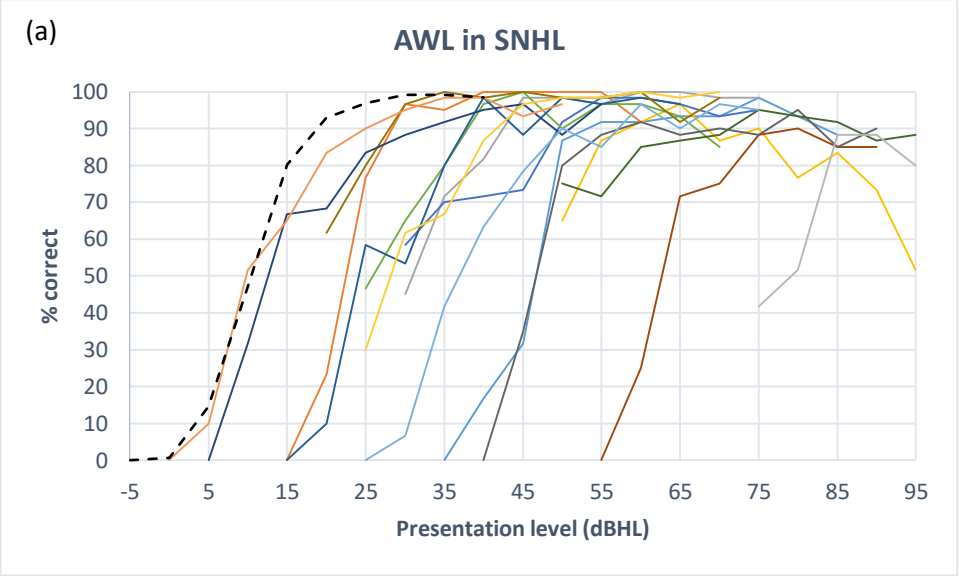
The rollover is also characteristic of the speech audiometry P-I function in sensorineural hearing loss cases, particularly those with retrocochlear damage. It is related to loudness decay, a phenomenon of which the hearing sensitivity lessens or 'decays' after a certain intensity level. In order to differentiate cochlear and retrocochlear damage, the calculation of rollover index is recommended (Mueller and Hall, 1996). A rollover index of 0.4 and above indicate retrocochlear hearing loss.

$$\text{Rollover index} = (\text{PBmax}-\text{PBmin})/\text{PBmin}$$

PBmax: Maximum speech recognition score or MSRS
PBmin: Minimum speech recognition score at an intensity level above the level of PBmax

As the current study did not distinguish cochlear and retrocochlear losses, it is unknown whether the index is applicable to the bisyllabic Malay word lists.

Maximum speech recognition score (MSRS) is another important characteristic that is observed in P-I function curves. It represents the participant's highest level of speech discrimination for that particular test material. In the current study, the range of maximum scores in the sensorineural hearing loss group is 88.3 to 100% with an average of 97.0% in AWL and between 92.5 to 100% (average 98.9%) in MWL. Sensorineural hearing losses are usually accompanied by diminished frequency discrimination abilities following damage to the hair cells and/or the auditory neural pathway. This abnormality is presented as decreased speech discrimination and lower-than-normal maximum speech recognition scores. In contrast to previous studies, the participants in the current study showed high levels of speech discrimination, including those with severe losses, both in AWL and MWL. The fact that the addition of nonsense words in AWL did not affect the maximum speech recognition scores considerably suggests that the perception of nonsense words that adopts Malay phonetics and phonology is similar to meaningful words. Linguistic cues provided through the Malay phonemes and phonetic rules may have contributed to the auditory discrimination process.



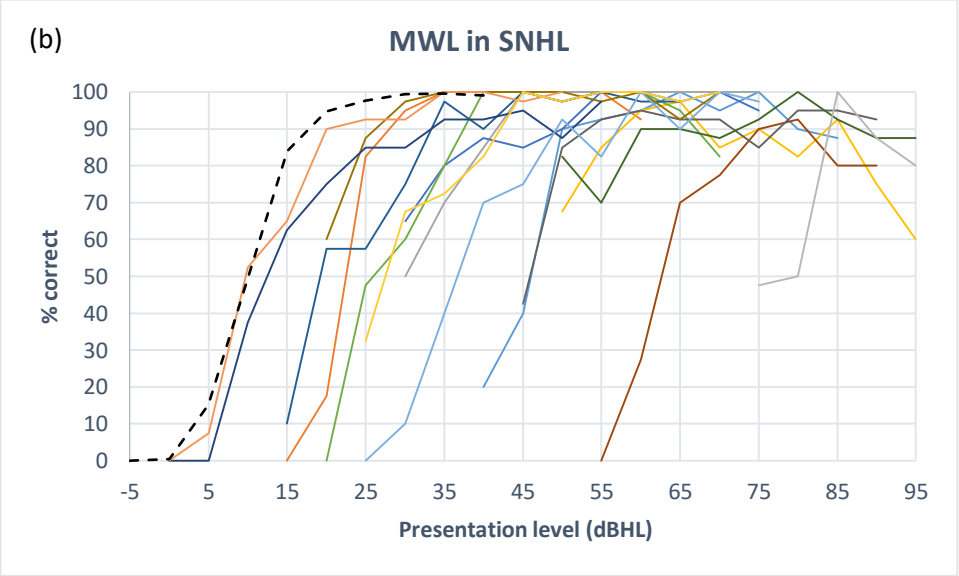


Figure 5.7 Performance-intensity functions of participants with SNHL using (a) AWL and (b) MWL. Dashed lines represent the average correct scores for normal hearing participants

A closer look at several cases with severe sensorineural hearing loss (SNHL) or worse showed a variety of hearing loss configurations. Three samples of cases are discussed below:

Case 1: HL4

Participant HL4 was a 26 year old male with unilateral steeply sloping hearing loss on the right ear (Figure 5.8). He reported that the hearing loss was sudden, following a motor-vehicle accident 6 months prior to the hearing test. He experienced otorrhea (bleeding in the ear) due to the accident; however, upon examination the ear drum was intact and tympanometry values were within normal limits. Participant HL4's case is of interest due to the suggestion of rollover in the P-I function (Figure 5.9). The MSRS for this participant was recorded at 65 dB dial, with 96.7% and 97.5% correct scores for AWL and MWL, respectively. Above 65 dB dial, the scores started to decrease gradually. The minimum percent correct scores (PBmin) at the level higher than the intensity for MSRS were 51.7% and 60% for AWL and MWL, respectively. These PBmin scores, crucial for the calculation of rollover index signifying cochlear hearing loss, were obtained at 95 dB dial, the highest presentation level (+40) set in the research design.

The rollover index was calculated following the formula described earlier. The rollover index for AWL was 0.465, well above the value recommended by Mueller and Hall (1996) for indication of retrocochlear disorder. The MWL, on the other hand, showed a rollover index of 0.385, slightly lower than the cut off value. Although there was a suggestion of retrocochlear disorder based on the P-I functions, the condition could not be confirmed without information from additional tests.

Case 2: Participant HL9

HL9 has a moderate SNHL on the left ear and profound SNHL on the right The average pure tone thresholds between 250 and 4000Hz on the left ear is 67 dB HL. Left ear was chosen as the test ear as it was the better ear (Figure 5.10). The configuration of the left ear hearing loss is relatively flat, with a slightly better threshold at 4000Hz. The speech audiometry P-I function showed a MSRS of 90% for AWL and 92% for MWL, both at 80 dB dial (Figure 5.11). Taking into consideration the Malay LTASS as described in section 4.3.4, the higher emphasis on lower and mid frequencies in Malay speech sounds as compared to English may have an effect on the speech perception of this participant.

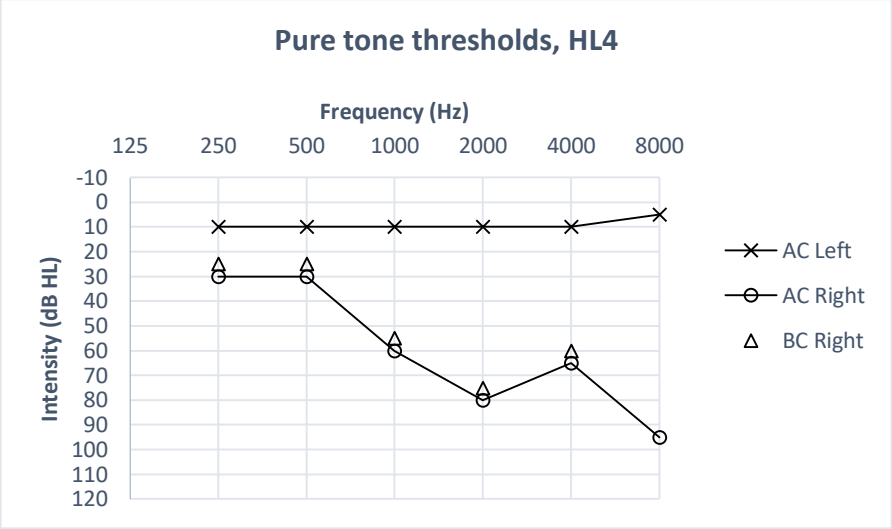


Figure 5.8 Pure tone audiogram of HL4

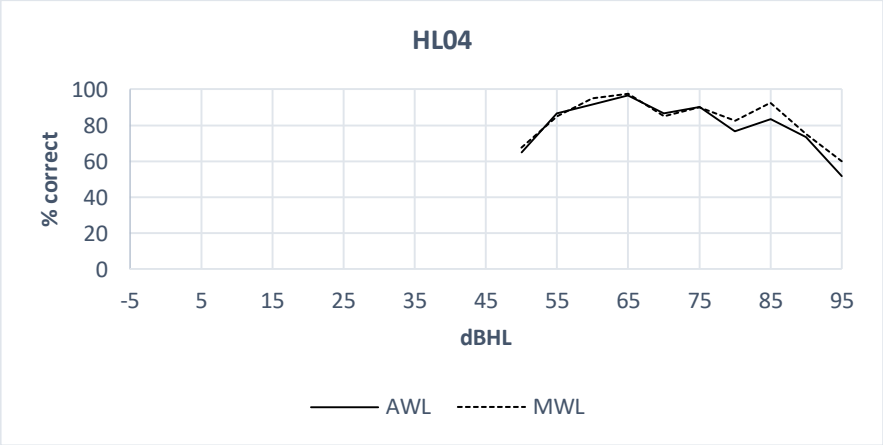


Figure 5.9 P-I function curve of HL4

The lower and mid frequencies also contain more energy, which makes the Malay speech sounds within those frequency regions ‘easier’ to hear compared to their English counterparts. In addition, the actual intensity level in dB HL of the speech stimuli was not known as the output intensity of the words was not measured. The output intensity is dependent on the input volume (also known as recording volume) of the speech stimuli, therefore if the input volume is high, the same occurs to the output intensity, and vice versa. In this case, there is a possibility of high input volume in the recording of the speech stimuli, affecting the output and therefore the level of sound that reaches the

listener's ear, which means 80 dB dial could be more intense (and therefore louder) than 80 dB HL.

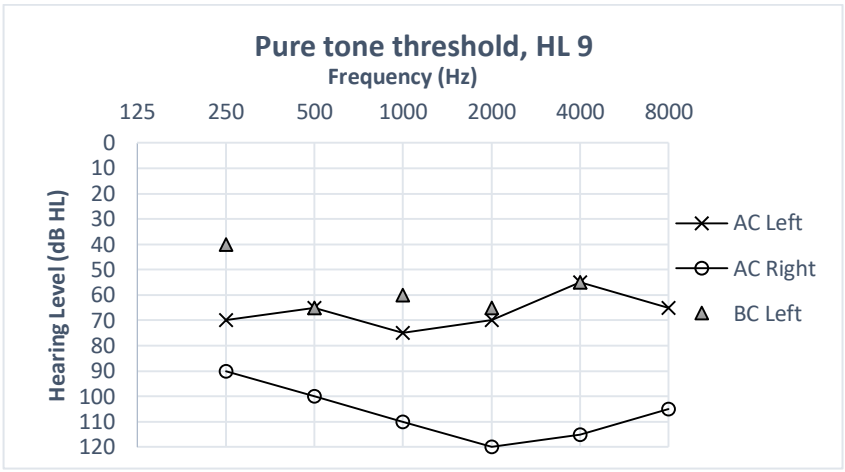


Figure 5.10 Pure tone audiogram of HL9

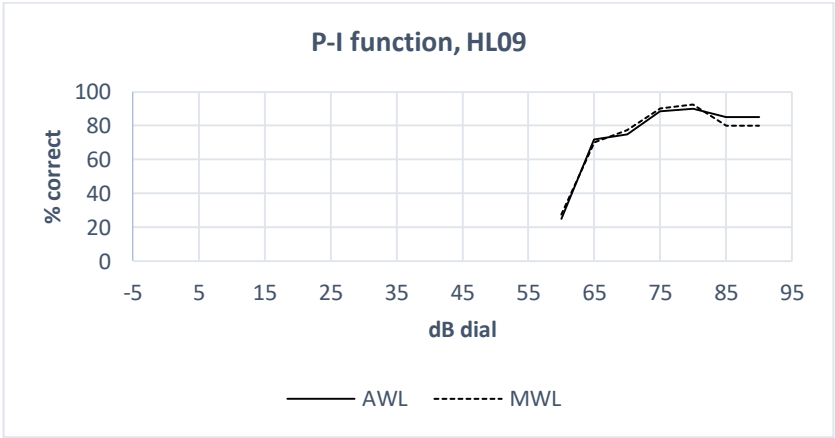


Figure 5.11 P-I function curve of HL9

Case 3: Participant HL15

Participant HL15 has left unilateral profound sensorineural hearing loss. He reported that the hearing loss was rapid and the onset was following a bout of vertigo circa 2003. The average pure tone threshold for frequencies 250-4000 Hz was 92 dB HL (Figure 5.12). Due to the difference in hearing levels between the two ears, masking was done for both pure tone audiometry and speech audiometry. Method of masking for the speech audiometry was as described by Yacullo (1999). The participant scored a maximum of 88.3% at 85 and 90 dB dial in AWL and 100% at 85 dB dial in MWL (Figure 5.13). There

are two possibilities that can be related to this occurrence. Apart from the input volume argument similar to Case 2, there is also a possibility that undermasking had occurred. The large gap between the test ear (left) and the non-test ear (right) might have allowed crossover of sound even after masking noise is presented to the non-test ear, especially at high stimulus intensity. Although care had been taken to ensure that the non-test ear was properly masked, undefined speech stimuli output intensity could be difficult to mask accurately. Although the use of insert phones are recommended for audiometry in patients with large interaural threshold difference, previous study has shown that there is a significant difference between the speech recognition scores obtained headphones and insert earphones at least at low presentation levels (Martin, Severence and Thibodeau, 1991). Based on this, the current study did not consider the use of insert earphones in exchange with headphones in the research design.

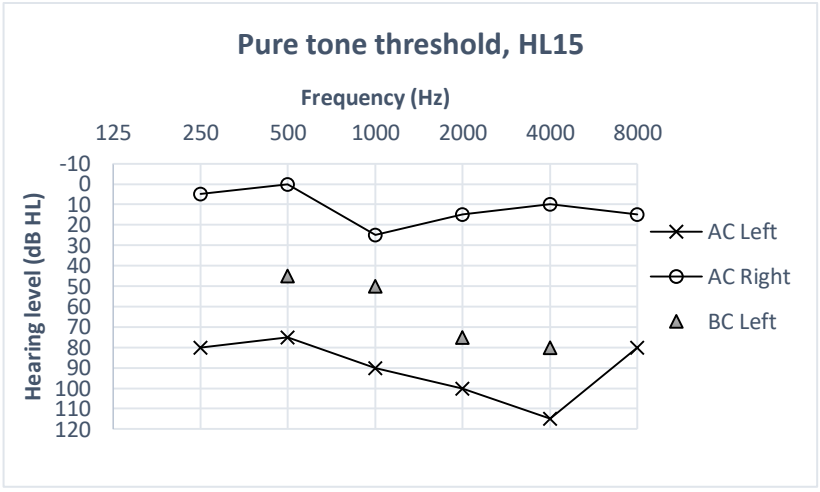


Figure 5.12 Pure tone audiometry of HL15 (masking was applied)

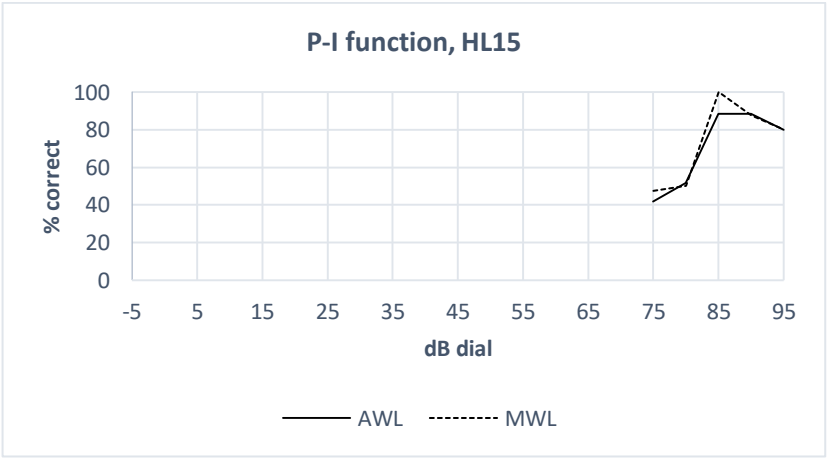


Figure 5.13 P-I function of HL15

5.3.4 Validity of bisyllabic Malay word lists in participants with conductive hearing loss

P-I function for both AWL and MWL in conductive hearing impaired participants showed a variety of curve patterns (Figure 5.14a and 5.14b). In general, the P-I function curves for participants in the CHL group were displaced towards the right on the x-axis compared to the P-I function curves of the normal hearing listeners. Two major types of patterns can be seen, curves with three distinct segments similar to the P-I function of the normal hearing participants, and curves with only two segments, a steep rise and a plateau. Most of the participants with mild and mild-moderate hearing losses showed three stages in their P-I function – initial gradual rise followed by steep rise and then plateau as the presentation level increases. Participants with moderate CHL and worse, on the other hand, showed two-stage curve development as the presentation level increases. The abbreviation of the normal P-I function curve in these cases may be contributed to the compression of the dynamic range of hearing, similar to the SNHL cases. Another aspect that can be considered is the limits of presentation levels in this study; the presentation levels were limited to a range of -5 dB to +40 dB relative to the listener's average hearing threshold. There is a possibility that the minimum level at which the listeners were able to hear any speech sound had not been reached. This can be seen on the P-I function curves for both AWL and MWL, in which several of the curves seem to be 'hanging' as the stimulus were not presented at a lower level in order to reach the lowest possible correct scores.

At higher presentation levels, the P-I function curves in the CHL participants showed the flat, plateau segment with no significant rollover observed. All of the participants showed a maximum speech recognition score of 98% and above in both AWL and MWL. These findings agree with the characteristics of CHL, which involves attenuation of sounds reaching the inner ear due to abnormalities in the conduction of sound without any damage to the frequency discrimination, as the cochlear hair cells in the inner ear are intact.

One participant showed an interesting P-I function with high scores, 68.3% and 67.5% for AWL and MWL respectively, at level of presentation of -5 dB below the 3-threshold average. This participant, HL14, had bilateral mild conductive hearing loss following a motor-vehicle accident 6 months prior to testing. The left ear showed type A

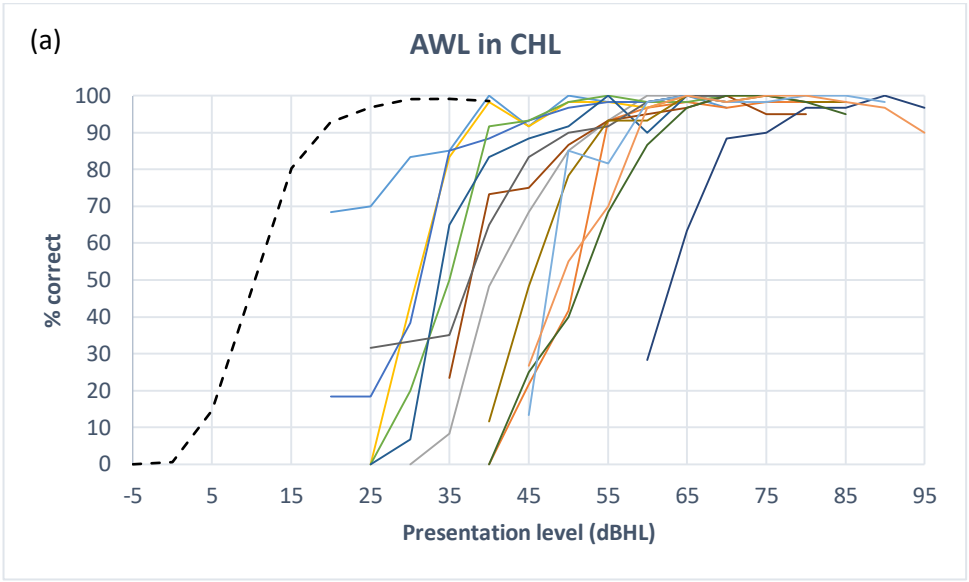
tympanogram indicating good middle ear function while the right ear showed type Ad tympanogram, suggesting a discontinuity of the ossicles. The high scores can be explained through two arguments; one, the presentation level was higher than expected. That is, the output level of the speech test items in dB dial, although set at 5 dB below the 3-threshold average, was more intense than the actual 3-threshold average minus 5 dB presented using pure tones ([(3-threshold average) - 5] dB HL). Another possibility is that HL15's speech sound perception could be better than what is reflected by the pure tone thresholds. Mild hearing losses, especially if acquired postlingually, are known to only minimally affect the listener's speech perception.

