A Study on Reusing Resources of Speech Synthesis for Closely-Related Languages

by

Nur Hana Samsudin

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
October 2017

“It does not matter how slowly you go as long as you do not stop."

Confucius

Thank you, mom

Abstract

A Study on Reusing Resources of Speech Synthesis for Closely-Related Languages

by Nur Hana Samsudin

This thesis describes research on building a text-to-speech (TTS) framework that can accommodate the lack of linguistic information of under-resource languages by using existing resources from another language. It describes the adaptation process required when such limited resource is used. The main natural languages involved in this research are Malay and Iban language. Malay represents a language with sufficient speech resources while Iban language represents a language with very limited resources. Overall thesis revolves around the two languages.

The thesis includes a study on grapheme to phoneme mapping and the substitution of phonemes. A set of substitution matrices will be presented which show the phoneme confusion in term of perception among respondents. The experiments conducted study the intelligibility as well as perception based on context of utterances.

The study on the phonetic prosody is then presented and compared to the Klatt duration model. This is to find the similarities of cross language duration model if one exists. Then a comparative study of Iban native speaker with an Iban polyglot TTS (with Malay as focal language) is presented. This is to confirm that the prosody, suprasegmental or the rhythm of Malay can be used to generate Iban synthesised speech.

The thesis concludes with the description of Iban polyglot TTS criteria with very minimal data using a very closely related language: Malay, as the main resource. The study is concluded with the respondents ratings and feedback.

The central hypothesis of this thesis is that by using a closely-related language resource, a natural sounding speech can be produced. The aim of this research was to shaow that by sticking to the indigenous language characteristics, it is possible to build a polyglot synthesised speech system even with insufficient speech resources.

Acknowledgements

Thank you for my dear supervisor, Mark Lee who has the disadvantage of having a PhD student like me and for sticking together until the end. I am truly grateful for the help he has given me as a PhD supervisor as well as a reliable friend to a foreign land without any clue as to how to survive. Thank you for your patience with my 'broken English writing skills'. Thank you for not getting angry with me when being kept pushing and asking. Thank you for trying to your best capabilities to help me to come out with this revised thesis. Thank you so very much.

This thesis would also not have been possible without the tremendous help from Dr Tan Tien Ping of the School of Computer Sciences, Universiti Sains Malaysia, USM. I am forever grateful for his patience, his guidance, his generous data and equipments sharing as well as his powerful technical skills when I needed it the most.

Thank you to the University of Birmingham specifically the School of Computer Science for providing the necessities, the equipment and the tremendous support throughout and post me being there.

This thesis would be too detached if not for the wonderful help I received from the whole Sarawak Language and Technology research group (SaLT) of Universiti Malaysia Sarawak, UNIMAS, especially to Sarah Samson Juan and Suhaila Saee. The team compassion and kindness as well as the blooming friendship will be forever cherished.

I am genuinely grateful to the faceless warriors of open academic resources which make searching for knowledge less painful even when I was put in a position to be outside of academic environment.

Thank you for my sponsor, the Ministry of Higher Education and the people of Malaysia. I am deeply thankful and grateful to be given the privilege. Special thank you to the ex-Dean, of the school of Computer Science, Prof Rosni Abdullah for being the boss, a friend and a supporter, despite having her hands full at all time.

Thank you to my beloved friends, in Birmingham, USM and those from neither, for being superbly supportive, for the compassion, for the unconditional friendship, for being the shoulders to cry on, for the constructive idea and for everything you all do. Really. Thank you. Thank you. Thank you. Life is less suffocating because of you all.

I sincerely, full-heartedly would like to thank my siblings for trying their best not to involve me too much with the crucial events as well as the never ending drama in the extended family after the passing of two prominent figures in our family.

This thesis will not be as what it is now if not for the constructive ideas, comments and tremendous guidance by both of my examiners. I am not exaggerate to say this thesis is not even half the quality it is now without them. Thank you to my internal examiner: Dr Peter Hancox and my external examiner: Prof Dr Nick Campbell for the inspiring and motivating hours in the viva room. I will forever be grateful for all their advice.

And I sincerely apologise if I missed anyone else. But thank you, is all I can say. Thank you very much.

Contents

List of Figures

List of Tables

Abbreviations

CBR	**C**ase **B**ased **R**easoning
CHATR	**ATR** Generic Speech Synthesis System
CV	**C**onsonant-**V**owel
CVC	**C**onsonant-**V**owel-**C**onsonant
HMM	**H**idden **M**arkov **M**odel
HTK	**H**idden Markov Model **T**ool-**K**it
HTS	**H**idden Markov Model for **T**ext-to-**S**peech Synthesis
ML	**M**ulti-**L**ingual
NLP	**N**atural **L**anguage **P**rocessing
RMSE	**R**oot **M**eans **S**quare **E**rror
SAPI	**S**peech **A**pplication **P**rogramming **I**nterface
SNR	**S**peech to **N**oise **R**atio
TTS	**T**ext-**t**o-**S**peech
VC	**V**owel-**C**onsonant

Chapter 1

Introduction

Speech processing has evolved to be a useful Natural Language Processing (NLP) application technology. However, the rate of progress differs from language to language. One of the main reasons for this imbalance is the level of maturity of NLP research done in certain languages for example, decades of research have been conducted in English, Spanish, French and German, but very little has been done for many under-developed languages.

A key issue in developing speech technology for any minority language is the scarcity of available resources. This research proposes to make use of available speech framework as guidelines to build polyglot speech synthesisers for insufficiently resourced languages using resources of a focal language. The source language and the target language being considered are isolect language. In other words, dialects will have high cognate and thus theoretically should be easily adaptable with the current state of the art in Text-to-Speech (TTS) technology. What this thesis looking at is, how when the language is not a dialect of another, but sufficiently closely related, a manipulation of speech data of the focal language could be used to create the target language synthesiser.

The research aims to study a mechanism that can be used as the basis of the creation of polyglot speech synthesis systems with very minimal changes done to the original framework despite having limited speech resources. The focus of the thesis is to provide a comparative evaluation and implementation of how such a situation can be achived that would correspond to the linguistic features of the focal language and the perceptual

features of speech. Using this information, speech can be synthesised for languages with limited resources.

1.1 Short Introduction to Speech Synthesis

Speech synthesis or text-to-speech refers to the production of speech from text. However in most advanced applications the text is not visible to the user, but is rather a part of the speech communication component, as shown in the speech circle in Figure 1.1.