5.3.5 Correlation between SRT and pure tone averages in normal hearing, SNHL and CHL groups

One of the questions in this study is "How does the speech reception threshold (SRT) obtained using bisyllabic Malay word lists correlate with the pure tone hearing thresholds?" In this section of the study, the correlation between SRT and pure tone hearing thresholds (HTL) are established in order to find out whether the SRTs is reflective of the HTL. There are cases in clinical settings in which the SRT is used to validate the HTL. However, in this study, the HTL served as a benchmark on which the SRT is compared.

To study the correlation between the SRT and HTLs, correlation analysis was done on several combinations of HTL frequencies. The justification for using combinations of frequencies lies on the frequency content of the speech sounds; speech sounds are made of a range of frequencies. The speech frequency spectrum ranges approximately between 125 to 8000Hz (Mueller and Hall, 1996), therefore, it is justifiable that SRT would best be reflected by a combination of pure tone frequencies.

For the purpose of comparison between the SRT for the speech audiometry, several combinations of pure tone (PT) averages for normal hearing participants were made. A summary of the average combinations and their means is shown in Table 5.7.These pure tone average combinations would be used in the correlation analysis of the speech reception thresholds.



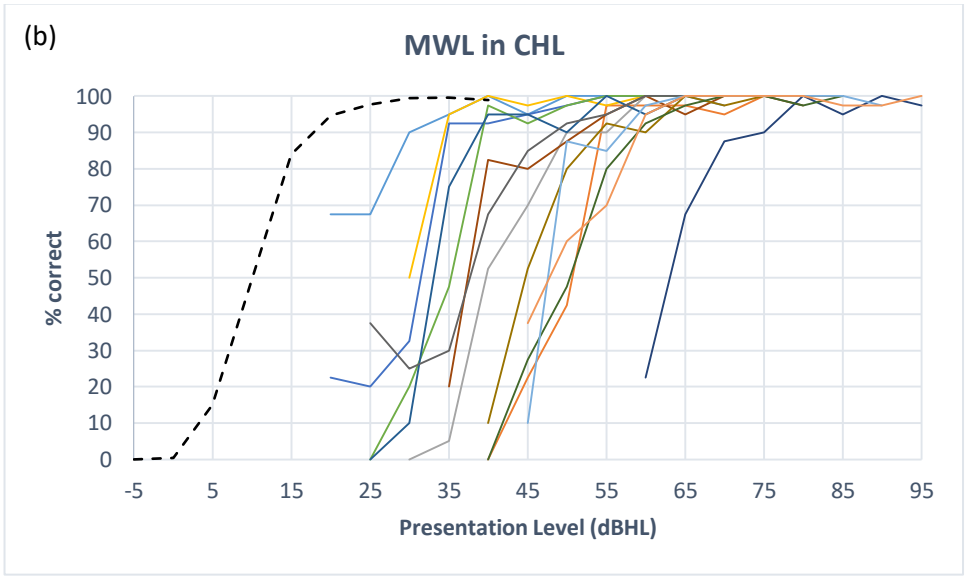


Figure 5.14 Performance-intensity functions of participants with CHL using (a) AWL and (b) MWL. Dashed lines represent the average correct scores for normal hearing participants.

Table 5.7 Summary of pure tone average combinations and their mean values

		Pure tone average combinations (kHz)						
		0.5, 1, 2	0.5, 1, 4	1, 2, 4	0.5, 1, 2, 4	0.25, 0.5, 1, 2	0.25, 0.5, 1, 2, 4	0.25, 0.5, 1, 2, 4, 8
Normal hearing	Mean (dB HL)	7	7	6	6	7	7	6
	SD	5	5	5	5	5	5	4
SNHL	Mean (dB HL)	39	42	45	42	38	41	42
	SD	21	19	22	20	20	20	19
CHL	Mean (dB HL)	41	40	39	40	41	40	41
	SD	14	14	14	14	11	12	12

Spearman Rank Correlation was selected as the statistical test based on the distribution of data and the strength of analysis. All tested combinations showed significant correlation between PTA HTLs and SRT. Strongest correlation can be seen in the SNHL group, both in all-words list (AWL) and meaningful words-only list (MWL), followed by the CHL group.

In the normal hearing group, strongest correlation between PTA and SRT can be seen in the 0.25, 0.5, 1 & 2 kHz and the 0.25, 0.5, 1, 2 & 4 kHz PTA combinations ($r=0.67$, $p<0.001$). Strongest correlation in MWL is seen in the 0.25, 0.5, 1, 2 & 4 kHz PTA combination ($r=0.65$, $p<0.001$). The highest correlation coefficients in SNHL group are seen in the 0.5, 1 & 2 Khz and the 0.25, 0.5, 1, 2 & 4 kHz combinations for both AWL and MWL ($r=0.95$, $p<0.001$). The 0.5, 1& 4 kHz combination showed the strongest correlation ($r=0.90$, $p<0.001$) in the CHL group using AWL. Using MWL, the CHL group showed strongest correlation between SRT and PTA average of 0.5,1, 2 & 4 kHz ($r=0.86$, $p<0.001$). Spearman's rho for all frequency combinations in all three groups using AWL and MWL are shown in Table 5.8 and Table 5.9 respectively.

It is important to observe that all of the frequency combinations showed significant correlation between SRT and HTL, and that the strongest correlation differs for different sets of hearing conditions and group of lists. In general, the correlation for AWL is stronger than MWL in participants from all three groups, although the difference is very small. The SNHL and CHL groups also demonstrated stronger monotonic correlations in both AWL and MWL compared to the normal hearing group. This difference in correlation can be explained through the range of hearing levels between the groups. The normal hearing group were represented by essentially a 20-dB range of hearing threshold levels (-5 to 15 dBHL), much smaller compared to the SNHL (110 dB range) and the CHL (70

dB range). The limited distribution, thus, the amount of variability, of the threshold levels of the normal hearing group could explain the lower correlation coefficients in the group as compared to the SNHL groups and CHL group (Goodwin and Leech, 2006). This does not necessarily mean that SRT and HTL among the normal hearing participants has weaker correlation as compared to the other two groups, it just that the larger variability in the hearing thresholds of the SNHL and CHL groups allow correlation between HTL and SRT to be better demonstrated.

5.3.6 Predictive analyses

Predictive analyses were done based on the findings of correlation analyses. The positive predictive value (PPV) and negative predictive value (NPV) calculations were used. These predictive values give insight on the probability of a test in giving correct diagnosis (Altman and Martin 1994); in this case, the probability of AWL and MWL in giving correct diagnosis of hearing impairment.

Based on the findings in section 5.3.5, the most commonly occurring PT average with the highest correlation is the 0.25, 0.5, 1, 2 & 4 kHz combination. The PT average showed the highest correlation for normal hearing and SNHL groups using MWL, and the SNHL group using AWL. It showed second highest correlation coefficient for normal hearing group using AWL and CHL group using MWL. Although it was the third strongest correlation in the CHL group using AWL, the correlation coefficient is still relatively high. This PT average combination does not agree with the commonly used pure tone average is speech audiometry, which is the combination of 500, 1000 and 2000 Hz. This finding may influence the calculations of initial presentation level and pure tone average-SRT agreement for AWL and MWL.

Scatterplots demonstrating the relationships between the HPL and the 0.25, 0.5, 1, 2 & 4 kHz pure tone average are displayed in Figure 5.15 (a)-(f). All the scatter plots showed positive gradients with linear association, indicating direct and linear correlations between the HPL and the pure tone average in all three groups, using either AWL or MWL. The data points for normal hearing participants showed moderate correlation in both AWL and MWL, which were reflected in the correlation coefficients in Table 5.8 and Table 5.9. Strong correlations are shown in both hearing loss groups using both AWL

and MWL, with data points tightly clustered along the gradient lines. The strong correlations were confirmed by the correlation coefficients seen in Tables 5.8 and 5.9.

Table 5.8 Non-parametric correlation of PT results vs SRT in AWL

Participant group	PT average	Spearman's rho		
		r_S	p	N
Normal hearing	0.5,1& 2 kHz	0.61	.001	25
	0.5,1& 4 kHz	0.62	.001	25
	1, 2 & 4 kHz	0.54	.006	25
	0.5,1, 2 & 4 kHz	0.60	.002	25
	0.25, 0.5, 1 & 2 kHz*	0.67	.000	25
	0.25, 0.5, 1, 2 & 4 kHz*	0.67	.000	25
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.49	.013	25
SNHL	0.5,1& 2 kHz*	0.95	.000	16
	0.5,1& 4 kHz	0.93	.000	16
	1, 2 & 4 kHz	0.81	.000	16
	0.5,1, 2 & 4 kHz	0.93	.000	16
	0.25, 0.5, 1 & 2 kHz	0.94	.000	16
	0.25, 0.5, 1, 2 & 4 kHz*	0.95	.000	16
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.90	.000	16
CHL	0.5,1& 2 kHz	0.85	.000	14
	0.5,1& 4 kHz*	0.90	.000	14
	1, 2 & 4 kHz	0.83	.000	14
	0.5,1, 2 & 4 kHz	0.87	.000	14
	0.25, 0.5, 1 & 2 kHz	0.84	.000	14
	0.25, 0.5, 1, 2 & 4 kHz	0.85	.000	14
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.84	.000	14

*Notes the pure tone combination with the strongest correlation

Table 5.9 Non-parametric correlation of PTA results vs SRT with MWL

Participant group	PTA	Spearman's rho		
		r_s	p	N
Normal hearing	0.5,1& 2 kHz	0.58	.002	25
	0.5,1& 4 kHz	0.61	.001	25
	1, 2 & 4 kHz	0.52	.008	25
	0.5,1, 2 & 4 kHz	0.58	.002	25
	0.25, 0.5, 1 & 2 kHz*	0.65	.000	25
	0.25, 0.5, 1, 2 & 4 kHz*	0.65	.000	25
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.48	.016	25
SNHL	0.5,1& 2 kHz*	0.95	.000	16
	0.5,1& 4 kHz	0.93	.000	16
	1, 2 & 4 kHz	0.81	.000	16
	0.5,1, 2 & 4 kHz	0.92	.000	16
	0.25, 0.5, 1 & 2 kHz	0.94	.000	16
	0.25, 0.5, 1, 2 & 4 kHz*	0.95	.000	16
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.90	.000	16
CHL	0.5,1& 2 kHz	0.84	.000	14
	0.5,1& 4 kHz	0.83	.000	14
	1, 2 & 4 kHz	0.80	.001	14
	0.5,1, 2 & 4 kHz*	0.86	.000	14
	0.25, 0.5, 1 & 2 kHz	0.83	.000	14
	0.25, 0.5, 1, 2 & 4 kHz	0.84	.000	14
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.81`	.000	14

*Notes the pure tone combination with the strongest correlation

The mean difference between HPL and PT average was analysed to study the distance between the HPL and PT average. For the normal hearing group, the 0.25, 0.5, 1, 2 & 4 kHz PT average is 4 dB and 3 dB lower than the HPL for AWL and MWL, respectively. Although the 0.25, 0.5, 1, 2 & 4 kHz PT average did not provide the lowest mean difference among the PT average combinations, it did yield the lowest standard deviation for both AWL and MWL measurements (Tables 5.10 and 5.11). The standard deviation was 3 for both AWL and MWL, giving the least distribution around the mean in comparison with other PT averages. This suggests that the speech reception thresholds,

calculated using HPL, are in good agreement with pure tone thresholds, calculated through the PT averages of 0.25, 0.5, 1, 2 & 4 kHz.

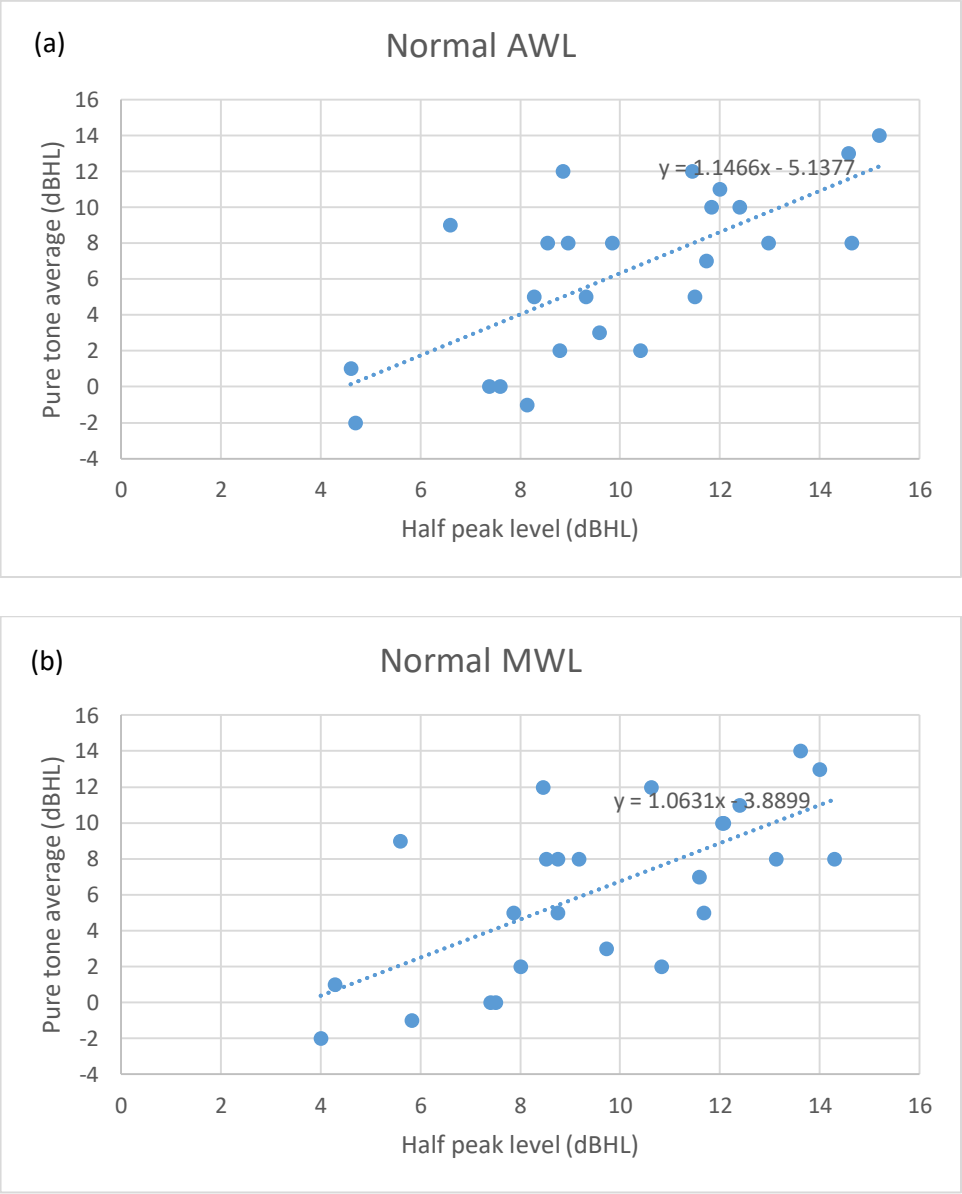
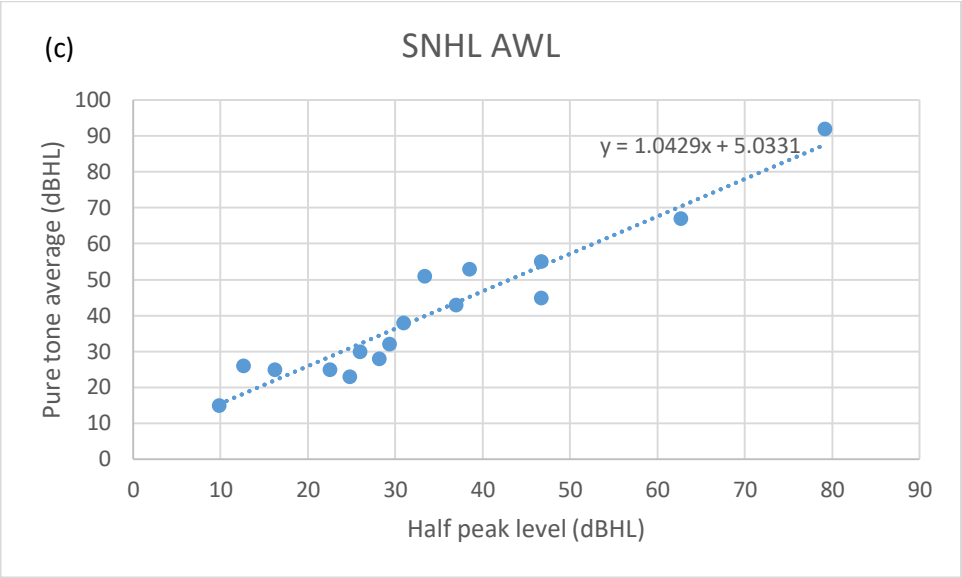


Figure 5.15 (a) & (b) Scatterplots displaying the relationship between the HPL and the 0.25, 0.5, 1, 2 & 4 kHz pure tone average for (a) normal hearing participants using AWL and (b) normal hearing participants using MWL



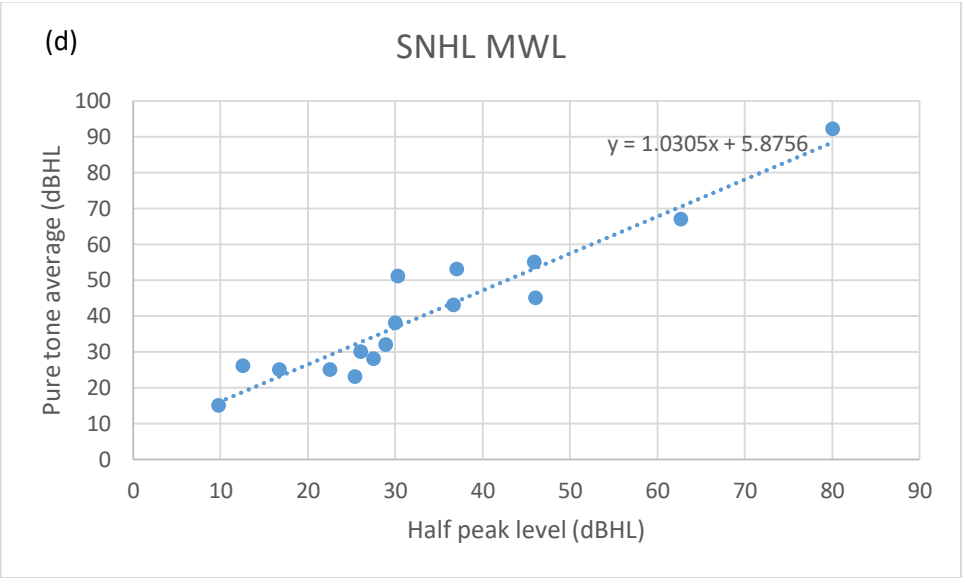
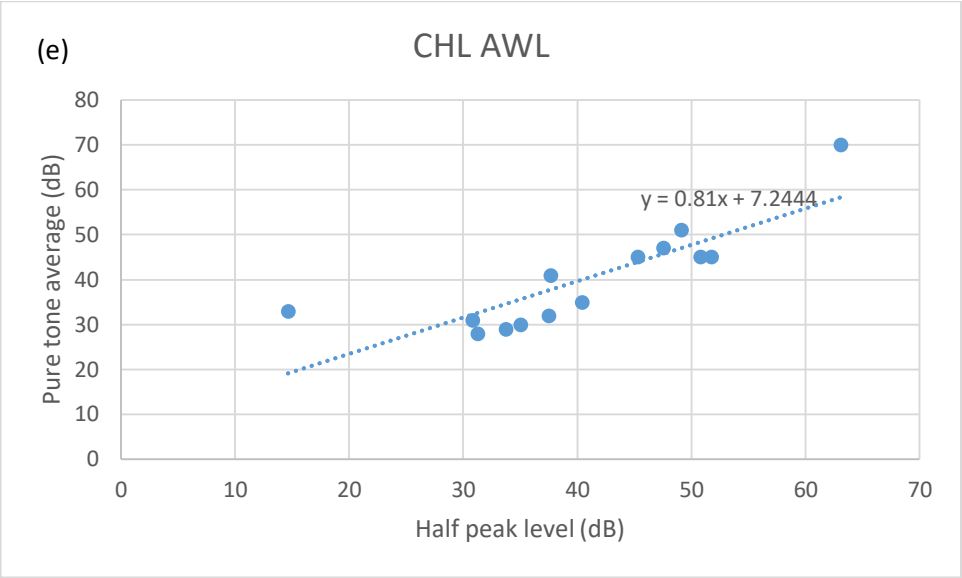


Figure 5.15 (c) & (d) Scatterplots displaying the relationship between the HPL and the 0.25, 0.5, 1, 2 & 4 kHz pure tone average for (c) participants with SNHL using AWL and (d) participants with SNHL using MWL



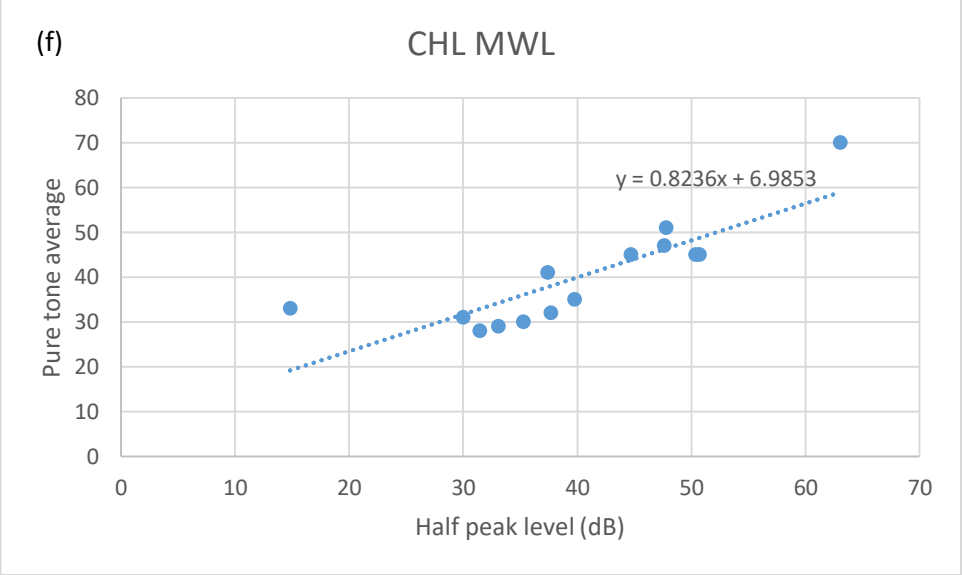


Figure 5.15 (e) & (f) Scatterplots displaying the relationship between the HPL and the 0.25, 0.5, 1, 2 & 4 kHz pure tone average for (e) participants with CHL using AWL and (f) participants with CHL using MWL

The SNHL group showed a larger mean difference between HPL and 0.25, 0.5, 1, 2 & 4 kHz PT average for both AWL and MWL, with the pure tone averages being higher than the HPL by 7 dB, as compared to the normal hearing group (Tables 5.12 and 5.13). The negative quantity implies that the PT average is, on average, higher than the HPL. The distribution around the mean was also larger for SNHL compared to the normal hearing group, as shown by the higher standard deviations for both AWL and MWL. However,

the standard deviations for both AWL and MWL calculated using 0.25, 0.5, 1, 2 & 4 kHz PT average are relatively lower than the standard deviations of other PT averages. This means that the agreement between HPL and pure tone average in SNHL is also best calculated using 0.25, 0.5, 1, 2 & 4 kHz PT average.