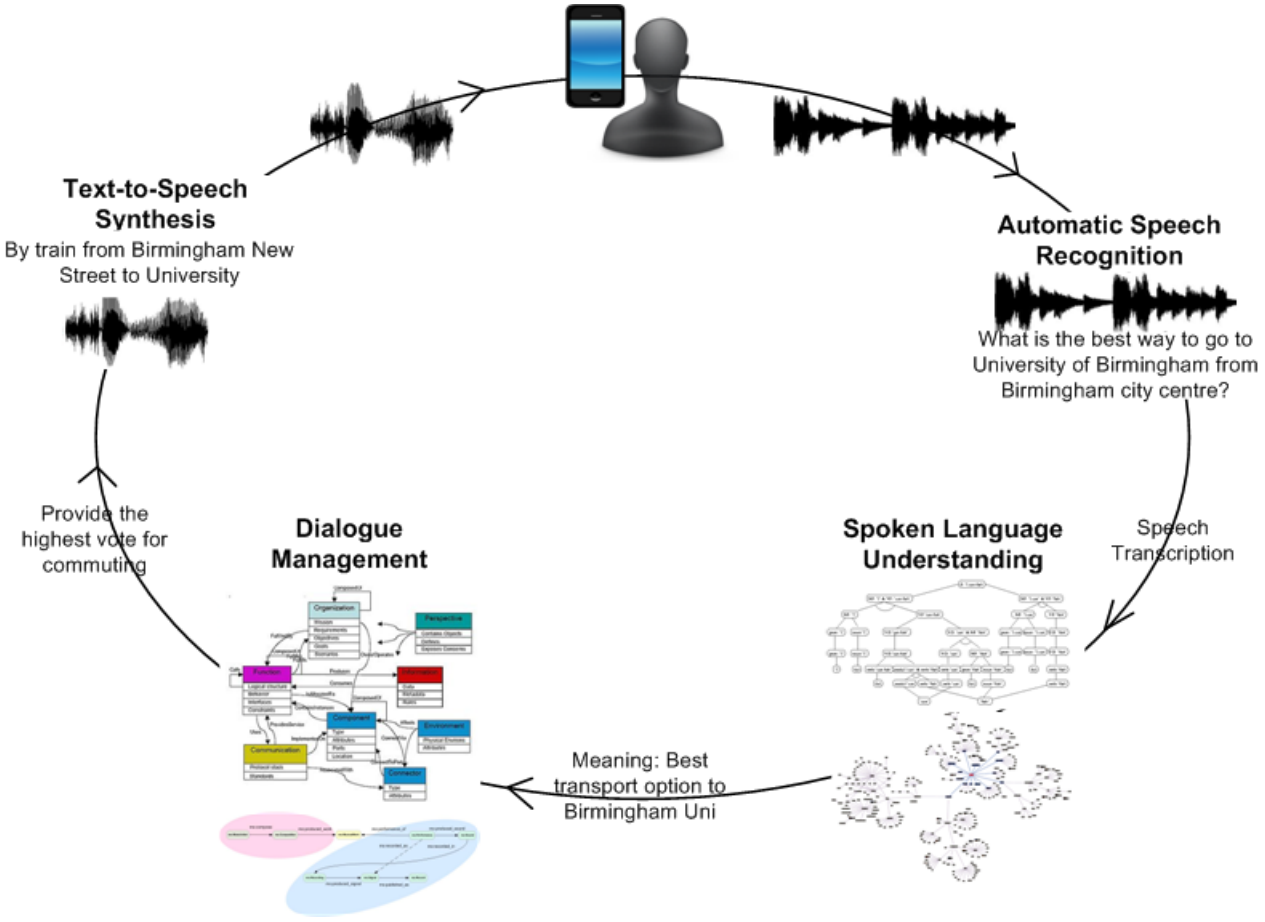

FIGURE 1.1: Text-to-speech synthesis in a speech circle.

The figure shows the speech synthesiser generating the answer requested. In a call service system, further information may be required and therefore the text-to-speech synthesis would generate a more detailed question instead of an answer.

A very general architecture of a TTS is shown in Figure 1.2. In the text analysis stage, the document structure detection identifies the beginning and end of the text/sentence construct (sentence, list, email format etc.) as well as paragraph structure. Then text

normalisation converts the non-orthographic text into graphemes. Linguistic analysis or syntactic and semantic parsing produces structural and semantic information about the sentences. These text analysis components can be used for many purposes other than TTS, including information retrieval, machine translation and text summarisation.

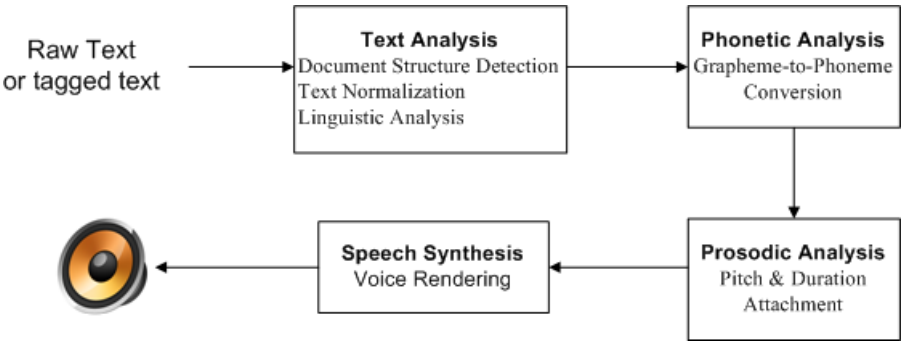

FIGURE 1.2: Basic system architecture of a TTS system(Huang et al., 2001).

1.1.1 History of Multilingual and Polyglot Speech Synthesis

The earliest multilingual speech synthesis system recorded is by Carlson and Granström, 1975. They use a synthesis by rules program for multiple languages: Swedish, Norwegian, American English, British English, Spanish, French, German and Italian. The main contribution of Carlson and Granström (1975) however is a special programming language to permit linguists to formulate synthesis rules which are then used for synthesis into speech. This language has the ability to refer to natural sets of phonemes through distinctive feature notation, making rule statements simple and easy to read.

Further studies that have evolved from the study of human articulation movement and the vocal tract transfer take into account phonetic science as well as the suprasegmental effect of phonetic sequence. (Heinz and Stevens, 1961) state that fricative consonants involve the generation of turbulence noise at a constriction in the vocal tract. The noise primarily excites the formants associated with the cavities in front of the constriction (Fant, 1960; Stevens,1972). Acoustic properties that distinguish the English fricatives from one another include the general spectral shape of the frication noise and the motions of the formants transitioning to the next sound (Klatt, 1987). Most of the formant transitions take place while aspiration is the sound source. The burst is slightly longer and more intense, and formant transitions are somewhat less distinct in voiceless plosives, making the burst a more potent cue to place an articulation.

In term of adjacent context on speech production, (Liberman et al., 1967) emphasised the encoded nature of speech, where the acoustic cues to identify phonemes were spread out in time so as to overlap with context-dependent cues of adjacent phonemes (Liberman et al., 1967). This is also supported by Cooper et al., 1952 that says the same plosive burst spectrum is heard as a different consonant depending on the vowel pattern that follows. In Klatt and Klatt, 1990, an analysis of reiterant imitations of nonsense sentences also showed the effect of arytenoids movement in producing a similar sound with different speech quality (prosody). The research however, mainly shows the difference of acoustic cues for different type of voice quality variations between different genders.

This information, although studied in a language specific environment, contributed to identifying the general features of the phonetics (and thus inclusive of phonology) and is very important in determining the core character of the speech that can be reused in different languages.

An early widely available TTS toolkit that have been developed would be Festival Speech Syntesis System (Taylor et al., 1998) using FestVox (Festival Voice) and MBROLA. The Festival Speech Synthesis System is a multilingual speech synthesis system developed by Alan W. Black at Centre for Speech Technology Research (CSTR) at the University of Edinburgh. The Festvox project on the other hands is a suite of tools by Alan W. Black and Kevin Lenzo for building synthetic voices for Festival. FestVox's aims to make the building of new synthetic voices more systematic and better documented. The goal is to make it possible for anyone to build a new voice. Both are free licensed tools. Festival offers a full text to speech system with various APIs, as well as an environment for development and research of speech synthesis techniques. Festival is also designed to support multiple languages, and comes with support for English (British and American pronunciation), Welsh, and Spanish. Voice packages exist for several other languages, such as Castilian Spanish, Czech, Finnish, Hindi, Italian, Marathi, Polish, Russian and Telugu.

MBROLA provides a free speech synthesiser engine with broad language corpora. MBROLA focus is on its synthesiser engine which can only be use with its own formatted voice database. It uses a diphone concatenation approach. The MBROLA project aims to create a multilingual speech synthesis by obtaining a set of speech synthesiser with as

many languages as possible and then provides them free for non-commercial applications. The recordings were provided freely by researchers around the world. The TTS engine is not dependent to one language. In fact, one can use any language database one would like to and can still uses the MBROLA engine. Up to now, data for 75 languages has been collected. Compared to Festival, MBROLA synthesiser is more rigid by design since it uses the patented TD-PSOLA algortihtm for it's diphone concatenation.

Contrary to MBROLA, Festival provides more flexibility and is still an evolving TTS toolset. Initially providing diphone concatenation synthesiser, the tool has progressed into a corpus-based and parametric speech synthesis approach. However, despite being very flexible, the Festival tool is not easily modified to tailor to one needs; it requires deep understanding of the tool's components as well as determination before it can be fully utilised.

CHATR is a generic speech synthesis system developed at the ATR, Japan. Similar to Festival, CHATR is designed in a modular way so that module parameters and the module to be used may be set and selected at a runtime. CHATR offers a useful research tool in which functionally equivalent modules may be easily compared. Similar to Microsoft's Speech Application Programming Interface (SAPI), it also can act as a simple system for those less interested in the specific details of speech synthesis but wish their computer to talk (Black and Taylor, 1994).

The latest and widely used multilingual speech synthesis approach is the HMM-based speech synthesis. The first HMM-based TTS were applied to English and appeared in 2002 (Tokuda et al., 2002). The HMM-based speech synthesiser has then widely been used by different researches to improved their TTS. Up to the moment this thesis is written, HMM-based approach is the most preferred multilingual TTS tool and still being improved.

1.2 Research Motivations and Objectives

The central hypothesis of this research is that it is possible to produce good quality speech synthesis for languages lacking extensive resources by using similar linguistic information, including data from another language. The aim of this research is to construct a TTS using related usable linguistic information and with optimal coverage of the phonemes of the target language. This method can then be used to adapt the

existing speech synthesis framework to the target language even with insufficient target language speech resources. Using the suggested representation, it is also possible to make use of available resources to create a resource-poor language TTS.

It The thesis will test several hypotheses. These are:

- How much acceptance of substituted phonemes can a listener bear in order for the speech to be deemed comprehensible?
- Can the context of the text help in overcoming the missing phoneme by providing a closer sound?
- If the language does not have a specific stress or accent pattern, how close is it to fit into Klatt's duration model which was originally tested for American and British English?
- How to test if one language is close to another based on the speech produced? Would it be sufficient and conclusive?
- Is it possible to create a synthesiser using a very limited target language resource or no target language resource at all?