In contrast to the normal hearing and SNHL groups, the CHL group showed the best agreement between HPL and 0.25, 0.5, 1, 2 & 4 kHz PT average, with mean difference of 1 dB and 0 dB for AWL and MWL respectively (Tables 5.14 and 5.15). Similar to the normal hearing group, the HPL is higher than the PT average in terms of dB level. The standard deviations, however, are both the second lowest among the PT averages, and larger compared to the normal hearing and SNHL groups. Again, this indicates good agreement between the speech reception thresholds and the pure tone thresholds in CHL.

Given the variation in the HPL-PT average difference, a reference range is also suggested. The range of two standard deviations should provide 95% confidence interval. Based on the mean HPL-PT average difference in Tables 5.10 and 5.11, the accuracy with 95% confidence limits for the agreement between HPL-PT averages are ±7 dB for both AWL and for MWL. For slightly less strict allowance for accuracy, a range of ±10 dB is suggested as it is more applicable in clinical situations.

An alternative method of measurement for reliability of the speech reception threshold giving the 95% confidence limit described by Boothroyd (1968), was carried out as a comparison. The standard deviation for correct scores at HPL for the normally hearing participants using AWL is 23%, giving a 95% confidence limit (± 2SD) for the correct scores at HPL at ±46%. Based on the gradient of the steepest part of the normal P-I function curve using AWL shown in section 5.3.2.2, which was 6.6%/dB, an error of 46% represents an error of 7dB at HPL. The slope was calculated over 3 presentation levels, and Boothroyd (ibid.) advised a reduction of the error in threshold by the factor of 1.73 (i.e. $\sqrt{3}$). Therefore, the accuracy of HPL with 95% confidence limit is calculated at ±4dB. A similar calculation for normally hearing participants using MWL gives a 95% confidence limit of ±4 dB. In a clinical setting, it is suggested that both measurements are rounded to ±5 dB. This confidence limit is similar to the pure tone audiometry threshold error allowance, but much stricter than the confidence limit calculated using the standard deviation values. This validates the use and the accuracy of AWL and MWL in measuring speech reception threshold.

Table 5.10 HPL-to-pure tone average differences for tested combinations of pure tone average in normal hearing participants using AWL

		HPL – PT average difference (dB)						
ID	**HPL**	**0.5, 1, 2 kHz**	**1,2, 4 kHz**	**0.5, 1,2, 4 kHz**	**0.25, 0.5, 1, 2 kHz**	**0.25, 0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2, 4, 8 kHz**	**0.5, 1, 4 kHz**
L01	12	7	3	4	7	5	5	3
L02	9	-5	-3	-4	-4	-3	-4	-3
L03	10	6	8	7	6	7	6	8
L04	13	5	6	6	4	5	6	6
L06	10	0	3	2	0	2	3	3
L07	12	-2	2	0	-2	-1	1	0
L08	15	0	0	0	1	1	3	0
L09	12	2	4	2	1	1	3	0
L10	9	3	4	4	3	4	4	6
L11	7	-5	-3	-3	-3	-2	-4	-2
P03	8	8	9	8	8	8	8	6
P04	10	9	9	8	9	8	9	7
P05	12	5	8	7	5	7	7	7
P06	7	6	7	7	6	7	7	11
P07	8	8	12	9	8	9	7	10
P08	9	7	11	9	5	7	8	7
P09	9	0	0	0	1	1	-1	-2
P10	15	6	8	7	6	7	6	6
P11	5	6	5	6	7	7	6	6
P12	8	5	2	2	6	3	-2	-0
P13	5	1	5	3	2	4	2	3
P14	15	1	3	2	1	2	3	1
P15	12	2	5	3	1	2	4	2
P16	9	1	1	0	2	1	1	-1
P17	12	-1	2	1	1	2	2	2
Mean	10	3	4	4	3	4	4	4
SD	3	4	4	4	4	3	4	4

Due to the presence of difference between the HPL and the PT average in normal hearing group, there is a question of whether correction factor should be added in testing the agreement between the HPL and the PT average. If so, a correction factor of 4 dB is suggested to be applied to the HPL to verify its agreement with the PT average when AWL is used. For MWL, a correction factor of 3 dB is suggested.

Table 5.11 HPL-to-pure tone average differences for tested combinations of pure tone average in normal hearing participants using MWL

		HPL – PT average difference (dB)						
ID	**HPL**	**0.5, 1, 2 kHz**	**1, 2, 4 kHz**	**0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2 kHz**	**0.25, 0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2, 4, 8 kHz**	**0.5, 1, 4 kHz**
L01	12	7	3	4	7	5	5	3
L02	9	-5	-3	-4	-4	-4	-4	-3
L03	10	6	8	7	6	7	6	8
L04	13	5	7	6	4	5	6	7
L06	9	-1	3	2	-1	1	3	3
L07	11	-3	1	-1	-3	-1	0	-1
L08	14	-1	-1	-1	0	0	1	-1
L09	12	2	4	2	1	1	3	1
L10	9	2	4	4	3	4	4	5
L11	6	-6	-4	-4	-4	-3	-5	-3
P03		7	9	7	7	7	8	6
P04	11	9	9	8	10	9	9	8
P05	12	5	8	7	5	7	7	7
P06	8	6	8	8	6	8	8	11
P07	6	6	9	7	6	7	5	8
P08	8	6	10	8	4	6	7	6
P09	9	0	0	0	1	1	-1	-2
P10	14	6	8	7	6	6	6	6
P11	4	6	4	5	7	6	6	6
P12	8	5	1	2	5	3	-2	-1
P13	4	1	4	3	2	3	2	3
P14	14	1	2	2	0	1	2	1
P15	1	2	5	3	1	2	4	2
P16	9	0	0	0	1	1	0	-1
P17	12	-1	2	1	1	2	2	2
Mean	10	3	4	3.	3	3	3	3
SD	3	4	4	4	4	4	4	4

Therefore, the suggested formula in investigating the agreement between HPL and PT average is as follows:

HPL = [(0.25, 0.5, 1, 2 & 4 kHz PT average) + CF] ± 7 dB

with

CF = correction factor; 4 dB for AWL and 3 dB for MWL

The sensitivity, specificity and positive predictive values of the bisyllabic Malay word lists were calculated based on the HPL-PT average difference of participants with hearing loss. Sensitivity provides the probability of being tested as having hearing loss when hearing loss is present, and therefore, signifies the ability of the word lists to correctly identify those with hearing loss (Parikh et al., 2008). It is calculated as:

Sensitivity = true positive/(true positive + false negative)

On the other hand, positive predictive value (PPV) and negative predictive value (NPV) give the proportion of participants showing positive test results, in this case HPL out of the normal range, who actually have hearing loss, and the proportion of participants showing negative results who actually do not have hearing loss, respectively (Parikh, ibid.). PPV and NPV are calculated as:

PPV = true positive/(true positive + false positive)

NPV = true negative/(true negative + false negative)

Predictive analyses, consisting of sensitivity, specificity, PPV and NPV were constructed based on the mean HPL-PT average difference of the hearing loss groups and the suggested confidence interval (CI). Three correction factor conditions were applied to the HPL-PT average agreement:

- Correction factor of +4 for AWL and +3 for MWL applied to normal hearing, SNHL and CHL groups
- No correction factor applied to any group
- Individual correction factors
    - AWL: Normal hearing and CHL groups, CF= +4; SNHL group, CF= -7
    - MWL: Normal hearing and CHL groups, CF= +3; SNHL group, CF= -7

Table 5.12 HPL-to-pure tone average differences for tested combinations of pure tone average in SNHL participants using AWL

ID	HPL	HPL – PT average difference (dB)						
		0.5, 1, 2 kHz	1, 2, 4 kHz	0.5, 1, 2, 4 kHz	0.25, 0.5, 1, 2 kHz	0.25, 0.5, 1, 2, 4 kHz	0.25, 0.5, 1, 2, 4, 8 kHz	0.5, 1, 4 kHz
HL01	47	2	12	5	-2	2	3	0
HL02	23	3	-11	-6	5	-3	-4	-9
HL03	31	-2	-14	-10	0	-7	-10	-12
HL04	39	-18	-30	-20	-12	-15	-22	-13
HL05	37	2	-8	-6	-1	-6	-13	-5
HL06	29	-1	8	1	-5	-3	-1	-4
HL07	26	-6	-7	-7	-3	-4	-4	-6
HL08	25	0	0	1	1	2	1	3
HL09	63	-7	-4	-4	-7	-4	-4	-2
HL10	10	-2	-10	-8	0	-5	-9	-9
HL13	47	0	-20	-11	0	-8	-8	-7
HL15	79	-9	-23	-16	-7	-13	-11	-13
HL19	16	-11	-9	-9	-10	-9	-6	-9
HL23	13	-4	-22	-15	-5	-13	-21	-12
HL24	28	0	-7	-3	3	0	-4	-2
HL26	33	-23	-25	-23	-17	-18	-14	-20
Mean	34	-5	-11	-8	-4	-7	-8	-8
SD	18	7	12	8	6	6	7	6

In general, the sensitivity, specificity and predictive values were higher with the application of less stringent accuracy limits of ±10 dB compared to ±7 dB. Better specificity, sensitivity and predictive values are found across the three correction factor conditions, with both AWL and MWL. This was expected as, with a wider confidence interval, more participants are included in the 'true positive' and 'true negative' groups. However, in this case, better sensitivity and specificity might result in less accurate representation of hearing loss through speech audiometry.

It is interesting to note, but not surprisingly, that the best sensitivity, specificity and predictive values are found with individual correction factors applied to the normal hearing, CHL and SNHL groups (Table 5.19). The sensitivity and NPV for the ±10 dB accuracy of more than 85%, and perfect specificity and PPV makes this test clinically very robust. MWL results produced excellent specificity (93%)and NPV (93%) than

AWL. The separate calculation for the SNHL, based on the mean HPL-PT average difference, contributes to a more precise prediction of HPL-PT average agreement.

Table 5.13 HPL-to-pure tone average differences for tested combinations of pure tone average in SNHL participants using MWL

		HPL – PT average difference (dB)						
ID	**HPL**	**0.5, 1, 2 kHz**	**1, 2, 4 kHz**	**0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2 kHz**	**0.25, 0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2, 4, 8 kHz**	**0.5, 1, 4 kHz**
HL01	46	1	11	5	-3	1	3	-1
HL02	23	3	-11	-6	5	-3	-4	-9
HL03	30	-3	-15	-11	-1	-8	-11	-13
HL04	37	-20	-31	-22	-13	-16	-23	-15
HL05	37	2	-8	-6	-1	-6	-13	-5
HL06	29	-1	7	0	-5	-3	-1	-5
HL07	26	-6	-7	-7	-3	-4	-4	-6
HL08	25	0	0	2	2	2	1	4
HL09	63	-7	-4	-4	-7	-4	-4	-2
HL10	10	-2	-10	-8	0	-5	-9	-9
HL13	46	-1	-21	-12	0	-9	-9	-8
HL15	80	-8	-22	-15	-6	-12	-10	-12
HL19	17	-10	-8	-8	-10	-8	-6	-8
HL23	13	-4	-23	-15	-5	-14	-21	-13
HL24	28	-1	-8	-4	3	-1	-4	-3
HL26	30	-26	-28	-26	-20	-21	-17	-23
Mean	34	-5	-11	-9	-4	-7	-8	-8
SD	18	8	12	8	6	6	7	6

Table 5.14 HPL-to-pure tone average differences for tested combinations of pure tone average in CHL participants using AWL

ID	HPL	HPL – PT average difference (dB)						
		0.5, 1, 2 kHz	1, 2, 4 kHz	0.5, 1, 2, 4 kHz	0.25, 0.5, 1, 2 kHz	0.25, 0.5, 1, 2, 4 kHz	0.25, 0.5, 1, 2, 4, 8 kHz	0.5, 1, 4 kHz
HL 14	15	-19	-24	-20	-17	-18	-20	-19
HL 17	40	5	5	5	5	5	0	-6
HL 18	31	-1	4	3	-4	0	3	-4
HL 20	31	8	15	9	1	3	3	8
HL 21	35	5	7	6	4	5	5	10
HL 22	34	2	10	6	1	5	4	5
HL 27	38	6	4	5	6	6	3	8
HL 16	51	4	11	7	3	6	8	21
HL 25	38	-1	-2	-2	-2	-3	-6	-2
HL 28	45	0	5	3	-2	0	-3	12
HL 30	52	7	3	4	9	7	5	7
HL 31	48	-2	-2	-2	1	1	-1	-26
HL 32	49	-4	1	-3	-2	-2	-6	3
HL 29	63	-15	-12	-12	-8	-7	-7	13
Mean	41	0	2	1	0	1	-1	2
SD	12	8	10	8	7	7	7	13

5.4 Discussion

This part of the study tried to answer the research questions “How would adult Malay speakers, both normal hearing and hearing impaired, perform using word lists that contain both meaningful and nonsense words?” and “Would speech audiometry material consisting of meaningful and nonsense words be able to reflect the speech hearing and discrimination abilities of its listener?” The aim of this part of the study was to clinically validate the bisyllabic Malay word lists in two main aspects; whether the word lists are able to reflect the different types of hearing conditions, and whether the word lists are able to reflect the hearing level. The following sub-sections discuss the findings presented in the results section

5.4.1 Clinical validity testing

Clinical validity of the bisyllabic Malay word lists is a critical element in the development process. It ensures that the word lists were fit to be used in a clinical setting, able to distinguish different patterns of hearing configuration and able to differentiate types hearing loss. A highly valid test gives the user confidence that any negative results do mean that the probability of the absence of the tested condition is high and any positive results mean that the condition has high probability to be present.

Table 5.15 HPL-to-pure tone average differences for tested combinations of pure tone average in CHL participants using MWL

		HPL – PT average difference (dB)						
ID	**HPL**	**0.5, 1, 2 kHz**	**1, 2, 4 kHz**	**0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2 kHz**	**0.25, 0.5, 1, 2, 4 kHz**	**0.25, 0.5, 1, 2, 4, 8 kHz**	**0.5, 1, 4 kHz**
HL 14	15	-19	-24	-20	-16	-18	-20	-19
HL 16	51	4	11	7	3	6	8	4
HL 17	40	5	5	5	5	5	0	5
HL 18	30	-2	3	3	-5	-1	3	7
HL 20	32	8	15	9	2	4	3	7
HL 21	35	5	7	7	4	5	5	7
HL 22	33	1	10	6	1	4	3	3
HL 25	37	-1	-3	-3	-3	-4	-6	7
HL 27	38	6	4	5	6	6	4	-2
HL 28	45	0	5	2	-3	0	-4	11
HL 29	63	-15	-12	-12	-8	-7	-7	18
HL 30	50	5	2	3	8	5	4	-23
HL 31	48	-2	-2	-2	1	1	-1	1
HL 32	48	-6	-1	-5	-4	-3	-7	-2
Mean	40	-1	2	0	-1	0	-1	2
SD	12	8	10	8	6	7	7	11

Table 5.16 A summary of means and two standard deviations for normal hearing group

	HPL-PT average difference	
	AWL	**MWL**
Mean	4	3
SD	3	4
2SD	7	7

To increase the validity of the developed bisyllabic Malay word lists in terms of detecting presence of hearing loss as well as eliminating the possibility having false positive results in normal hearing listeners, the study was designed to assess three groups of participants – normal hearing participants, participants with sensorineural hearing loss and participants with conductive hearing loss. Each of these three groups not only present different levels of hearing ability but also different characteristics of hearing acuity, such as dynamic range (the range between the threshold of hearing to the maximum sound level that is comfortable to hear) and frequency discrimination ability. These differences would also translate into different patterns of P-I function produced by speech stimuli.

Table 5.17 Predictive values for HPL-PT average agreement with two accuracy limits and applied correction factor

		Correction factor (CF) applied	
		AWL (CF=4 across all groups)	**MWL (CF=3 across all groups)**
CI ±7	sensitivity	0.53	0.60
	specificity	0.96	1.00
	PPV	0.94	1
	NPV	0.63	0.68
CI ±10	sensitivity	0.67	0.73
	specificity	1.00	1.00
	PPV	1	1
	NPV	0.71	0.76

Table 5.18 Predictive values for HPL-PT average agreement with two accuracy limits and no correction factor

		No CF applied	
		AWL	**MWL**
CI ±7	sensitivity	0.73	0.83
	specificity	0.84	0.88
	PPV	0.85	0.89
	NPV	0.72	0.81
CI ±10	sensitivity	0.83	0.83
	specificity	1.00	1.00
	PPV	1.00	1.00
	NPV	0.83	0.83

Table 5.19 Predictive values for HPL-PT average agreement with two accuracy limits and individual correction factors for SNHL and normal hearing/CHL

		Individual CF applied	
		AWL (CF= 4 for normal hearing and CHL, CF=-7 for SNHL)	**MWL (CF= 3 for normal hearing and CHL, CF=-7 for SNHL)**
CI ±7	sensitivity	0.73	0.80
	specificity	0.96	1.00
	PPV	0.96	1.00
	NPV	0.75	0.81
CI ±10	sensitivity	0.90	0.93
	specificity	1.00	1.00
	PPV	1.00	1.00
	NPV	0.89	0.93

This study aims to establish the characteristics of P-I function produced by Malay bisyllabic words using two sets of word combinations, a mix of meaningful and nonsense words and a set of meaningful words only. Earlier studies have established the characteristics of P-I function associated with different types of hearing loss, for example, reduced maximum speech recognition score (MSRS) in sensorineural loss, the presence of roll-over in retrocochlear hearing loss and elevation of threshold without decreased MSRS in cases of conductive hearing loss. Earlier in the introduction chapter, it was shown that Malay has a unique phonology and phonetic system. The differences in phoneme content are believed to influence the Malay LTASS, as shown in the Chapter 4. There is a considerable difference in the acoustic content, especially at higher frequencies, between Malay LTASS and Universal LTASS as well as Cox's and Moore's (1988) English LTASS. It was not known whether these differences, as well as having a mix of meaningful and nonsense words in a list, affect the patterns of P-I function. There is a possibility that the speech cues in Malay are more dependent on low frequency signals rather than high frequency signals, unlike the speech cues for English.

The method used for the validity testing is by looking at the correlation between the speech reception thresholds and the pure tone hearing threshold average, which is selected as the standard in the current study. Again, due to the differences in the acoustic content of Malay language as seen in the LTASS, the commonly used combination of

500, 1000 and 2000 Hz pure tone threshold average might not be suitable to be used as a basis of estimation of the speech audiometry readings. Therefore, several combinations of pure tone threshold averages were compared to the speech reception thresholds in order to find the best fit correlation. Due to the large differences at higher frequencies between Malay and the universal/English LTASS, it was interesting to know whether pure tone thresholds at high frequencies affect the correlation between the speech reception thresholds and the threshold averages.

All participant groups went through the same research process. The invitation to participate in the study was based on self-assessment of normal hearing for the normal hearing participant group. The participants with hearing impairment, on the other hand, were invited based on their previous hearing assessments. The following preliminary assessments – health history, otoscopy and tympanometry, as well as pure tone audiometry were done in the same manner for all of the participants in order to minimise bias. Slight variation was applied to the speech audiometry procedure; normal hearing participants were presented with several pre-set levels of presentations. In contrast, the hearing impaired groups, due to the individual variations of hearing loss type and degree, were presented with speech stimuli in accordance to their levels of hearing loss.

The following sections discuss the findings of each of the three groups.

5.4.2 Construct validity through clinical validity testing in normal hearing participants

Two measurements were made for the normal hearing participants group, the ‘all-word lists’ (AWL) which employed all 15 words (10 meaningful and 5 nonsense) contained in each list and the ‘meaningful-word only lists’ (MWL) which utilised only the meaningful words in each list. The reason for having these two measurements is to study the similarities and differences in the findings between the commonly-used meaningful word lists (MWL) and the novel mixed words lists (AWL). To reduce subject bias, it was decided that the measurements of both AWL and MWL were to be taken from the same participants. As discussed in the review of methods, obtaining data for both measurements from the same individual participants should eliminate the need for subject matching and reduce the error of natural variability.

The findings show that both AWL and MWL produced P-I functions consistent with the P-I functions produced in other speech audiometry material (Ashoor and Prochazka, 1982; Mukari and Said, 1991, Nissen et al., 2007; Nissen et al., 2011). The upper limits of speech reception thresholds, calculated based on the HPL, were less than 20 dB HL in both AWL and MWL, signifying that the word lists are able to identify normal hearing. The MWL shows better representation of normal hearing, with the upper limit of HPL less than 15 dB HL. The confidence interval of ±7 calculated based on the standard deviation is comparable to the suggested value for good HPL-PT average agreement (American Speech-Language-Hearing Association, 2016).

Pure tone average is a commonly used measurement in audiology, particularly in speech audiometry. It has been used as a basis of determining presentation levels of speech stimuli as well as in the validation of speech audiometry, although in clinical setting the pure tone averages are instead validated by the speech audiometry thresholds (Boothroyd, 2008). There are no consensus in the literature for which frequency or combination of frequencies used as a comparison to the speech audiometry threshold; Boothroyd (1968) tried using the HTL at 1000 Hz and found that the correlation between the HTL and the speech audiometry threshold were lacking for hearing loss configurations other than flat hearing loss. . Commonly used average is the 3-frequency average, that is, the mean threshold of 500, 1000 and 2000 Hz (Lau and So, 1988; Hazan, Fourcin and Abbelton, 1991; Nissen et al., 2007; Wang et al., 2007; Boothroyd, ibid.; Han et al., 2009; Nissen et al, 2011). In the current study, the validity of the speech audiometry material is measured through its correlation with the pure tone average. Several averages were tested for correlation to find the average that best fit the speech reception threshold (SRT), the level at which 50% recognition probability for the words is obtained.

In order to measure the correlation, pure tone audiometry was carried out on each participant. This was also used to determine whether or not the participants meet the inclusion criteria for having normal hearing. All 26 recruited participants were found to have hearing thresholds of ≤15 dB HL across the frequencies on the tested ear. The average thresholds with various frequency combinations were also calculated. The mean HTLs ranged from 4 to 9 dB HL for frequencies between 0.25 to 8 kHz. The mean 3-frequency average for 500, 1000 and 2000 in previous studies varied greatly, with the current study showing 3-frequency average at 7 dB HL. A summary of comparison

between the current finding and the findings from previous studies are given in Table 5.20.

Table 5.20 Summary of 3-frequency (500, 1000 and 2000 Hz) pure tone threshold averages for normal hearing participants

Study	3-frequency pure tone threshold average in dB HL (standard deviation)
Current study	7 (5)
Nissen et al. (2011)	4.5 (3.0)
Nissen et al. (2007)	5.0
Lau and So (1988)	15.3
Nissen et al. (2005)	3.0 (2.7)
Wang et al. (2007)	5.8

Clinically, outcomes to speech audiometry using words as stimuli are presented in the form of performance-intensity function (P-I function), a curve displaying the correct score achieved for each of the tested presentation intensity levels. Several assumptions can be made in reference to the P-I function, mainly the speech hearing threshold and the maximum level of speech recognition that can be achieved by the listener. The speech hearing threshold, also known as speech reception (or recognition) threshold (SRT) is defined as the level giving 50% recognition probability for the test items, in this case, the phonemes. On a P-I function curve, this is marked as the half peak level (HPL). The maximum level of speech recognition, also known as Maximum Speech Recognition Score (MSRS) or PBmax, is defined as the maximum score achieved on a P-I function. MSRS is used to estimate the listener's speech recognition performance, and therefore, provide the estimate of the listener's maximum ability to understand speech (Gelfand, 2009). The overall shape of the P-I function curve would also provide an indication of the type of hearing loss the listener might have.

In general, the P-I function curves of all normal hearing participants (excluding participant L05), in both AWL and MWL, followed the shape usually found in speech audiograms using word stimuli (Boothroyd, 1968; Lau and So, 1988; Mukari and Said, 1991; Nissen et al., 2005a; Harris et al., 2007). The S-shape function reflected on how speech is perceived; the flatter curve at lower presentation levels corresponds to stimulus intensity levels that are below the listener's audibility. The speech signal is spread over a range

of frequencies and intensities. As the level is raised, the speech components with the highest amplitudes start to be audible. The higher the presentation level, the wider the range of energy gets over the threshold of initial audibility, therefore, more speech acoustic signal is heard. In the case of normal hearing listeners, full score is attained when all of the stimulus's speech signals are audible. Any increase of intensity above this point does not result in change in the amount of speech audibility and, therefore, the scores, which results in the plateau at high presentation intensities.

The characteristics and the components of the P-I function curve, including the HPL and the MSRS, were analysed to compare the performance of normal hearing listeners to AWL and MWL. It is interesting to find that the P-I function curves of AWL and MWL to be strikingly similar. The MWL curve showed slightly higher correct scores between stimulus levels of 10 to 25 dB dial, a range that corresponds to the steeper slopes in the P-I function curve. The largest difference is found 15 dB dial presentation level, with only 3.6% difference in correct scores.

As the MWL showed slightly higher scores compared to AWL especially at the steeper part of the P-I function curves, it was also expected that the HPL for the MWL is lower than the AWL. Two calculations of the HPL were made; one was the actual HPL of the P-I function curve, and the other was calculated as the mean of the individual HPLs of the participant in the group. Although the HPL, thus the speech reception threshold (SRT), using MWL is slightly lower than AWL, the measurements showed that the differences are less than 0.5 dB, which may be treated as negligible.

The findings of the P-I function and the HPL contradict the anticipated outcome. Previous studies have found that, due to the unfamiliarity of the words, the P-I function obtained using unfamiliar words showed much more gradual slope reaching to the plateau (Zakrzewski et al., 1975; Gelfand et al., 1992; Cheesman and Jamieson, 1996; Boothroyd, 2008). Correct scores were also lesser with nonsense test items compared to meaningful test items in equivalent presentation levels (Hume, 2002). This resulted in a function curve that is displaced toward the right. The unfamiliarity of the test items, lack of redundancy and lack of contextual cues result in higher presentation level required to gain a score equivalent to responses using meaningful test items, therefore, increasing the HPL. This pattern is persistent even with nonsense test items that are phonemically and structurally equivalent to their meaningful counterpart (Zakrzewski et al. 1975). Having a mix of meaningful and nonsense words as test items was expected to affect the P-I function and HPL in a similar way, although possibly not as severe as the effect

given by a list of all nonsense words. However, the findings in this clinical validity study has shown that the outcome of AWL and MWL is not significantly different. A closer look at the HPLs of other studies even showed great variability between sets of word lists constructed using familiar words (Figure 5.21), which further suggests that there may be factors other than word familiarity that affect the average HPL value.

Table 5.21 Summary of half peak level averages for normal hearing participants

Study	Mean HPL
Current study (AWL and MWL)	10 dB dial
Nissen et al. (2011)	8.7 dB HL
Nissen et al. (2007)	4.4 dB HL
Lau and So (1988)	21.5 dB SPL
Nissen et al. (2005)	5.4 dB HL
Wang et al. (2007)	6.4 dB HL

The reference ranges for both AWL and MWL showed HPL with almost identical variations. The HPL reference range, with 95% confidence interval, for both AWL and MWL was calculated at 10 ± 6 dB dial. The variation of HPL is similar to the acceptable variation of pure tone thresholds in clinical setting, which is 'threshold ± 5dB'. This suggests that, based on the random variability of the HPL for both AWL and MWL, speech audiometry using the bisyllabic Malay word lists is a valid test of hearing and the variation is consistent to the conventional clinical validation for hearing tests. The actual value of HPL was not necessarily identical to that of the pure tone threshold average, as the units are different and the output intensity of the word lists were not calibrated to dB HL, the unit used for pure tone audiometry.

There are several reasons that may explain the differences in the outcomes of this study and the findings of previous studies, mainly focusing on the structure of the test items. A major difference between the test items used in previous studies and reports (Miller, Heise and Lichten, 1951; Zakrzewski et al., 1975; Cheesman and Jamieson, 1996) and the current study is the number of syllables per item. The test material of the previous studies, as well as the established Closed-Response Nonsense Syllable Test (CUNY NST), had used monosyllables as opposed to the bisyllabic words that were used in this study. Zakrzewski et al. (1975), used phonemically balanced monosyllables formed with several types of word structures (CV, CVC, CCV etc.) while Cheesman and Jamieson

(ibid.) presented words which varies only in the initial consonant, C (CIl). This suggests that the bisyllabic structure used in the current bisyllabic Malay word lists, particularly the nonsense words, may have provided the listeners additional acoustic and contextual cues. The words in the Malay bisyllabic word lists, for both meaningful and nonsense, are constructed following the Malay phonetic rules. The additional syllable in the current test items, i.e. bisyllable vs monosyllable, as well as the information from the limited syllable combinations following the Malay phonetic rules may have provided the listener with contextual cues that can aid the perception of test items. As an example, the nonsense words in the current word lists were designed to follow the general pattern of Malay vowel occurences that restricts the second vowels for open syllables in CVCV words to [i], [ə] and [u][1] (Teoh, 1994). This limits the choices of vowels that can be 'selected' by the participants and increases the probability of being correct in their response. Another possible factor is the level of fidelity in the phoneme distribution of the words lists to the phoneme distribution of Malay CVCV words in general. Phonetic balance and phonetic distribution that highly resemble the balance of phonemes in CVCV words of the corpus sourced from the daily newspapers (as discussed in the earlier chapter) further limits the selection of phonemes in the response to the test items. Higher occurences of certain phonemes and the usage of Malay phonemes only increases the probability of the participant giving correct responses. These factors may have affected the performance of the participants in a way that the performance for nonsense words is similar to that of meaningful words.

A comparison between the maximum score or MSRS of AWL and MWL showed that both wordlists were able to elicit almost 100% correct response in normal hearing participants. Again, the MSRS difference between AWL (99.13%) and MWL (99.6%) is very small. These high scores at suprathreshold presentation levels are expected from normal hearing listeners; good frequency discrimination in normal hearing participants allows full audibility of the words at high enough presentation intensity levels. In both versions of word lists, the MSRS was reached at stimulus presentation level of 35 dB dial. However, on average the plateau started at a lower level than 35 dB dial; AWL began to plateau at 30 dB dial while MWL started to plateau at 25 dB dial. The 5 dB difference between AWL and MWL can be contributed to MWL which are totally constructed of meaningful words, as opposed to AWL. It suggests that, although both

[1] The phoneme [a] is also present in the second vowel position for open syllables in CVCV words, although not as common as the other three phonemes. Examples include 'busa' (foam), and 'massa' (pronounced as /masa/; meaning 'mass')

AWL and MWL reached half peak levels (HPLs) at approximately the same stimulus intensity levels, AWL required slightly higher stimulus presentation level to reach total audibility. This is consistent with previous findings that stated that nonsense test items required higher presentation levels in order to achieve similar performance, i.e. correct scores, as their equivalent but meaningful items (Miller, Heiser and Lichten, 1951; Webster, 1972). The mixture of nonsense and meaningful words in AWL is proposed to be the reason for the minimal difference in the onset of plateau.

A more interesting point was the fact that both AWL and MWL were able to elicit almost 100% correct scores at suprathreshold levels. There were several contradicting reports regarding the issue; there are previous studies that found full audibility and 100% or almost 100% scores for MSRS using nonsense syllables (Webster, 1972; Cheesman and Jamieson, 1996) while others reported reduced MSRS for nonsense syllables as compared to other types of test items. (Miller, Heise and Lichten, 1951; Mendel and Danhauer, 1997). Modified CUNY NST, a widely used nonsense syllable speech audiometry material, reported maximum correct scores of 91.9% and 92.6% for its VC and CV test items, respectively (Gelfand et al. 1992). Again, two major differences between the current study and the previous studies are the number of syllables per test item as well as the mix of nonsense and meaningful words in the lists. As with the case of the HPL, the bisyllabic test items, as opposed to the monosyllables in the previous studies, may have facilitated the listeners by giving extra acoustic and contextual cues. The majority of meaningful words in the AWL (ratio of meaningful words to nonsense words is 2:1) may also played a part in elevating the performance of listeners as compared to lists that are wholly assembled of nonsense syllables.

Based on the P-I function curves of the normal hearing participants, it can be said that both versions of the bisyllabic Malay word lists are equally able to reflect the progress of audibility of speech in normal hearing, demonstrate the aspect of full audibility at suprathreshold levels through the MSRS and demonstrate normal reference ranges that agree with the accepted variation in hearing tests. With the exception of the correlation between HPL and pure tone thresholds, these findings suggest that the word lists are clinically valid. The clinical validity of the HPL based on pure tone thresholds will be discussed further below.

5.4.3 Performance-Intensity function in participants with sensorineural hearing loss

The P-I functions of participants with sensorineural hearing loss (SNHL) were more diverse in shape than the P-I function of normal hearing participants. This is mainly due to the varied severity and configuration of SNHL within the group. The wide range of degrees of SNHL was important in determining the correlation between the level of hearing loss and the characteristics of P-I function. The variety of hearing loss configuration further helped the aim to clinically validate the bisyllabic Malay word lists.

Three characteristics of the P-I function were looked into in order to answer the questions of whether the results obtained using the word lists were able to predict the speech hearing acuity and ability. The P-I function curve was analysed for its general shape. The HPL, which served as an indication of speech hearing threshold, and the MSRS, which was expected to reflect the listener's ability to distinguish phonemes, were also looked into.

5.4.3.1 P-I function curve in general

The shape of the P-I function curves generally follows the "S" shape found in normal hearing participants. The shape started with the bottom, flat segment, corresponding to the level at which the stimulus could not yet be heard by the listener. The following steeper segment correlates to the level at which the highest amplitudes in the speech signal was audible to the listener. Full audibility, or in some cases, the highest level of audibility achievable by the listener, were represented by the top and again flat segment, also called the plateau.

There were several deviations in curve shape seen in the SNHL group compared to the normal hearing group. Several participants revealed missing lower flat segment, showing curves that began directly with the steep segment instead. Some of the participants showed curves that seemed to be 'floating' due to their lowest correct scores being more than 0%. There were also curves that presented with rollover effect, of which the scores decline at stimulus levels higher than the level that produced the maximum score. These three marked differences can be attributed to the features of SNHL. SNHL were found to cause reduced dynamic range, with greater reduction of dynamic range as the loss

progresses (Pascoe, 1988). The range between the hearing thresholds to a level that cause discomfort becomes narrower as the hearing loss becomes worse. This means that the progress from no audibility to maximum audibility requires less increase in intensity in listeners with SNHL as compared to normal hearing listeners, which could explain the lack of the lower gradual segment in the P-I function of several of the participants. Absence of this 'tail' may indicate that the word lists are sensitive to the difference in dynamic range of listeners with normal hearing and those with SNHL. Further investigation on the correlation between dynamic range and the P-I function would be necessary to confirm this.

Although no literature can be found discussing this characteristic, several curves appeared to be 'floating', as the correct scores at the lowest stimulus presentation level were higher than the scores found in the normal hearing group. For these cases, the minimum stimulus level that was set in the research design was high enough for at least part of the stimulus to be audible to the participant, even though the level were estimated to be below the speech threshold. Sloping hearing losses, of which the hearing at certain frequencies is better than the others, tend to show this condition, as there were segments of the speech stimulus that were intense enough to be audible. HL4 (Case 1), for example, was a participant with a sloping hearing loss, with the lower frequencies having considerably better thresholds than the higher frequencies. The pure tone average at 500, 1000 and 2000 Hz that was used to calculate the minimum speech presentation level was much higher than the thresholds at the lower frequencies (250 and 500 Hz). This, combined with the findings of LTASS that showed greater emphasis in lower frequencies for Malay speech sounds, may have resulted in the participant being able to hear the speech stimulus at the minimum presentation level. The P-I function in HL4 was shown to be missing the steeply sloping segment altogether, as well as producing correct scores of more than 50% even at the lowest speech presentation level, which restrict precise calculation of the HPL and, therefore, SRT. It is suggested that this characteristic be taken into consideration when conducting speech audiometry using the current word list; it could be utilised as an indication of the configuration of SNHL, whether it is flat or sloping. On the other hand, it would indicate the use of different combination of pure tone threshold average instead of the 500, 1000 and 2000 Hz combination in calculating the initial speech presentation level in order to produce a more extensive P-I function, which allows better HPL measurement, and therefore more accurate estimate of the speech perception of the listener. The findings of le Andrade et al. (2013) indirectly supports this suggestion; it was found that, for those with upward- or downward-sloping audiograms,

the pure tone averages of 500, 1000, 2000 and 4000 Hz were the most significant in predicting the SRT.

Rollover effect is an established method of determining the possibility of having SNHL of retrocochlear origin. Indication of retrocochlear hearing loss is based on the rollover index, taking into account the MSRS and the minimum score obtained at a level higher than the level for MSRS. Several of the P-I function in the SNHL group displayed rollover effect. This was, to a certain degree, expected in both subcategories of SNHL, which are cochlear hearing loss and retrocochlear hearing loss (Jerger and Hayes, 1977; McArdle and Hnath-Chisholm, 2015). However, one case, HL4, stood out for having significant rollover suggestive of retrocochlear disorder according to the value recommended by Mueller and Hall (1996). Although confirmation of the disorder was not possible due to lack of supporting diagnostic data, this finding suggests that the bisyllabic Malay word lists has the capability in detecting different types of SNHL. This is especially important in clinical assessment as it would provide supporting data in order to determine further steps in assessment and management of the hearing impairment.

5.4.3.2 Half peak level (HPL) and maximum speech recognition score (MSRS)

As shown in the findings, the HPL obtained using both AWL and MWL seemed to be consistent with the hearing level. The higher the severity of hearing loss, the more staggered to the right the P-I function curve, which corresponds to higher HPL. This is expected of speech audiometry curves in cases of hearing loss as described by previous studies (Hood and Poole, 1971; Jerger and Hayes, 1977; Hong, 1984; Wang et al., 2007; Han et al., 2009). A higher stimulus intensity is needed to overcome the attenuation caused by the loss of hearing, therefore shifting the P-I function curve towards the right on the horizontal intensity axis and increasing the HPL. This finding suggests that the HPL correlates with the pure tone hearing thresholds, and therefore, indicates that the word lists are able to provide an impression of the speech hearing ability of the listener. The significance of the correlation between HPL and pure tone thresholds are discussed in detail below.

The MSRS in participants with SNHL are, however, better that expected. Due to the nature of SNHL, the frequency discrimination ability is reduced as a result of damage in the cochlea and/or the neural pathway of the hearing system. This decrease in frequency

discrimination, in addition to increased threshold, affects the phoneme discrimination and recognition during speech audiometry. In a more severe losses, the effect of frequency discrimination is more pronounced, resulting in lower MSRS as compared to a milder loss. In the current study, the MSRS obtained using both AWL and MWL were found to be high, including those with severe hearing losses, as seen in sample case of participant HL9 and, to some extent, participant HL4. Several factors may contribute to this finding; the length and the structure of the word as well as the scoring system may have resulted in higher peak scores. The bisyllabic, phonetically-balanced form of the test items in the current word lists might have provided the listener with additional acoustic and contextual cues for the stimuli. Phonetic balance of the word lists which closely imitate the phonetic balance of Malay CVCV words in general would have given some degree of probability in terms of the phoneme choices. As an example, /r/ which has greater distribution in Malay consonants, occur more frequently in the word lists compared to /c/, a less common Malay consonant. This would indirectly provide a Malay-speaking listener with some contextual cues, thus increasing the probability of a correct response. Similarly with the phonetic balance, the word structure that follows the rule of phoneme combination in Malay may also have an effect to the MSRS. Limited combination of phonemes, especially in the final syllable, would also give contextual cues to the listener. Lastly, the phoneme scoring used in the current word lists would allow for lesser weight per test item as compared to word scoring, as each phoneme in the list carries its own score and, therefore considered as one test item. For example, the word 'BUKU' carries 10% of the total score in a list of 10 CVCV words. A response of 'BURU' would contribute 0% to the total score in word scoring system instead of 7.5% in phoneme scoring system. Although it seems that the phoneme scoring allows for more 'lenient' measure, it has to be stressed that the objective of MSRS is to measure the phoneme recognition abilities of the listener and not comprehension. The phonemic scoring applied in the current wordlists is also favourable as it allows for finer assessment of speech frequency discrimination.

5.4.4 Performance-Intensity function in participants with conductive hearing loss

Conductive hearing loss is a type of hearing loss caused by mechanical impairment in the outer and middle ear. The mechanical impairment disturbs the sound energy transfer into the inner ear and attenuates the sound. Because the disorder is mechanical and not

affecting the cochlea, the loss is characterised by reduction of sensitivity to sound intensity without any decrease in frequency discrimination ability. Therefore, although increase in HPL is expected in listeners with conductive hearing losses, the speech discrimination, represented by the MSRS, should show high scores, similar to those with normal hearing.

Utilising AWL and MWL resulted in similar P-I function curves in participants with CHL. This shows that the difference in the content of the word lists did not majorly affect the speech audiometry results. Both versions showed the general S-shaped curves similar to those seen in participants with normal hearing and sensorineural hearing loss.

However, there are several differences noted in the CHL group compared to the normal hearing and SNHL group. As expected, the curves are displaced towards the right on the horizontal axis as compared to the P-I function of normal hearing participants, consistent with the findings of previous studies (Boothroyd, 1968; Hood and Poole, 1971; Hong, 1984). The displacement is consistent with the increase of hearing thresholds due to the attenuation caused by the abnormality in the mechanical transmission of sound. The displacement would cause the increase of the HPL, which should be proportional to the increase in hearing level. The correlation between the HPL and the hearing threshold, represented by pure tone hearing thresholds, are discussed in the following section.

The plateau section is evident in all of the P-I function curves. The exceptionally high scores for the maximum point plateau, all above 98% correct, signify full phoneme discrimination abilities in the listeners and indicate the intact frequency discrimination feature of normal cochlear function, both characteristic of conductive hearing loss. No rollover is seen in the findings, as opposed to the findings of participants with SNHL. Again, these features are consistent with previous findings by Boothroyd (1968), Hood and Poole (1971) and Hong (1984).

Upon closer inspection, some of the P-I function curves seem to be ‘floating’. This feature was not seen in the normal hearing group; however it is present in the SNHL group. The reason for the ‘floating’ curve was that the lowest stimulus presentation level was not low enough to reach the point of speech inaudibility for the listeners and therefore the scores did not reach zero. The difference in hearing configuration may have resulted in the scores obtained at very low presentation levels. As outlined in the research design, the presentation levels were set based on the 3-frequency pure tone threshold average. A sloping or rising pure tone audiogram, where some frequencies have better thresholds and therefore are more sensitive than others, may result in some of the phonemes being

audible to the listeners despite the low presentation levels. However, in a clinical setting, the lowest score is of lower clinical and diagnostic importance as opposed to the HPL (or SRT), MSRS and presence of rollover, therefore, 'floating' P-I function curves should be acceptable as long as the HPL are achieved.

5.4.5 Correlation between the speech reception thresholds (SRT) and pure tone hearing thresholds (HTL)

The clinical validity of the bisyllabic Malay word lists were further confirmed by measuring the correlation between the speech reception thresholds (SRT) and the pure tone hearing threshold averages. In this case, the half peak level (HPL) of the P-I function represents the SRT, while the HTL is represented by a combination of thresholds at several frequencies. An agreement between the SRT and PTA shows that the word lists are robust enough to reflect the hearing acuity, particularly the speech hearing ability.

Several combinations of pure tone thresholds were tested against the SRT in order to find the pure tone combination with the best fit against the SRT. The commonly used combinations in speech audiometry testing are the 500, 1000 and 2000 Hz threshold average (Carhart, 1951; Lau and So, 1988) and 500, 1000, 2000 and 4000 Hz threshold average (Wang et al., 2007, Neumann et al., 2012; Weißgerber et al., 2012) However, considering that the differences of frequency emphasis found between the Malay LTASS and the English as well as the universal LTASS, there is a possibility that better correlation could be found with a different set of combination.

Based on Spearman's Rank correlation, all pure tone (PT) average combinations tested showed statistically significant correlation between SRT and pure tone hearing thresholds, both using AWL and MWL, in all three participant groups. This significant correlation answers the question of whether or not the Malay word lists are able to reflect the hearing level, with higher PT threshold showing higher SRT and vice versa. The strength of the correlation, however, varies between combinations, as well as between participant groups.

An apparent feature of the PT combination showing the highest correlations in normal hearing participants and participants with SNHL is the inclusion of 250 Hz to the normal PT average used in the previous studies (Wang et al., 2007, Neumann et al., 2012; Weißgerber et al., 2012). The inclusion is consistent with and reflects the finding Malay

LTASS in the previous chapter, which showed considerably higher energy in the frequency range of 100-400 Hz as compared to the universal LTASS and English LTASS (Cox and Moore, 1988; Byrne et al., 1994). The lower frequency emphasis in Malay is further enhanced by the high percentage of vowels, of which spectral frequencies concentrates in the lower frequency region.

It is, however, interesting to find that the PT combination with the highest correlation in participants with CHL is different from the other two groups. Although there are evidences of very good correlation with PT averages that include 250Hz, the PT averages with the strongest correlation in CHL are 500, 1000 and 4000 Hz for AWL and 500, 1000, 2000 and 4000 Hz for MWL. These findings contradict the findings of normal hearing and SNHL groups. Closer inspection of the pure tone thresholds of the CHL participants did not yield any significant patterns or configurations of hearing loss (for example, rising audiogram) that may explain the difference. A possible explanation is, although the Malay speech sounds are more emphasised at the lower frequencies, in CHL cases where the frequency discrimination ability is not affected, the higher frequency speech sounds tend to give more impact towards the speech perception.

Groups with SNHL and CHL generally showed stronger correlations in all PT average combinations between SRT and PTA combinations compared to the normal hearing group. This difference may mainly be contributed by the difference in research design between one that was applied to the normal hearing participants and one applied to the participants with hearing loss, as well as the statistical analysis used in measuring the correlation. Normal hearing participants were presented with a standard set of presentation intensities for the speech stimuli, as opposed to the hearing loss participants, where the presentation levels were dependent on the pure tone thresholds. The variation in the individual participant's hearing thresholds in the normal hearing group, together with the rank correlation analysis, may have influenced the SRT and affected the statistical outcome.

One of the clinical application for SRT is to validate pure tone hearing thresholds (Boothroyd, 2008), as opposed to the current research design where pure tone thresholds are used to validate the SRT. The question is, with the contradictory findings of the PT averages with the strongest correlation, which is the best PT average combination to be used in pure tone threshold validation? The findings of correlation analysis suggest that the best PT average combination is 250, 500, 1000, 2000 and 4000

Hz for normal hearing listeners and listeners with SNHL, and 500, 1000, 2000 and 4000 Hz for listeners with CHL.

5.4.6 Half peak level (HPL) – pure tone (PT) average agreement

To conclude the clinical validity assessment, the sensitivity, specificity and predictive values of the HPL of the HPL-PT average agreement were calculated to determine whether the bisyllabic Malay speech audiometry word lists are robust enough to be utilised in a clinical setting. The agreement between pure tone threshold and speech reception threshold adds to the information regarding the patient's hearing ability, thus supports the decision making in audiological diagnosis and management (Boothroyd, 1968). It is measured by the HPL-PT average difference; the smaller the difference, the better the agreement between the HPL and the PT average.

A decision was made to have a single PT average frequency combinations for the calculation of HPL-PT average difference for all groups based on the opinion that one of the desirable features of a speech audiometry is its ability to allow the tester to complete the test within a short amount of time (Hirsh et al., 1952; Boothroyd, 1968; Harris et al., 2007). The correlation between the speech reception thresholds (SRT) and pure tone hearing thresholds (HTL) suggested two separate PT average combinations for normal hearing and SNHL groups and CHL group, respectively, as discussed in the previous section. Building from that idea, the PT average for 250, 500, 1000, 2000 and 4000 Hz was selected as the frequency combination for which the HPL-PT average agreement is calculated. The justification for this selection was that the PT average with this combination showed, in general, very high correlation to the HPL across the participant groups.

The findings of mean HPL-PT average difference suggest that there are disparities between the HPL and the PT average. Two interesting outcomes were found; firstly, HPL is higher, and, therefore, worse, than PT average in normal hearing participants, almost the same in CHL participants, and lower, which means better, in SHNL listeners. Secondly, the standard deviation is larger in hearing impaired groups than in the normal hearing group. The frequency combination used to calculate the PT average does not account for hearing configurations other than flat, which could explain why there is a

discrepancy between the HPL-PT average difference in normal hearing group and the groups with hearing loss. The HPL to PT average difference in CHL group, although smaller than the difference found in the normal hearing group, also indicate that the hearing for speech relative to pure tone hearing is better in CHL than in normal hearing. A sloping hearing loss may have caused the HPL to be lower and, therefore, better than the PT average, as the speech with low frequency content is heard first and at lower intensity compared to high-frequency speech sounds, thus affecting the P-I function curve. This outcome has been described in an earlier study by Boothroyd (2008) where the calculated phoneme recognition threshold is inconsistent with the SRT in a case of severe high frequency SNHL. The effect of sloping hearing loss on the SRT-pure tone threshold agreement was also discussed by American Speech-Language-Hearing Association (1988), who suggested using two-frequency pure tone average, and Gelfand's and Silman's 1985 and 1993 studies (cited in Gelfand, 2009, p. 143) who suggested the single best pure tone threshold as the reference for SRT – pure tone threshold agreement.

5.4.7 Predictive analyses

Having two separate correction factors (CF) for normal hearing and CHL groups (CF=3 for both groups) and SNHL group (CF=-7) gives the best sensitivity, specificity and predictive values for the bisyllabic Malay word lists in comparison to no correction factor or applying the same correction factor (CF of +4 or +3) across all groups. Building from the idea of having reference speech reception threshold level (RSRTL) as the zero point for the measurement of speech reception threshold level (ISVR, 2003) as well as different PT average frequency combinations for different hearing loss types and configurations, the sensitivity, specificity and predictive value findings illustrate that separate correction factors for different types of hearing loss may also be a possible method of HPL-PT average agreement calculation. Very high sensitivity (>0.86) and specificity (1) with the standard error set at ±10 dB indicate that this method of HPL-PT average agreement calculation allows for very good identification of the presence of hearing loss as well as proves that the word lists is applicable in cross-checking pure tone thresholds. The predictive values are also very high, which means that the bisyllabic Malay word lists are reliable and valid as an indicator of the level of hearing for speech.

The sensitivity, specificity and predictive values are also very high for HPL-PT average agreement with no correction factor, particularly in separating those with normal hearing

and those with hearing loss (without accounting for the type of loss). The values are particularly better for MWL compared to AWL, a trend that is consistent throughout all three methods of calculation. The high specificity, sensitivity and predictive values can be contributed to the compromise between the correction factors of normal hearing/CHL and SNHL, as opposed to the application of single correction factor across all types of hearing loss, although the values are nor as high as having specific correction factors for normal hearing/CHL and SNHL.

There is no specific validity studies based on sensitivity, specificity and predictive values done on any speech audiometry material found in the literature. However, sensitivity, specificity and predictive measures determines whether or not the test is valuable to clinicians, depending on how accurately the test correctly detects the presence of disease or abnormality and how the test correctly identifies patients who are disease-free (Lalkhen and McCluskey, 2008). Based on other established audiometric tests such as screening pure tone audiometry, which showed sensitivity of 0.87 and specificity of 0.8 (Sabo et al. 2000), tympanometry in predicting hearing impairment in otitis media cases, which showed sensitivity of more than 0.9 (Group, 1999), and auditory brainstem response in detecting hearing loss in neonates, with sensitivity and specificity of ≥0.81 and ≥0.91, respectively (Hyde et al., 1990), it can be said that the level of sensitivity and specificity showed by the bisylllabic Malay word lists is comparable with current audiometric tests.

The 95% confidence interval was calculated as ±7 dB based on the standard deviation of the HPL-PT average difference of the normal hearing group. This is much stricter than other intervals suggested in the literature; Bess and Humes (2003) recommended ±10 dB agreement between HPL and PT average, Chaiklin and Ventry (1964) as cited by Gelfand (2009) suggested SRT-PT average discrepancy of ±12 dB as acceptable, while Boothroyd (1968) allowed up to 15 dB difference between the pure tone thresholds and speech reception thresholds. Having more stringent cut-off decreases the sensitivity and specificity of the bisyllabic Malay word lists in detecting hearing loss, especially in the SNHL group, which may cause unnecessary alarm for non-organic loss in cases of cross-validation for pure tone audiometry. A more lenient cut-off point of ±10 dB is recommended as it improves the sensitivity, specificity and predictive values. In speech audiometry using AWL, the milder cut-off point results in an improvement of as much as 10% and 19%, for the HPL-PT average agreement AWL for hearing loss in general and

SNHL, respectively. Moreover, the ±10dB is comparable to the range suggested in the literature.

In conclusion, the findings from the analyses of HPL-PT average agreement, P-I function and predictive values show that AWL and MWL have the ability to determine speech reception thresholds, reflect hearing level and, to an extent, suggest the type of hearing loss. This validates the use of AWL and MWL in measuring speech hearing ability in clinical use.

5.5 Limitations of research and future study

The findings of this study, based on the ability of speech audiometry using bisyllabic Malay word lists to separate different levels and types of hearing loss, suggest that both of the AWL and MWL sets are clinically valid. Speech reception thresholds using either set of word lists strongly correlate with the pure tone hearing thresholds, thus able to provide a measure of speech perception for both normal hearing participants and participants with hearing problems. Although care has been taken to ensure that the sample size provides good statistical power, all of the participants were residents of Kuantan. The study has tried to minimise the bias by using Standard Malay language; however, linguistic differences such as accents might still have a small effect to the results, suggesting that the results are best generalised to the population of the area. It is recommended that the speech audiometry material be tested in other states in Malaysia in order to have more comprehensive information regarding the speech audiometry material. The linguistic differences between different races in Malaysia may also have an impact on the performance; although Malay is the national language, there is a part of the population whose first language is based on their ethnicity, such as Chinese, Indian or Iban. Higher number of participants, with more varied types and levels of hearing losses, would be desirable. The study could also be extended to older children and teenage listeners, as the material is fairly elementary.

The evaluation of hearing impaired groups showed that the bisyllabic Malay word lists was able to reflect on the hearing thresholds and, to a certain extent, reflect on the speech audibility of the listeners and the limitations posed by their impairment. However, evidence on the ability of the word lists to separate different types of hearing loss is severely limited. The P-I functions of the hearing impaired participants showed the characteristics related to certain types of hearing loss (for example, retrocochlear hearing loss), however, the diagnoses could not be established without other supporting

evidence. A study concentrating on a particular, and preferably diagnosed, type hearing impairment, supported by other diagnostic tests, should be considered in the future.

A concession for masking has to be made for cases with unilateral or assymetrical hearing loss. As shown in sample case of participant HL15, the HPL and MSRS did not correspond with the expected MSRS. There is a possibility of stimulus crossover being audible on the non-test ear due to insufficient masking, thus making the response not exclusively based on what is on the test ear. Although current study utilised the masking approach recommended by Yacullo (1999), there is no confirmation that the level of masking is adequate for the current speech stimuli. Two further investigations are suggested based on this limitation; one, measurement of the intensity of the bisyllabic Malay words in the unit of dB HL, and two, the measurement of effective masking level in regard to the word lists.

CHAPTER 6 CONCLUSION

6.1 Introduction

The study was set out to develop a clinically valid speech audiometry material to test the hearing for speech and assess speech recognition in Malay speaking adults. It has identified the need for a verified and validated phonetically balanced speech material for the Malay speaking population. There is also a gap in knowledge regarding speech recognition assessment using speech material containing a mix of meaningful and nonsense words, as to simulate familiarity level of everyday speech. This chapter concludes the whole thesis: the theoretical model of development of speech audiometry word lists, future research and limitations regarding the developed material.

The findings of this study are integrated to synthesise a valid and standardised speech recognition test material. The research design is adapted to produce a test protocol suitable to be performed for hearing assessment. There are two test formats, the combination of meaningful and nonsense word lists, known as all-word lists (AWL), and shorter, meaningful words only lists (MWL). The P-I function of normal hearing and hearing impaired clients from this study forms the normative data on which future comparisons can be made.

The Bisyllabic Malay Speech Audiometry (BMSA) word lists are compiled into a test kit which includes the introduction to the test, a summary of the development of the test, the recommended test procedure and interpretation of results. The word lists in written form, included in the scoring sheets, as well as the recorded stimuli are also included in the test kit.

The standardised BMSA is fit to be introduced into the Malaysian clinics as part of the hearing assessment battery. The required equipment are an audiometer with external input channel and a compact disk player, all of which are readily available and easily accessible in an audiology clinic. The test does not require special training to perform, therefore, it can easily be carried out by any qualified audiologist. The application of speech audiometry for Malay adult speakers will add to the information needed for better diagnosis and management of their hearing loss.

6.2 Theoretical implication

In theory, there are two major contributions given by the current study. The first contribution involves the method of development of bisyllabic word list. A new framework on the development of word lists for speech reception threshold test is constructed based on the development process of the word lists. There are three main components that are essential in the development of a speech audiometry, or in this case, speech recognition test, material. In order to develop a clinically sensitive material, first, the purpose and structure of the test has to be identified. These factors influence the design and development of the test material. After the material is constructed, the next step would be to verify and validate the test items in order to arrive to a final set of speech audiometry material. This verified and validated set is then clinically tested in order to measure its normative values, such as standard error and reference range, and its effectiveness in achieving the intended purpose. The development of bisyllabic Malay speech audiometry wordlists followed these steps in order to produce a clinically valid speech audiometry material. However, in order to overcome limitations and fulfil the requirements of each step, several modifications and additions are put into the research design. The following are the conclusions made regarding the theoretical model of the development of BMSA material in regards to the three main components of the development and the limitations they impose. Figure 6.1 summarises the development of speech reception threshold test framework. Sections 6.2.1, 6.2.2 and 6.2.3 discusses in detail the processes involved in the development.

The second major theoretical contribution of the current study is a product that was synthesised from the development process of the word lists and comes in the form of speech audiometry test kit. The test kit consists of two parts – a booklet containing the background of the BMSA word lists and a compact disk containing the audio file for the word lists. Section 6.2.4 discusses the BMSA test kit further.

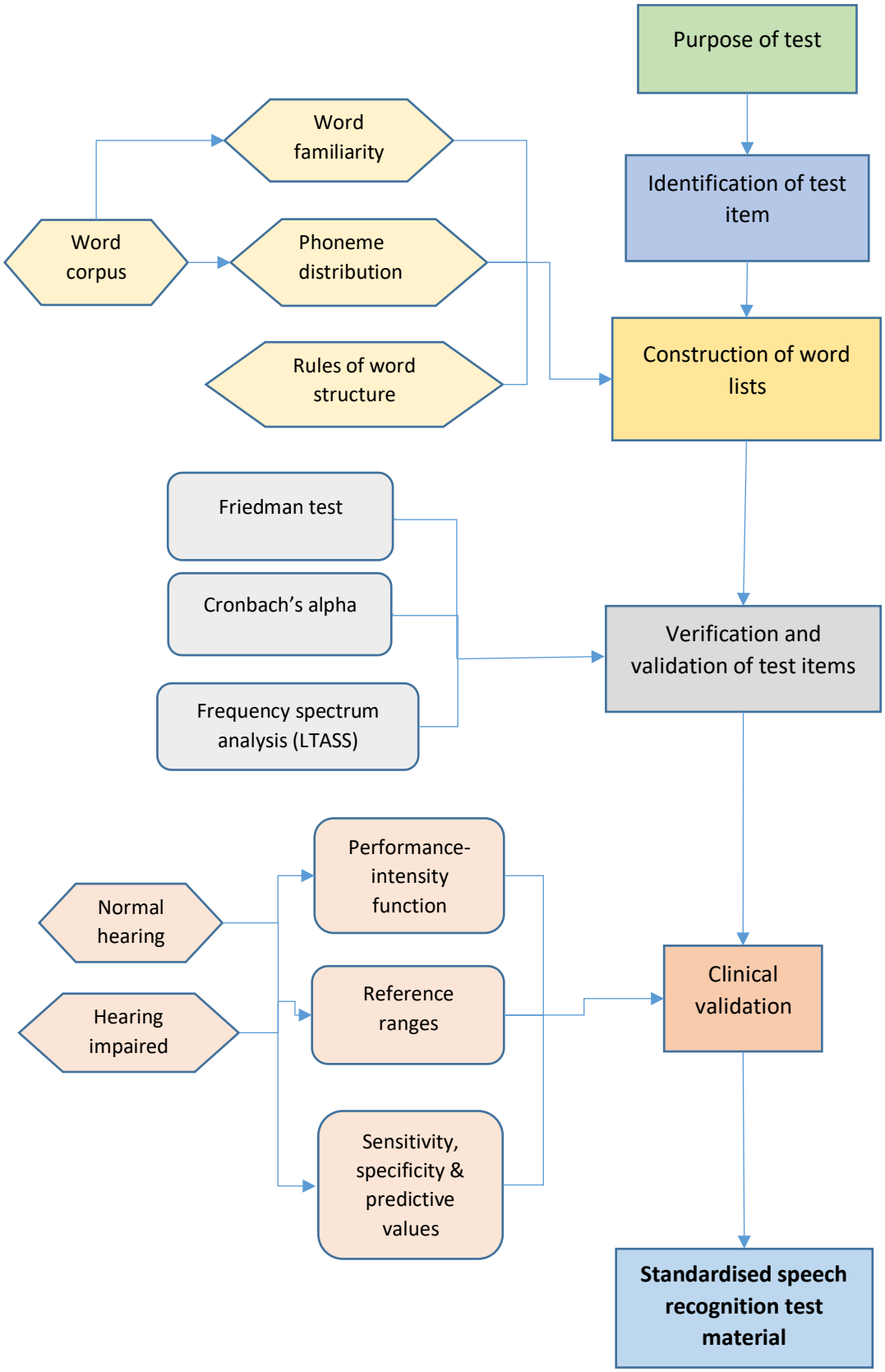


Figure 6.1 Summary of the theoretical framework of the development of BMSA

6.2.1 Construction of word lists

The purpose and structure of the test is decided upon identifying a need of a standardised speech recognition test material in Malay following a review of the available speech audiometry materials in Malay. The mixed mode of word familiarity offers a set of word lists that better reflect the variety of word familiarity in natural speech through the addition of nonsense words which provide minimal contextual cues, while still keeping to the bisyllabic nature of speech recognition test material. To allow clinicians who prefer the traditional format of lists that contain only meaningful words, an alternative all-meaningful BMSA word lists

There are several prerequisites in order to design phonetically balanced word lists containing meaningful and nonsense words. Knowledge on the distribution of phonemes is essential in forming the basis for phonetic balance. Rules of word structure are important in constructing nonsense words that resemble true words. The meaningful words in the BMSA lists are assembled according to the familiar- or frequent-word principle, similar to the previously published speech recognition test materials (Boothroyd, 1968; Ashoor & Prochazka, 1982; Harris, et al., 2007; Wang, et al., 2007; Han, et al., 2009), therefore, a corpus of familiar and/or frequent words as the source of test items is also crucial.

In this study, the use of text analysis software for printed media offered online by Dewan Bahasa dan Pustaka (Council for Language and Publication) is designed to overcome the absence of Malay word corpus. With the consideration of adults being the target clients for BMSA, it is thought that daily newspapers, rather than books or textbooks, are more suitable in terms of the vocabulary as the source for the study of phoneme distribution of phonemes and construction of the word pool of potential test items. The words taken from daily newspapers published within a period of five years were filtered for the target word structure (consonant-vowel-consonant-vowel), and the resultant word pool served as the corpus from which the distribution of phonemes are calculated, and the most frequent words are ranked. As word frequency correlates with word familiarity (Tanaka-Ishii and Terada, 2011), the frequency of occurrence of words from the corpus is used to reflect familiarity. This method can be an option in the development of word lists whereby the target language have little or no prior studies of word frequency and/or phoneme distribution, or is in need of a word corpus.

Ideally, a phonetically-balanced word list would contain the same distribution of phonemes found in normal speech. However, the ability to have the same distribution is limited by the quantity of words or test items included in the list. To improve the distribution, the list can be lengthened to allow more words and, therefore, more phonemes, to get the right percentage; however, raising the number of words also means increasing the test time, which is not desirable in speech audiometry. In this study, the problem is tackled by getting a balance between the distribution of phonemes and the length of list. Phonemes rarely used (less than 1% occurrence) in CVCV Malay words are excluded. This allows for shorter lists in the set, while keeping with the distribution of more common Malay phonemes in the lists.

A novel aspect of BMSA is the combination of meaningful and nonsense words in the lists. The selection of meaningful words was taken from a pool of frequently occurring words as found in the corpus. On the other hand, the nonsense words in BMSA are purposely constructed to fit the phonetic balance. They are also designed to 'sound' similar to true Malay words. To achieve this, the construction of words follows the phonetic rules on Malay CVCV words, such as vowel limitations for initial and final open syllables, as well as keeping check on the distribution of phonemes for the particular lists so that the final phoneme distribution matches the intended percentage.

The word lists was recorded digitally in a recording studio using a male voice and saved in a digital form. There is no carrier phrase used in the recording, and a 5-second interval is inserted between the words in each list. The input intensity, which is the intensity of the words in the recording, are digitally manipulated so that they are equal between words, and equal to the calibration tone. This is to minimise the possibility of having any lists being softer or louder than the others and jeopardise the equal difficulty between the lists.

The completed and recorded BMSA lists were then assessed for their homogeneity and equal difficulty.

6.2.2 Verification of the word lists

Verification of the word lists is important as it ensures homogeneity and equal difficulty among the word lists, therefore, allows the lists to be interchangeable. This ultimately improves the reliability of the wordlists and the measures obtained by using them.

There are three assessments that contribute to the verification of the BMSA word lists. Analysis of variance (ANOVA) on the scores obtained at a specified level is the most common method of verification of the homogeneity of the lists. Two new methods of verification are introduced in the study; one, internal consistency analysis using intraclass correlation coefficient, also known as Cronbach's alpha, and two, homogeneity of acoustic content.

Intraclass correlation coefficient, to date of writing, has never before described in homogeneity analysis for speech audiometry material. Both ANOVA and Cronbach's alpha in this study utilises the correct scores obtained at 15dB presentation level for a group of normal hearing listeners. While ANOVA is the more popular method in verification of speech audiometry word lists, it is thought that Cronbach's alpha provide a more accurate measure of homogeneity between the BMSA lists. Cronbach's alpha assesses the word lists based on the individual scores and how consistent an individual's performance throughout the set of lists, as opposed to ANOVA which compares the mean scores of the lists.

Another new method introduced in the study is the verification of the acoustic content of the BMSA word lists. The long term average speech spectrum (LTASS) is used as the reference point for the frequency spectra of the BMSA lists. Intensity comparisons at the various frequency points between the spectra of the lists and the LTASS showed that the acoustic content of the lists resembles that of average Malay speech. This further verifies that the BMSA word lists represent the frequency spectrum of Malay language.

Measurement of Cronbach's alpha and comparison of acoustic content between the lists and the LTASS provide additional or alternative methods of verification of word lists, which can be considered in future developments of speech audiometry material.

6.2.3 Clinical validation

The effectiveness of BMSA word lists in assessing speech recognition is validated through three analyses; evaluation of performance-intensity (P-I) function, evaluation of reference range, and assessment of sensitivity, specificity and predictive values.

The half peak level (HPL) and maximum speech recognition score (MSRS) calculated from the P-I functions of normal hearing and hearing impaired clients are found to

correlate with the pure tone hearing thresholds of the clients. This confirms that the BMSA word lists are able to reflect on the hearing for speech in both normal hearing and hearing impaired listeners. The standard deviation and reference range of the HPL are also comparable with the values of available speech audiometry material. These finding suggests that the BMSA word lists are clinically valid and can be included in hearing assessment battery.

The sensitivity, specificity and predictive values of BMSA word lists are high, suggesting that the word lists are reliable in both detecting hearing loss for as well as discriminating normal hearing and hearing loss.

The clinical validation of BMSA also suggests that the combination of meaningful and nonsense words in the test material has the ability to assess speech hearing in a manner that is comparable with previously published speech audiometry material that uses only meaningful words.

6.2.4 Clinical implication: Bisyllabic Malay Speech Audiometry test kit

Another theoretical implication of the current study is the clinical contribution through the Bisyllabic Malay Speech Audiometry (BMSA) word lists. The earlier findings of the current study, which are the word lists, the verification results and the clinical validation results are consolidated to produce a prototype Bisyllabic Malay Speech Audiometry test kit.

The test kit serves as a guideline for audiologists to perform speech reception threshold test in Malay using the BMSA word lists. The content of the kit consists of the rationale of performing speech reception threshold test, particularly in Malay; a summary on the process of developing the BMSA word lists; description of the BMSA word lists; instructions on how to administer and interpret the test; and a compact disk containing the audio file for the word lists. The booklet for BMSA can be found in Appendix I and the audio file is in the accompanying compact disk.

The BMSA is relatively easy to administer, considering that the test kit only requires room set up and equipment that are readily available in most audiology clinics and the method of presentation and interpretation is fairly straightforward. Therefore, the test should be accessible to most of the audiology clinics in Malaysia. It is hoped that with the production of the word lists, BMSA can be included in the routine hearing assessment battery in

Malaysia. The information acquired from the speech audiometry should aid and enhance the diagnosis and management of hearing loss in adult Malay speakers.

6.3 Limitations of research

A limitation of the study was the small sample size of the three participant groups. Although care was taken to ensure that the sample size reaches the targeted sample power, caution must be taken when inferring future test results of the normative data and reference range. However, the wide range of hearing loss covered in the current study should give an indication of what to be expected in the speech audiometry results. A more comprehensive normative data can be constructed with longer data collection period as well involvement of more, nationwide testing centres.

The combination of meaningful and nonsense words as test items in the all-words lists (AWL) allow for reduced contextual cues, therefore lessen the effect of guessing, in speech testing while providing comparable diagnostic values to other published speech recognition tests. A possible application for the AWL is in phoneme confusion studies, which was not part of the objective of the current study. Phoneme error matrix, which allow categorisation and identification of phoneme confusions, can be utilised in the study using the word lists developed in the current study.

6.4 Future research

More comprehensive diagnostic information will facilitate better interpretation of results and more accurate diagnosis. Exploration of the following areas will strengthen the understanding on the use of BMSA in assessing speech hearing:

i. P-I function of specific types of hearing losses: certain types of hearing losses, such as retrocochlear loss caused by eighth nerve disorder, generate specific patterns of P-I functions (Jerger, 1977). Qualitative and quantitative studies on the patterns of these types of hearing losses will provide more definitive understanding of the effect of the loss on the hearing of speech.
ii. The phonemic scoring of the BMSA allows for more in-depth analysis of the hearing of which speech sounds are affected by the hearing loss. A study of the patterns of affected speech sounds in specific types and configuration of hearing loss will aid in the management of hearing loss.

REFERENCES

Abdul Rahman, H. (1988). Dasar pendeskripsian system fonologi Bahasa Melayu. In Onn, F. (ed.)(1988) *Bunga Rampai Fonologi Bahasa Melayu*. Petaling Jaya: Fajar Bakti

Altman D.G. and Bland J.M. (1994). Statistics Notes: Diagnostic tests 2: predictive values. *British Medical Journal,* 309, p.102

Alusi, H., Hinchcliffe, R., Ingham, B., Knight, J.J. and North, C. et al. (1974) Arabic Speech Audiometry. *Audiology*, 13, pp.212–230.

American Speech-Language-Hearing Association. (1978) Guidelines for manual pure-tone threshold audiometry. *American Speech-Language- and Hearing Association*. pp.297-301

American Speech-Language-Hearing Association. (1988). Determining Threshold Level for Speech [WWW]. Available from www.asha.org/policy [accessed: 6/4/2014]

American Speech-Language-Hearing Association. (2011) Audiology Information Series: Type, degree and configuration of hearing loss [WWW]. Available from: http://www.asha.org/uploadedFiles/AIS-Hearing-Loss-Types-Degree-Configuration.pdf [accessed 28/01/2015]

American Speech-Language-Hearing Association. (1978) Hearing Loss: Beyond Early Childhood – Assessment [WWW]. *American Speech-Language- and Hearing Association*. Available from: http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935335&section=Assessment [accessed 1/3/2016]

Ashoor, A. A., & Prochazka, T. (1982). Saudi Arabic speech audiometry. *International Journal of Audiology*, 21(6), 493–508

Audit Bureau of Circulations (2011). *Circulation Figures: Newspapers* [WWW]. Audit Bureau of Circulations Malaysia. Available from: http://abcm.org.my/reports-archives/ [accessed 28/06/2011]

Bess, F.H. & Humes, L. (2003). *Audiology: The Fundamentals*. Baltimore: Lippincott Williams & Wilkins.

Bland, J.M. & Altman, D.G. (1997) Cronbach's alpha. *British Medical Journal*, 314, p.572.

Bochner, J., Garrison, W., & Palmer, L. (1986). A closed-set sentence protocol for assessing speech discrimination in deaf individuals: the speech sound pattern discrimination test. *Ear and hearing*, *7*(6), pp.370-376.

Boothroyd, A. (1968). Developments in speech audiometry. *British Journal of Audiology*, 2(1), pp.3–10.

Boothroyd, A. (2008). The Performance/Intensity Function: an underused resource. *Ear and hearing*, 29(4), pp.479–491.

Boyle, G.J. (1991) Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences,* 12(3), pp.291–294.

British Society of Audiology (2011) *Recommended Procedure bone-conduction threshold audiometry with and without masking* [WWW]. British Society of Audiology. Available at: http://www.thebsa.org.uk/docs/Guidelines/BSA_RP_PTA_FINAL_24Sept11.pdf. [Accessed: 17/1/2014]

Byrne, D. (1986) Effects of frequency response characteristics on speech discrimination and perceived intelligibility and pleasantness of speech for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 80(2), pp.494–504.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R. & Hagerman, B. (1994) An international comparison of long-term average speech spectra. *Journal of Acoustic Society of America*, 96(4), pp.2108–2120.

Carhart, R. (1951) Basic principle of speech audiometry. *Acta Oto-laryngologica*, 40, pp.62–71.

Chaiklin, J. B., & Ventry, I. M. (1964). Spondee threshold measurement: A comparison of 2-and 5-dB methods. *Journal of Speech and Hearing Disorders*, *29*(1), pp.47-59.

Cheesman, M.F. & Jamieson, D.G. (1996) Development, evaluation and scoring of a nonsense word test suitable for use with speakers of Canadian English. *Canadian Acoustics*, 24(1), pp.3–11.

Colorado Hands & Voices (2013) *The anatomy of the ear* [WWW]. Colorado Hands & Voices. Available from: http://www.cohandsandvoices.org/resources/coGuide/06_TheEar.htm. [Accessed: 30/8/2016]

Cornelisse, L.E., Gagnńe, J.P. & Seewald, R.C. (1991a) Ear level recordings of the long-term average spectrum of speech. *Ear and hearing*, 12(1), pp.47–54.

Cornelisse, L.E., Gagne, J.P. & Seewald, R.C. (1991b) Long-term Average Speech Spectrum at the chest-level microphone location. *Journal of Speech Language Pathology and Audiology*, 15(3), pp.7–12.

Cox, R.M. & Moore, J.N. (1988) Composite speech spectrum for hearing and gain prescriptions. *Journal of speech and hearing research*, 31(1), pp.102–107.

Department of Statistics Malaysia (2016). *Current population estimates, Malaysia, 2014-2016* [WWW]. Department of Statistics Malaysia. Available from: https://www.statistics.gov.my/index.php?r=column/cthemeByCat&cat=155&bul_id

=OWIxdEVoYlJCS0hUZzJyRUcvZEYxZz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09 [Accessed 24/8/2016]

diphthong. (n.d.). Collins English Dictionary - Complete & Unabridged 10th Edition [WWW]. Available from: http://dictionary.reference.com/browse/diphthong [Accessed 24/03/2011]

Dirks, D.D., Takayana, S. & Moshfegh, A. (2001) Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners. *Journal of the American Academy of Audiology*, 12(5), pp.233–244.

Egan J. (1948). Articulation testing methods. *Laryngoscope*, 58, pp.955-991.

Elango, S., Purohit, G.N., Hashim, M & Hilmi, R. (1991). Hearing loss and ear disorders in Malaysian schoolchildren. *International Journal of Paediatric Otorhinolaryngology*, 22, pp. 75-10

EllenBR (2010). *Speech banana and sound & way beyond* [WWW]. Cochlear Limited. Available from: http://www.cochlearcommunity.com/EllenBR/weblog/5471.html [Accessed: 28/1/2015]

FIRST YEARS (2006). Ling's Six-sound Test [WWW]. Available from http://firstyears.org/c1/u2/6soundchart.htm [Accessed 15/05/2011]

Frisina, R.D. (2009). Age related hearing loss. *Annals of New York Academy of Sciences*, 1170, pp.708–717.

Fromkin, V. & Rodman, R. (1998). *An Introduction to Language, 6th Ed.*. London:Harcourt Brace

Fu, Q.J., Zhu, M. & Wang, X. (2011). Development and validation of the Mandarin speech perception test. *Journal of the Acoustical Society of America*, 129(6), p.EL267-EL272.

Gelfand, S. (2009). *Essentials of Audiology* 3rd Ed., New York: Thieme.

Gelfand, S.A. (2010*). Hearing: An introduction to psychological and physiological acoustics: 5th Ed*. London: Informa Healthcare

Gelfand, S.A., Schwander, T., Levitt, H., Weiss, M., Silman, S. (1992). Speech recognition performance on a modified nonsense syllable test. *Journal of Rehabilitation Research and Development*, 29(1), pp.53–60.

GN Otometrics (2016). MADSEN Itera II - clinical diagnostic audiometer [WWW]. Available from: http://www.otometrics.com.au/Hearing-assessment/Audiometers/diagnostic-audiometer-madsen-itera [Accessed 4/9/2016]

Goodwin, L. & Leech, N. (2006). Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education*, 74 (3), 251-266

Google Maps (2016). *Map of Malaysia*. [WWW]. Google. Available from: https://www.google.com/maps/place/Malaysia/@4.0892925,100.570574,5z/data=!3m1!4b1!4m5!3m4!1s0x3034d3975f6730af:0x745969328211cd8!8m2!3d4.210484!4d101.975766 [Accessed 24 /7/2016].

Gramley, V. (2010). *IPA* [WWW]. Universität Bielefeld. Available from: http://www.uni-bielefeld.de/lili/personen/vgramley/teaching/HTHS/IPA.html [Accessed 13/3/2011]

Group, MrC. M.-C. O. M. S. (1999). Sensitivity, specificity and predictive value of tympanometry in predicting a hearing impairment in otitis media with effusion. *Clinical Otolaryngology & Allied Sciences*, *24*(4), 294-300.

Hall, J.W. & Mueller, H.G. (1996) *Audiologists's Desk reference Volume 1: Diagnostic Principles, procedures and protocols*. Singular: London

Han, D., Wang, S., Zhang, H., Chen, J., Jiang, W., Mannell, R., Newall, P. & Zhang, L. (2009) Development of Mandarin monosyllabic speech test materials in China. *International Journal of Audiology*, 48(5), pp.300–311.

Hannley, M. & Jerger, J. (1981) PB rollover and the acoustic reflex. *Audiology*, 20, pp.251–258.

Harris, R. W., Nissen, S. L., Pola, M. G., McPherson, D. L., Tavartkiladze, G. & Eggett, D. L. (2007). Psychometrically equivalent Russian speech audiometry materials by male and female talkers. *International journal of audiology*, 46(1), 47-66.

Hazan, V., Fourcin, A. & Abbeiton, E. (1991) Development of phonetic labeling in hearing-impaired children. *Ear and hearing*, 12(1), pp71–84.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders*, 17(3), 321-37.

Hirsh, I.J., Davis, H., Silverman, S.R., Reynolds, E.G., Eldert, E., Benson, R.W. (1952) Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders*, 17(3), pp.321–37.

Hood, J. & Poole, J. (1971) Speech audiometry in conductive and sensorineural hearing loss. *Sound*, 5, pp.30–38.

Hudgins C.V., Hawkins J.E., Karlin J.E. & Stevens S.S. (1947). The development of recorded auditory tests for measuring hearing loss for speech. *Laryngoscope*, 57 (1), 57-89

Humes, L.E. (2002). Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *The Journal of the Acoustical Society of America*, 112(3), p.1112.

Hyde, M. L., Riko, K., & Malizia, K. (1990). Audiometric accuracy of the click ABR in infants at risk for hearing loss. *Journal of American Academy of Audiology*, *1*(2), 59-66.

Hyman, L.M. (1975). *Phonology: theory and analysis*. New York: Holt, Rinehart and Winston

International Islamic University Malaysia. (2014). Audiology testing protocol. *International Islamic University Malaysia.* p.19

James, C., Bowsher J. and Simpson, P.J. (1991). Speech audiometry: digitization effects and the non-equivalence of isophonemic word lists, *British Journal of Audiology,* 25(2), 111-121.

Jensen Hearing (2016). *Jensen Hearing services brochure* [WWW]. Jensen Hearing. Available at: http://www.jensenhearing.com/category/hearing-aids/audiology_centres/

Jerger, J. & Hayes, D. (1977) Diagnostic speech audiometry. *Archives of Otolaryngology*, 103(4), pp.216-222.

Jerger, J., Speaks, C. & Trammell, J.L. (1968. A new approach to speech audiometry. *Journal of Speech and Hearing Disorders*, 33(4), p.318.

Kalikow, D.N., Stevens, K.N. & Elliott, L.L. (1977) Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of Acoustical Society of America*, 61(5), pp.1337-1351

Katz, J., Chasin, M., English, K. M., Hood, L. J., & Tillery, K. L. (2015).*Handbook of clinical audiology*, *7th Ed*. Philadephia: Wolters Kluwer

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

Ladefoged, P. (1982) *A Course in Phonetics*. San Diego: Harcourt Brace Jovanovich.

Lalkhen, A.G. & McCluskey, A. (2008) Clinical tests : sensitivity and specificity. *Continuing Education in Anesthesia, Critical Care & Pain*, 8(6), pp.221–223.

Lalkhen, A.G. and McCluskey, A. (2008). Clinical Tests: Sensitivity and Specificity. *Continuous Education in Anaesthesia, Critical Care and Pain*, 8 (6), pp. 221-223.

Lau, C.C. & So, K.W. (1988) Material for Cantonese speech audiometry constructed by appropriate phonetic principles. *British journal of audiology*, 22(4), pp.297–304.

Lehiste, I. & Peterson, G.E. (1959) Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America*, 31(3), pp.280–286.

Levitt, H. and Resnick, S.B. (1978) Speech reception by the hearing impaired: Methods of testing and development of materials. *Scandinavia Audiology Suppl*. 6, pp.107-109

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T. & Medler, D.A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, 15(10), pp.1621-1631.

Loh Guan Lye Specialist Centre (2016). *Audiology Unit* [WWW]. Loh Guan Lye Specialist Centre. Available at: http://www.lohguanlye.com/fa-audiology-unit.php

Lyregaard, P (1997). *Towards a theory in speech audiometry test*. In Speech audiometry 2nd Edition (Martin, M., ed). London: Whurr

Magnusson, L. (1995) Reliable clinical determination of Speech Recognition Scores using Swedish PB Words in speech-weighted noise. *Scandinavian audiology*, 24(4), pp.217–223.

Markides, A., 1979. The Effect of Content of Initial Instructions on the Speech Discrimination Scores of Hearing and Hearing-Impaired Children. *British Journal of Audiology*, 13, pp.113–117.

Martin, F.M. & Clark, J.G. (2010) *Introduction to Audiology, 10th Ed*. London:Pearson

Martin, M. (1997). *Speech Audiometry 2nd edition*. London: Whurr

McArdle, R., & Hnath-Chisolm, T. (2015) Speech audiometry.In Katz, J. et al, (eds.) *Handbook of clinical audiology*, 6th Ed., Philadelphia: Wolters Kluwer. pp. 64-79.

McGraw, K.O. & Wong, S.P. (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), pp.30–46.

Mendel, L. L., & Danhauer, J. L. (1997). *Audiologic evaluation and management and speech perception assessment*. San Diego: Singular.

Mendel, L.L. (2008) Current considerations in pediatric speech audiometry. *International journal of audiology*, 47(9), pp.546–53.

Miller, G.A., Heise, G.A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5), p.329.

Ministry of Health Malaysia (2007). *National Hearing and Ear Disorder Survey.* [WWW]. Malaysia's Health. Available from: http://www.moh.gov.my/images/gallery/publications/mh/Malaysia%20Health%2007-2.pdf [Accessed: 27/8/2016]

Moore, B.C.J. (2007) *Cochlear hearing loss: physiological, psyhcological and technical issues*. 2nd ed., Chichester: Wiley.

Mukari, S.M. & Said, H. (1991) The development of Malay speech audiometry. *Medical Journal of Malaysia*, 46(3), pp.262–268.

National Acoustics Laboratory. (2015) *Hearing assessment* [WWW]. National Acoustics Laboratory. Available from: http://www.nal.gov.au/hearing-assessment_tab_behavioural_testing.shtml [accessed: 17/07/2015]

National ORL Registry (2013). *The annual report National ORL Registry hearing and otology related disease/cochlear implant – Vol 1 (January 2010 – December 2011)*. [WWW]. National ORL Registry – Hearing and Otology Related Disease/Cochlear Implant. Available at: https://app.acrm.org.my/ORL [Accessed: 16/1/2016]

Neumann, K., Baumeister, N., Baumann, U., Sick, U., Euler, H. A., & Weißgerber, T. (2012) Speech audiometry in quiet with the Oldenburg Sentence Test for Children. *International journal of audiology*, *51*(3), 157-163.

Nielsen, J.B. & Dau, T. (2009) Development of a Danish speech intelligibility test. *International Journal of audiology*, 48(10), pp.729–41.

Niemeyer, W. (1965) Speech Audiometry with Phonetically Balanced Sentences. *International Journal of Audiology*, 4(2), pp.97–101.

Nilsson, M., Soli, S.D. & Sullivan, J. (1994) Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. The Journal of the Acoustical Society of America, 95(2), pp.1085–99.

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

Nissen, S.L., Harris, R.W. & Slade, K.B. (2007) Development of speech reception threshold materials for speakers of Taiwan Mandarin. *International journal of audiology*, 46(8), pp.449–58.

Nissen, S.L., Harris, R.W., Channell, R.W., Conklin, B., Kim, M. & Wong, L. (2011) The development of psychometrically equivalent Cantonese speech audiometry materials. *International journal of audiology*, 50(3), pp.191–201.

Nissen, S.L., Harris, R.W., Jennings, L.J., Eggett, D.L. & Buck, H. (2005) Psychometrically equivalent mandarin bisyllabic speech discrimination materials spoken by male and female talkers. *International Journal of Audiology*, 44(7), pp.379-390.

Nissen, S.L., Harris, R.W., Jennings, L.J., Eggett, D.L. & Buck, H. (2005b) Psychometrically equivalent trisyllabic words for speech reception threshold testing in Mandarin. *International Journal of Audiology*, 44(7), pp.391–399.

Noh, H. & Lee, D.H. (2012a) Cross-language identification of long-term average speech spectra in korean and english: toward a better understanding of the quantitative difference between two languages. *Ear and hearing*, 33(3), pp.441–3.

Noh, H., & Lee, D. H. (2012b). How does speaking clearly influence acoustic measures? A speech clarity study using long-term average speech spectra in Korean language. *Clinical and experimental otorhinolaryngology*, *5*(2), 68-73.

Onn, F. (ed.)(1988) *Bunga Rampai Fonologi Bahasa Melayu*. Petaling Jaya: Fajar Bakti

Ousey, J., Sheppard, S., Twomey, T. & Palmer, A.R. (1989) The IHR-McCormick Automated Toy Discrimination test-description and initial evaluation. *British Journal of Audiology*, 23(3), pp.245–249.

Palmer, A.R., Sheppard, S. & Marshall, D.H. (1991) Predictions of hearing thresholds in children using an automated toy discrimination test. *British Journal of Audiology*, 25, pp.351–356.

Perfect ENT Hearing and Speech Centre (2011). *Services: hearing consultation* [WWW]. Perfect ENT Hearing and Speech Centre. Available at: http://www.hearingaids.com.my/services1.html

Peters, R.W., Moore, B.C. & Baer, T. (1998) Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *Journal of the Acoustical Society of America*, 103(1), pp.577–87.

Peterson, G.E. & Lehiste, I. (1962) Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, 27(February), pp.62–70.

PHG Foundation (2015). *Clinical validity* [WWW]. PHG Foundation. Available at: http://www.phgfoundation.org/tutorials/acce/3.html

Prasansuk, S. (2000) Incidence/prevalence of sensorineural hearing impairment in Thailand and Southeast Asia. *Audiology*, 39(4), pp.207–211.

Radford, A., Atkinson, M., Britain, D., Clahsen, H. & Spencer, A. (1999). *Linguistics: an introduction*. Cambridge: Cambridge University Press

Ramkissonn, I., Proctor, A., Lansing, C.R. & Bilger, R. (2002) Digit Speech Recognition Thresholds (SRT) for Non-Native Speakers of English. *American Journal of Audiology*, 11(June), pp.23–28.

Robinson, J (n.d.) *Received Pronunciation* (WWW) British Library. Available from: http://www.bl.uk/learning/langlit/sounds/find-out-more/received-pronunciation/ [Accessed 16/3/2011]

Runge, C.A. & Hosford-Dunn, H. (1985) Word recognition performance with modified CID W-22 word lists. *Journal of speech and hearing research*, 28(3), pp.355–362.

Sabo, M. P., Winston, R., & Macias, J. D. (2000). Comparison of pure tone and transient otoacoustic emissions screening in a grade school population. *Otology & Neurotology*, *21*(1), 88-91.

Scollie, S.D. (2008). Children's speech recognition scores: the Speech Intelligibility Index and proficiency factors for age and hearing level. *Ear and hearing*, 29(4), pp.543–556.

Shapiro, I. (1976) Hearing aid fitting by prescription. *Audiology*, 15, p.163.

Stelmachowicz, P.G., Pittman, A.L., Hoover, B.M., Lewis, D.E. & Moeller, M. (2004) The Importance of High-Frequency Audibility in the Speech and Language

Development of Children With Hearing Loss. *Archives of Otolaryngology, Head and Neck Surgery,*130(5):556-562.

Stevens, G., Flaxman, S., Brunskill, E., Mascarenhas, M., Mathers, C.D. & Finucane, M. (2013) Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries. European journal of public health, 23(1), pp.146–152.

Tab a Doctor (2016). *Sunway Medical Centre - Speech & Hearing Centre - Medical Services* [WWW]. Tab a Doctor. Available at: https://www.tabadoctor.com/sunway-medical-centre/speech-hearing-centre/medical-services

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), pp. 96-116.

Tavakol, M. & Dennick, R. (2011) Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, pp.53–55.

Teoh, B.S.. (1994). *The sound system of Malay revisited*. Kuala Lumpur: Dewan Bahasa dan Pustaka

Tillman, T.W. & Carhart, R. (1966) *An expanded test for speech discrimination utilizing CNC monosyllabic words. Northwestern University Auditory Test No. 6.* [Technical report] SAM-TR. USAF School of Aerospace Medicine, (6), pp.1–12.

Townsend, T. & Bess, F., (1980). Effects of age and sensorineural hearing loss on word recognition. *Scandinavian Audiology*, 9, pp.245–248.

Universiti Malaya Medical Centre (2014). *Department of Otorhinolaryngology: Introduction* [WWW]. Universiti Malaya Medical Centre. Available at: http://www.ummc.edu.my/department/department.asp?kodjabatan=20

Varošanec - Škarić, G. (2003) Voice Assessment Before and After Phonetic Voice and Pronunciation Exercises. In *Proceedings of the 15th International Congress of Phonetic Sciences*. pp. 2153–2156.

Ventry, I.M. (1979) Effects of conductive hearing loss: fact or fiction. *Journal of Speech and Hearing Disorders*, 45(2), pp.143–157.

von Hapsburg, D., Champlin, C. & Shetty, S.R. (2004). Reception thresholds for sentences in bilingual (Spanish/English) and monolingual (English) listeners. *Journal of the American Academy of Audiology*, 15(1), pp.88–98.

Wang, S., Mannell, R., Newall, P., Zhang, H., & Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *International Journal of Audiology*, 46(12), 719-31

Webster, J.C. (1972). *Compedium of speech testing material and typical noise spectra for use in evaluateing communications equipment.* Naval Electronics Laboratory Center

Weißgerber, T., Baumann, U., Brand, T. & Neumann, K. (2012) German Oldenburg Sentence Test for Children: A Useful speech audiometry tool for hearing-impaired children at kindergarten and school age. *Folia Phoniatrica et Logopaedia*, 64, pp.227–233.

Wilson, R.H. et al. (1976) Northwestern University Auditory Test No. 6: Normative and Comparative Intelligibility Functions. *Journal of the American Audiology Society*, 1(5), pp.221–228.

World Health Organization (2015) *Deafness and hearing loss* [WWW]. World Health Organization. Available from: http://www.who.int/mediacentre/factsheets/fs300/en/ [accessed: 26/09/2015]

Yacullo, W. (1999) Clinical Masking in Speech Audiometry: A Simplified Approach. *American Journal of Audiology*, 8, pp.1–12.

Yiap, K.H., (1984). Disyllabic Malay word lists for speech audiometry. *Medical journal of Malaysia*, 39(3), pp.197–204.

Zakrzewski, A., Jassem, W., Pruszewicz, A., & Obrebowski, A. (1975). Identification and discrimination of speech sounds in monosyllabic meaningful words and nonsense words by children. *Audiology*, 14, 21-26.

Zakrzewski, A., Jassem, W., Pruszewicz, A., & Obrebowski, A. (1976). Speech Audiometry for Children and Subjective Probability of Polish Words. *Audiology,* 15, 228-231.

Zatorre, R.J., Evans, A.C., Meyer E. & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*(5058), p.846-849.

Zatorre, R.J., Belin, P. and Penhune, V.B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), pp.37-46.

APPENDIX I

Bisyllabic Malay Speech Audiometry Word List

Test Kit

(Prototype)

Chapter 1 Introduction and rationale

Speech audiometry is an established assessment for the purpose of quantifying an individual's ability to hear speech (American Speech-Language-Hearing Association, 1988). There are several forms of speech audiometry, which can be categorised based on the material used, method of administration and/or target population.

Due to the speech element of the test, material constructed based on the native language of the target population is preferred. Studies have shown that speech tests conducted not in the listener's native language negatively affect the results (Hapsburg and Pena, 2002; Ramkissoon, et al., 2002; Hapsburg, et al., 2004).

The bisyllabic Malay speech audiometry (BMSA) word lists offers a validated speech audiometry material to assess the speech perception ability, particularly the threshold of intelligibility of speech and quantification of speech discrimination ability, of Malay speakers. It is designed to be used for normal hearing and hearing-impaired adults ages 18 and above. BMSA has a unique characteristic of the inclusion of nonsense words within the test items in each list in order to more closely reflect everyday speech. The set may also be converted to the shorter all-meaningful word lists to allow for quicker assessment time.

1.1 The need for speech recognition test

Wilson and Margolis (1983) and Boothroyd (2008) outlined several rationales for measuring speech recognition. They include:

i. Quantification of hearing sensitivity
   Communicative problems are a major concern in hearing impairment. Speech recognition test provides a valid measurement of speech sensitivity and, in extension, the communication process

ii. Pure tone threshold verification
   Speech reception thresholds provide a means to verify pure tone audiometry results, for example in cases of non-organic hearing loss or poor pure tone audiometry technique

iii. Estimate of auditory resolution

iv. Reliability for difficult-to-test patient

Speech recognition test may provide additional information regarding the hearing acuity in cases when the client responds poorly to pure tones

1.2 Speech perception test developed in Malay language

With the start of audiology services in Malaysia and Singapore, several speech audiometry material targeted for Malay-speaking population have been published since. There are two sets of word lists that are aimed to assess speech hearing thresholds and suprathreshold intelligibility by means of maximum speech recognition score. They are:

- *Disyllabic Malay word lists for speech audiometry (Hong, Y.K., 1984)*
  This set is one of the earliest Malay speech audiometry material found in literature. It is aimed to assess speech hearing threshold and suprathreshold intelligibility. The set is made up of ten lists with ten bisyllabic words each. The set was designed to be phonetically balanced based on everyday spoken Malay. Scoring was based on syllables with 5% scores for each syllable. Homogeneity and test-retest reliability was established. However, normative speech hearing threshold and its relationship with pure tone audiometry was not identified.

- *Malay speech audiometry word lists (Mukari and Said, 1991)*
  This set of word lists is also designed to assess speech hearing threshold and speech discrimination. The lists are phonemically balanced, consonant-vowel-consonant-vowel bisyllabic Malay words. There are 25 lists with 10 words in each list. The scoring is based on phonemic scoring, with each phoneme carrying a score of 2.5%. The interlist intelligibility difference, normal discrimination curve and mean normal speech threshold was established. However, the relationship between the speech intelligibility measures (threshold and discrimination score) and pure tone audiometry were not established. Repeated usage of words in some of the list also raises a question on the possibility of memory and/or practice effect.

There are also other speech audiometry materials available in Malay language. These materials, however, carry different purposes, rather than assessing speech hearing threshold and discrimination score, and may employ different methods of assessment:

- Single and double dichotic digit tests in Malay (Mukari, et al., 2006)
- Malay Hearing-in-noise test (Quar, et al., 2008)
- Malay Speech Intelligibility Test (MSIT) for Deaf Malaysian Children (Yusof, et al., 2013)

Chapter 2 Development of the bisyllabic Malay speech audiometry (BMSA) word lists

2.1 **Determination of test structure**

In the review of literature and existing speech audiometry material available worldwide, speech audiometry is identified as one of the basic and routine tests included in the audiometric test battery, especially in adults. Much of the materials of established speech recognition tests, such as CNC lists, AB word lists and spondaic word lists, are based on phonemes and words as the test items (Lehiste and Peterson, 1959; Boothroyd, 1968; American Speech-Language-Hearing Association, 1988). A review of material for speech recognition test in other languages also revealed that most of words used are of single- or two-syllable words (Ashoor and Prochazka, 1982; Lau and So, 1988; Harris, et al., 2007; Wang, et al., 2007; Caldwell, 2009; Han, et al., 2009). After a preliminary study of the distribution of Malay words, bisyllabic Malay words were chosen over single syllable words due to their wider selection of words. Malay monosyllables were also found to be mostly colloquial and/or abbreviation of words. The test is also structured to have phonemic scoring, with each phoneme carrying a score, as it increases the test items and reduces the test score variability (Boothroyd, ibid.).

2.3 **Item construction**

A literature search on Malay word corpus revealed no data available on Malay spoken or written word inventory or their frequency of occurrence. Therefore, for the purpose of item construction, a word corpus was constructed using a program provided by Dewan Bahasa and Pustaka based on words that appeared in Malay main daily newspapers, Utusan Malaysia and Berita Harian, together with their Sunday editions, over a period of 5 years (2006-2010). The words were ranked according to frequency of occurrence and the top 350 words were shortlisted for the test items. Words that were proper nouns and culturally or religiously inappropriate were excluded.

The distribution of phonemes in CVCV Malay words were also calculated based on the word corpus. This distribution provides the basis of phonetic balance in the word lists. However, not all phonemes in Malay language were able to be included in the lists due to the limitation posed by the number of words in each list. Vowels ‘a’, ‘I’, ‘e’, ‘ə’, ‘u’ and ‘o’, and consonants ’b’, ‘c’, ‘d’, ‘f’, ‘g’, ‘h’, ‘j’, ‘k’, ‘l’, ‘m’, ‘n’, ‘p’, ‘r’, ‘s’, ‘t’, ‘w’ and ‘y’. All these phonemes occur more than 1% in the corpus. Words in the final 15 lists were carefully chosen (from the shortlisted 350) to simulate this phonetic balance. Nonsense words were then added to the lists, while retaining to the phonetic balance.

2.4 **Verification – internal consistency and homogeneity**

The homogeneity of the word lists were verified through three methods – Friedman Test, internal consistency analysis using Cronbach’s alpha and long-term average speech spectrum (LTASS).

The consistency of speech recognition using the word lists were analysed using Friedman test due to the non-normal distribution of some of the word list scores. There was statistically no significant difference in correct scores achieved using any of the word lists, X^2 (14) = 19.584, p= 0.144. This result showed that the choice of word list used in testing had no effect on the speech audiometry scores in normal hearing participants, and therefore, interchangeable.

A test of homogeneity using intraclass correlation coefficient (Cronbach's alpha) was applied to the scores. The ICC value of 0.78 indicates strong internal consistency among the 15 lists, suggesting that all of the lists are homogenous and have equal difficulty.

Repeated measures ANOVA on the frequency spectra of the lists with Malay LTASS revealed no significant difference between the lists and LTASS, F=1.229, p>0.05. This indicates consistency between the frequency spectra of the lists and the LTASS and suggests that the lists reflect the average frequency spectrum of Malay language.

2.5 Clinical validation

Correlation analysis between the half peak level (HPL) of the speech recognition curve and pure tone threshold (PT) average shows that, in consideration of both normal hearing and hearing impaired listeners, the HPL correlates best with the PT average of 250, 500, 1000, 2000 and 4000 Hz for both AWL (r = 0.667 to 0.951) and MWL (r = 0.649 to 0.946).

Table 2.1 Non-parametric correlation of PT results vs SRT in AWL

		Spearman's rho		
Subject group	**PT average**	**r_S**	**p**	**N**
Normal hearing	0.5,1& 2 kHz	0.61	.001	25
	0.5,1& 4 kHz	0.62	.001	25
	1, 2 & 4 kHz	0.54	.006	25
	0.5,1, 2 & 4 kHz	0.60	.002	25
	0.25, 0.5, 1 & 2 kHz*	0.67	.000	25
	0.25, 0.5, 1, 2 & 4 kHz	0.67	.000	25
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.50	.013	25
SNHL	0.5,1& 2 kHz	0.95	.000	16
	0.5,1& 4 kHz	0.93	.000	16
	1, 2 & 4 kHz	0.81	.000	16
	0.5,1, 2 & 4 kHz	0.93	.000	16
	0.25, 0.5, 1 & 2 kHz	0.94	.000	16
	0.25, 0.5, 1, 2 & 4 kHz*	0.95	.000	16
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.90	.000	16
CHL	0.5,1& 2 kHz	0.85	.000	14
	0.5,1& 4 kHz*	0.90	.000	14
	1, 2 & 4 kHz	0.83	.000	14
	0.5,1, 2 & 4 kHz	0.87	.000	14
	0.25, 0.5, 1 & 2 kHz	0.84	.000	14
	0.25, 0.5, 1, 2 & 4 kHz	0.85	.000	14
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.84	.000	14

*Notes the pure tone combination with the strongest correlation

Table 2.2 Non-parametric correlation of PTA results vs SRT with MWL

Participant group	PTA	Spearman's rho		
		r_S	p	N
Normal hearing	0.5,1& 2 kHz	0.584	.002	25
	0.5,1& 4 kHz	0.607	.001	25
	1, 2 & 4 kHz	0.517	.008	25
	0.5,1, 2 & 4 kHz	0.581	.002	25
	0.25, 0.5, 1 & 2 kHz	0.646	.000	25
	0.25, 0.5, 1, 2 & 4 kHz*	0.649	.000	25
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.475	.016	25
SNHL	0.5,1& 2 kHz	0.945	.000	16
	0.5,1& 4 kHz	0.929	.000	16
	1, 2 & 4 kHz	0.805	.000	16
	0.5,1, 2 & 4 kHz	0.920	.000	16
	0.25, 0.5, 1 & 2 kHz	0.937	.000	16
	0.25, 0.5, 1, 2 & 4 kHz*	0.946	.000	16
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.898	.000	16
CHL	0.5,1& 2 kHz	0.835	.000	14
	0.5,1& 4 kHz	0.829	.000	14
	1, 2 & 4 kHz	0.804	.001	14
	0.5,1, 2 & 4 kHz*	0.855	.000	14
	0.25, 0.5, 1 & 2 kHz	0.830	.000	14
	0.25, 0.5, 1, 2 & 4 kHz	0.841	.000	14
	0.25, 0.5, 1, 2, 4 & 8 kHz	0.813	.000	14

A comparison between HPL and PT average of 250, 500, 1000, 2000 and 4000 Hz showed mean differences of 3.67 dB (SD = 3.37) and 3.25 dB (SD = 3.74) for AWL and MWL respectively (Table 2.3). The standard deviation for the normal hearing groups can be used to estimate the range of tolerance for speech audiometry using BMSA; ±6.74 (with 95% confidence) dB for AWL and ±7.84 dB for MWL. This range of tolerance is comparable to the range applied to pure tone audiometry, which is ±5 dB.

Table 2.3 Mean HPL and mean HPL-PT average difference for normal hearing clients, clients with sensorineural hearing loss and clients with conductive hearing loss using AWL and MWL

		AWL			MWL		
		Normal	SNHL	CHL	Normal	SNHL	CHL
HPL	Mean (dB dial)	10	34	41	10	34	40
	SD	3	18.11	11.87	3	18	12
HPL – PT Difference	Mean	4	-7	1	3	-7	0
	SD	3	6	7	4	6	7

Predictive analyses suggest that speech recognition test using BMSA word lists has excellent sensitivity and specificity and very high predictive values (Table 2.4). To achieve this, separate correction factors are given to normal hearing and conductive hearing impaired listeners, and

listeners with sensorineural loss. The tolerance range are also set at ±10 dB as it improves sensitivity and specificity. MWL provides better sensitivity and specificity towards hearing loss compared to AWL.

Table 2.4 Predictive values for HPL-PT average agreement separate correction factors for SNHL and normal hearing/CHL

		Variable CF applied					
		AWL (CF= 4 for normal hearing and CHL, CF=-7 for SNHL)			MWL (CF= 3 for normal hearing and CHL, CF=-7 for SNHL)		
		General	SNHL	CHL	General	SNHL	CHL
CI ±10	sensitivity	0.90	0.94	0.86	0.93	0.94	0.93
	specificity	1.00	1.00	1.00	1.00	1.00	1.00
	PPV	1.00	1.00	1.00	1.00	1.00	1.00
	NPV	0.89	0.96	0.93	0.93	0.96	0.96

General: normal hearing vs all types of HL; SNHL: normal hearing vs SNHL; CHL: normal hearing vs CHL; CF: correction factor; CI: accuracy/confidence interval with 95% confidence level

Chapter 3 Description of BMSA

BMSA is a pre-recorded open-set speech audiometry material for the purpose of assessing speech reception threshold (SRT) and speech discrimination. The test items are phonetically-balanced based on the Malay phoneme distribution and consonant-vowel-consonant-vowel (CVCV) in structure. There are two types of words in BMSA – meaningful words and nonsense words. The nonsense words, although not carrying any meaning in standard Malay language, were constructed according to Malay phonetic rules and, therefore, may resemble meaningful words.

There is no carrier phrase in BMSA. An interval of 5 seconds is given between each word to allow the client time to respond.

There are two test formats available in BMSA – the all-word lists (AWL), integrating meaningful and nonsense words in each list, and the meaningful-word lists (MWL), consisting of only meaningful words in each list. Both formats contain 15 lists each, with 15 words (10 meaningful words, 5 nonsense words) in each AWL list and 10 words (all meaningful) in each MWL list.

BMSA employs phonemic scoring; each phoneme carries one mark.

Chapter 4 Test administration and scoring procedures

The following are the recommended procedure for test administration and scoring of BMSA.

4.1 Testers

The BMSA is designed to be used by audiology practitioners such as audiologists, audiology technicians and audiology students, and other professionals assessing speech reception. Testers should have formal clinical training in performing hearing assessments.

4.2 Test environment, equipment and setting

BMSA is designed to be performed in a sound-treated audiometric booth or room. The specifications for audiometric booths or rooms can be found in several published standards such as ANSI 3.1-1999 or BS EN ISO 8253-1:1998.

An audiometer, transducers and a CD player are required in the administration of the BMSA. The audiometer should be calibrated and meet the requirements relevant for diagnostic audiometers. The CD player should be compatible with the audiometer in channelling external speech stimuli.

Test can be done in either single room (tester and client in a same room) or two room (tester and client in separate rooms) settings. Double room setting would require extra equipment to allow the tester to hear the responses.

4.3 Test administration

4.3.1 Test format and scoring forms

Two test formats are available in BMSA – all word lists (AWL) which consists of 10 meaningful and 5 nonsense words per list, or meaningful-word lists (MWL) which contains 10 meaningful words per list. The forms contain the test items for each format with the words in each list arranged according to the order of presentation (Appendix 1).

4.3.2 Randomisation

All lists in BDSA have been tested and verified for homogeneity and equivalence in terms of difficulty. Therefore, the lists are interchangeable and should produce the same results irrespective of their order of presentation.

The 'random' presentation setting on the CD player, if available, can be used to randomise the order of presentation of the list.

4.3.3 Procedure and scoring

BMSA is designed to measure speech reception threshold (SRT). SRT is defined as the level at which the listener correctly identifies 50% of the test items (ISVR, 2003). SRT is sometimes used interchangeably with half-peak level (HPL) or half-optimum speech reception threshold level (HOSRTL), defined as the speech hearing level at which half of the maximum speech recognition score is obtained (ISVR, ibid.). The word lists can also be used in the assessment of

suprathreshold speech perception by measuring the maximum speech recognition score (MSRS), the highest correct score obtained by the client on the speech recognition curve.

The recommended procedure for measuring SRT using BDSA adopts the decreasing method suggested by Chaiklin and Ventry (1964) and Boothroyd (1968).

1. Equipment set up

Set up the audiometer, CD and CD recorder for speech audiometry according to the instructions in the audiometer manual. Use the 1 kHz calibration tone included in the CD to set the output on the VU meter to 0.

2. Instructions

Instructions should be given in a manner that is suitable for the client. Information regarding the nature of the test, structure of the stimuli and mode of response must be relayed to the client. Client should also understand that the stimuli may be presented at very faint levels and he/she is encouraged to guess.

An example of instruction to the client:

"Anda akan mendengar beberapa patah perkataan. Perkataan-perkataan tersebut mungkin berbunyi kuat ataupun perlahan. Sila ulang setiap perkataan selepas anda mendengarnya, tidak kira samada ianya membawa maksud ataupun tidak. Anda akan diberikan masa untuk mengulang selepas setiap perkataan dibunyikan. Anda juga digalakkan untuk meneka."

(*You will be presented with several words. The words may be loud or soft. Please repeat the word you hear after it is presented, no matter if it carries any meaning or not. You will be given time to repeat the word after it is presented. You are also encouraged to make a guess*.)

3. Determination of initial presentation level

The initial presentation level should be at a level well above the client's speech hearing threshold. To estimate the speech hearing threshold and calculate the initial presentation level:

i. Calculate the average pure tone hearing thresholds for 250, 500, 1000, 2000 and 4000 Hz.
ii. Add 30 dB to the pure tone threshold average (PT average). This value is the starting level or initial presentation level.

4. Familiarisation

i. Select the format of BMSA to be utilised. Select the lists to be used in the test and include one list for familiarisation.
ii. Present the stimulus from the familiarisation list at the initial presentation level (PT average + 30dB). If the response is correct, continue presenting 2 to 3 more stimuli to ensure that the initial presentation level is well above the speech threshold. If the response is incorrect, increase the presentation level in 10-dB increments until a correct response is obtained. This level is the starting level.

5. Plotting the performance-intensity (P-I) function

i. Identify on the speech audiometry score sheet the lists that are going to be used in the session.

ii. Note the presentation levels for each list. Present one list at the selected initial presentation level/starting level. Most listeners will get a full score (100% correct) at this level. Mark the responses on the score sheet:

Stimulus	Response
P A D I	PADI
P A D ɨ	PADU
~~P~~ A ~~D~~ ɨ	AGA
~~P A D ɨ~~	KEJU

iii. Continue presenting the following lists at 10 dB decrements (starting level-10, starting level-20, starting level-30 etc.).

iv. <u>If the speech recognition test is to measure speech reception threshold only</u>, the test is terminated when the correct score is equal to or less than 30% (42 or more incorrect phonemes in AWL, 28 or more incorrect phonemes in MWL) and the highest score is between 95 to 100%.

<u>If the test is to measure maximum speech recognition score (MSRS) in addition to speech reception threshold, or the highest score is less than 95%,</u> continue decreasing the presentation level in 10 dB decrements, until the correct score is equal to or less than 30% (42 or more incorrect phonemes in AWL, 28 or more incorrect phonemes in MWL).

To find the MSRS and/or seek for the presence of rollover, if the client scored 100% (or close to 100%) at the starting level (5ii), present a new list at 20 dB above the starting level (starting level + 20) and record the responses. Present the list at 10 dB above starting level if +20 dB is uncomfortably loud for the client.

If the client did not score 100% correct score at the starting level, present a list at 20 dB above starting level (starting level + 20) and record the responses. Increase the presentation level by 20 dB (starting level + 40), or by 10 dB (starting level + 30) if the 20 dB increment is uncomfortably loud for the client, and present a new list.

6. Scoring

This test employs phonemic scoring. Each correctly repeated phoneme is given a score of one (1) while incorrect phonemes are given 0. The maximum score for each test item (word) is 4.

Example:

Stimulus	Response	Score
P A D I	PADI	4
P A D ɨ	PADU	3
~~P~~ A ~~D~~ ɨ	AGA	1
~~P A D ɨ~~	KEJU	0

The maximum score per list is 60 for AWL and 40 per MWL.

Plot the results on the speech audiogram.

7. Calculation of half peak level (HPL), maximum speech recognition score (MSRS) and rollover index

Calculation of HPL and MSRS are done in reference to the speech audiogram.

To calculate HPL

i. Identify the highest correct score along the speech audiogram, i.e. the peak of the speech audiogram curve. This is the maximum speech recognition score (MSRS)
ii. Divide the score by two. This gives the half peak score
iii. Identify the presentation level corresponding to the score. This is the half peak level

To calculate the rollover index

i. Identify the MSRS and its presentation level
ii. Identify the lowest correct score obtained at a level higher than the MSRS level. This is PBmin
iii. Calculate the rollover index:

$$\text{Rollover index} = (\text{MSRS} - \text{PBmin})/\text{PBmin}$$

Chapter 5 Interpretation of BMSA

The client's speech recognition can be assessed based on two approaches – the normative data and the calculation of HPL-PT average agreement.
Rollover index (RI) is used to differentiate cochlear and auditory nerve (VIII cranial nerve) dysfunction. The RI value for BMSA has yet been established; however, a summary of published rollover index values based on other speech audiometry material has been compiled by Mueller and Hall (1995).

1. Normative data

The following tables (Tables 5.1 and 5.2) are the normative data for the percentage of correct scores at various presentation levels for AWL and MWL, based on 25 normal hearing adults. Comparisons on correct scores based on the normative data can be made using the reference range (with 95% confidence interval).

Table 5.1 Mean correct scores for normal hearing participants using bisyllabic Malay speech audiometry, AWL

Presentation level (dB dial)	Correct score (%)	
	Mean	Range (95% confidence interval)
-5	0	0
0	0.6	0 – 5.40
5	14.6	0 – 49.24
10	47.4	1.03 – 93.77
15	80.3	56.24 – 100
20	92.9	84.07 – 100
25	96.8	91.36 – 100
30	99.1	96.69 – 100
35	99.1	96.75 – 100
40	98.5	94.63 - 100

Table 5.2 Mean correct scores for normal hearing participants using bisyllabic Malay speech audiometry, MWL

Presentation level (dB dial)	Correct score (%)	
	Mean	Range (95% confidence interval)
-5	0	0
0	0.4	0 - 3.17
5	15.1	0 - 54.27
10	49.7	0 - 100
15	83.9	61.3 - 100
20	94.7	84.36 - 100
25	97.7	91.59 - 100
30	99.5	97.46 - 100
35	99.6	97.24 - 100
40	99	95.77 - 100

2. Half peak level – pure tone threshold average agreement

To determine the HPL-PT average agreement e.g. for the purpose of validating pure tone results or estimating pure tone thresholds in hard-to-test clients, the following formula can be used:

AWL Format

For normal hearing and conductive hearing loss

HPL = [(0.25, 0.5, 1, 2 & 4 kHz PT average) + 4] ± 10 dB

For sensorineural hearing loss

HPL = [(0.25, 0.5, 1, 2 & 4 kHz PT average) - 7] ± 10 dB
with

MWL Format

For normal hearing and conductive hearing loss

HPL = [(0.25, 0.5, 1, 2 & 4 kHz PT average) + 3] ± 10 dB

For sensorineural hearing loss

HPL = [(0.25, 0.5, 1, 2 & 4 kHz PT average) - 7] ± 10 dB

MALAY BISYLLABIC SPEECH AUDIOMETRY SCORE SHEET (AWL)

LIST 1	SCORE	LIST 2	SCORE	LIST 3	SCORE	LIST 4	SCORE	LIST 5	SCORE	LIST 6	SCORE	LIST 7	SCORE
LEVEL		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**	
LAGI		JANI		HATI		DUSI		SANA		RUDA		HAWA	
BAHU		TALI		BINA		REDA		SEMA		RELA		PATI	
RATU		CUBA		TUJU		GEPA		SURI		MOJE		BENI	
DABI		BEKU		RAHI		HOBI		BERI		KOPI		DEPU	
SIGU		GULA		FERI		GURU		KALI		DAYA		LALU	
NASI		SATU		WIRA		LALI		MERI		FASA		MUDA	
MEJA		KIRA		SEPU		RULI		BATU		BELI		BIRO	
KACA		HILA		BEKI		PENA		HOKI		BULA		SAYA	
TEPI		DARI		KAMU		KETI		JADI		SERI		KAJI	
MONI		BOGA		CELA		CABA		KATU		KAMI		BUKA	
SUDI		SEDI		SAPI		SITU		RADU		HAJI		SURA	
BARU		SEPI		DOSA		MASA		TEMU		TAGI		TISU	
RUGA		TURU		LAKU		JATI		PAYA		CATU		KIMI	
KELI		MANA		LUMA		CURI		BUTI		KOSI		TUMA	
KETU		LOJI		DOTA		KALA		LAGU		GUNA		RAGA	
TOTAL		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**	

Notes:

_____

_____

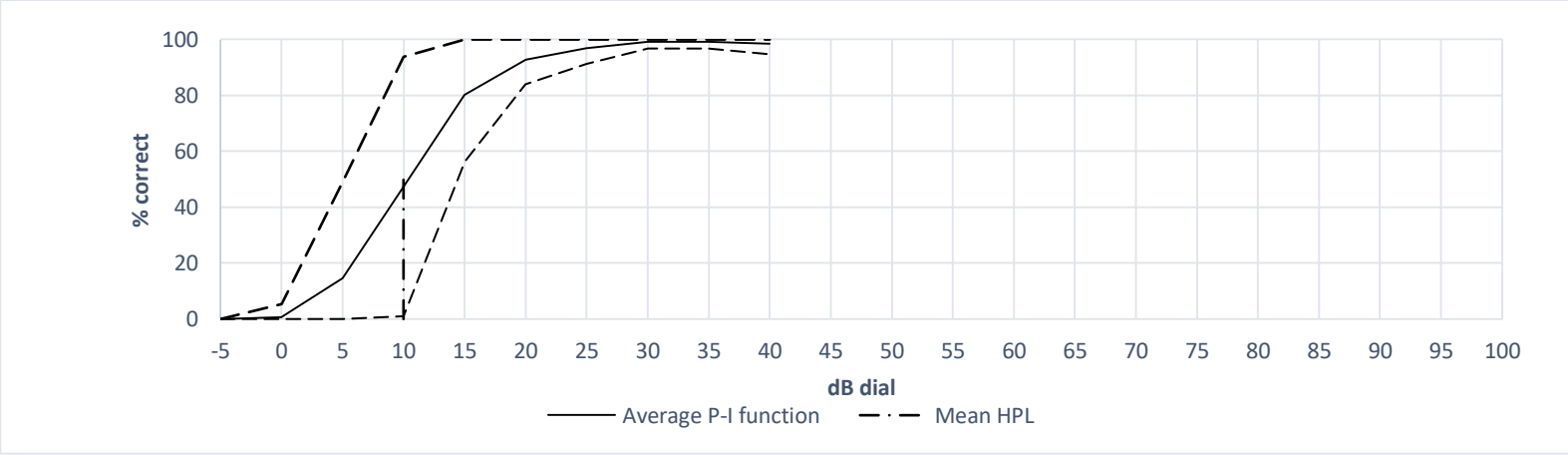
_____

_____

_____

LIST 8	SCORE	LIST 9	SCORE	LIST 10	SCORE	LIST 11	SCORE	LIST 12	SCORE	LIST 13	SCORE	LIST 14	SCORE	LIST 15	SCORE
LEVEL		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**	
BOMA		PADI		TANI		LORI		BAJA		KENA		KADI		TIBA	
JIWA		BESI		TEGA		SIFU		PAJI		GARI		SUPA		NILA	
KASA		KEJU		KUBU		SETI		TIPU		LUKA		RUJI		PADA	
SUHU		CUTI		MAKA		RIDA		LUCU		SOYA		SILA		TERI	
BACA		LITA		LOYA		DULU		TUKI		WARI		KERA		KOLE	
DUPI		TARI		SEGI		MOKE		MIGA		MAHA		MALU		WAJA	
PETI		JATU		PILU		SATE		DESA		SEBA		DANA		MIKU	
MUTU		SEGA		BIRU		BILU		GUNI		TEPU		BIMI		MAHU	
KARI		BOYA		SUMI		BUMI		KOMA		DUTA		TORI		SARA	
GURA		GALA		ROKI		PECU		SAWI		KUMU		TOPI		SIKU	
LIGA		LESU		DATA		NAGA		ROBA		NOJI		RAJA		CUTI	
CABU		NERU		LUBA		JARI		LARI		JELI		CETI		SONA	
NADA		MUKA		PEDU		KATA		KITA		PAGA		DOBA		RUGI	
SUKI		HARI		JASA		BAPA		BARA		DURI		SAGU		LIDA	
LOBI		KOBI		CARI		KAYA		SENU		MISI		WALU		KABA	
TOTAL		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**	

Speech audiogram



MALAY BISYLLABIC SPEECH AUDIOMETRY SCORE SHEET (MWL)

Name:……

ID:……

Date:……

Tester:……..

LIST 1	SCORE	LIST 2	SCORE	LIST 3	SCORE	LIST 4	SCORE	LIST 5	SCORE	LIST 6	SCORE	LIST 7	SCORE
LEVEL		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**	
LAGI		TALI		HATI		REDA		SANA		RELA		HAWA	
BAHU		CUBA		BINA		HOBI		SURI		KOPI		PATI	
RATU		BEKU		TUJU		GURU		BERI		DAYA		LALU	
NASI		GULA		FERI		LALI		KALI		FASA		MUDA	
MEJA		SATU		WIRA		PENA		BATU		BELI		BIRO	
KACA		KIRA		KAMU		SITU		HOKI		SERI		SAYA	
TEPI		DARI		CELA		MASA		JADI		KAMI		KAJI	
SUDI		SEPI		SAPI		JATI		TEMU		HAJI		BUKA	
BARU		MANA		DOSA		CURI		PAYA		CATU		TISU	
KELI		LOJI		LAKU		KALA		LAGU		GUNA		RAGA	
TOTAL		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**	

PT Average:

Freq (Hz)	HTL (dBHL)
250	
500	
1000	
2000	
4000	
Total	
PT ave.	

Notes:

_____

_____

_____

_____

_____

LIST 8	SCORE	LIST 9	SCORE	LIST 10	SCORE	LIST 11	SCORE	LIST 12	SCORE	LIST 13	SCORE	LIST 14	SCORE	LIST 15	SCORE
LEVEL		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**		**LEVEL**	
JIWA		PADI		TANI		LORI		BAJA		KENA		KADI		TIBA	
KASA		BESI		KUBU		SIFU		TIPU		GARI		RUJI		NILA	
SUHU		KEJU		MAKA		DULU		LUCU		LUKA		SILA		PADA	
BACA		CUTI		LOYA		SATE		DESA		SOYA		KERA		KOLE	
PETI		TARI		SEGI		BUMI		GUNI		MAHA		MALU		WAJA	
MUTU		BOYA		PILU		NAGA		KOMA		TEPU		DANA		MAHU	
KARI		GALA		BIRU		JARI		SAWI		DUTA		TOPI		SARA	
LIGA		LESU		DATA		KATA		LARI		JELI		RAJA		SIKU	
NADA		MUKA		JASA		BAPA		KITA		DURI		CETI		CUTI	
LOBI		HARI		CARI		KAYA		BARA		MISI		SAGU		RUGI	
TOTAL		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**		**TOTAL**	

Speech audiogram