1.3 Thesis Statement

This research aims to reuse a TTS framework to speed up the production of a polyglot speech synthesis systems as a generalisation approach for languages with limited speech resources together with other existing language resources. This approach focused on closely-related languages and representing linguistic information of the source language which is closely corresponds to the linguistic features of the target language and the perceptual features of the speech.

1.4 Terms Used

The following terms will be used extensively throughout the thesis. This is still following the standard definition but is directed to the thesis' scope.

1. Multilingual vs polyglot

   A **multilingual speaker** grow up using more than two languages while they are

still in language acquisition years. While prosody is obtained during the infant years (Höhle et al., 2009; Saffran et al., 2001), the speech learning is progressive (Saffran et al., 2001). A multilingual speaker is an individual who has learnt the second language sufficiently or already a bilingual by the age of seven (Clark, 2000). A **polyglot speaker** is one that is already sufficiently fluent in one language before learning a second or the subsequent languages. Therefore, the speaker uses the first language as the point of reference for mastering the subsequent language(s). For a TTS, a **multilingual system** will have different algorithms, rules and speech data for different languages while a **polyglot system**, has a primary language which is the focal language of the synthesiser (Traber et al., 1999). The other languages will use this primary TTS language as the core of the system while having the freedom to add data, rules or information processing.

2. Resource language vs target language

   This study on reusing resources of speech synthesis for closely related language revolves around using the available language resources to be used or modified to create different language synthesisers. In such cases, the language from which the data originates is referred to as the **resource language**, while the language the system can generate is referred to as the **target language**. This thesis may also interchangeably refer to focal language and synthesised language respectively depending on the context.

3. Phoneme vs phone

   In orthographic form, a word consists of a set of letters. These letters represent a particular sound of the specific language. When two sounds can be used to differentiate words, they are said to belong to different **phonemes**. Therefore one can say that a phonetic representation of a word constitutes a sequence of phonemes. Each phoneme usually corresponds to a phone. A **phone** is an instance of a phoneme in an actual utterance. It is a speech sound which could be of any sound produced by the human vocal tract which is found as part of the speech production. Phones have more variation than letters do. An **Allophone** is a speech sound viewed from the perspective of its membership of a phoneme. The allophones of a phoneme form a set of sounds that (1) do not change the meaning

TABLE 1.1: The transformation from ortographic to phonemic to phonetic transcription

Orthographic	Phonemic Transcription	Phonological Rules	Phonetic Transcription
controlling	kəntrəʊliŋ	schwa deletion between plosive and nasal	kntrəʊliŋ
covering	kʌvəriŋ	schwa deletion before /r/	kʌvriŋ
English	ŋgliʃ	deletion of /g/ after /ŋ/	ŋliʃ

of a word, (2) are all very similar to one another, and (3) occur in phonetic contexts different from one another (Ladefoged and Johnson, 2010).

For example, the words <pin> and <spin> both have a phoneme [p] which is phonemically similar. However they are from different pairs of allophones where the [p] in <pin> is aspirated, /p^h/ while <spin> is not. Similar to the allophones of [p] in the word <pop>, there are phonetic variations that cannot be used to distinguish words (Ladefoged and Johnson, 2010). The variation is there because of the phonetic context.

4. Phonetics and phonology

   The production of speech from text is influenced by these two linguistics features. **Phonetics** is the study of speech sounds and their production, classification and transcription (Huang et al., 2001) while **phonology** is the area of linguistics that describes a systematic way that sounds are differently realised in different environments (Jurafsky and Martin, 2008).

   The **phonetic representation** is written in the International Phonetic Alphabet (IPA) giving a symbolic representation of phones(Jurafsky and Martin, 2008). IPA is an evolving standard with the goal of transcribing the sounds of all human language (Jurafsky and Martin, 2008).

   An example illustrating the effect of phonetics and the influence of phonology in pronunciation are presented in the word transcription in Table 1.1. To transcribe from orthographic to phonemic transcription, the corresponding phoneme of the letters in the word are represented. The actual phone used to produced each sound has been processed phonetically.

   This research focused on two languages which use a **phonetic spelling system**: Malay and Iban. A phonetic spelling system is a system of spelling in which each letter represents invariably the same spoken sound (Ladefoged and Johnson, 2010). Languages like English or French have different phonemic transcriptions

from written text due to the changes made to pronunciation over the centuries, while spelling has remained basically the same (Divay and Vitale, 1997; Ladefoged and Johnson, 2010).

5. Standard Language of Malay vs Standard Malay

   The Malay language evolved from 18th century were originated from Johore-Riau which then becoming the Standard Malay. In the late 19th century, Standard Language of Malay is introduced to form a better uniformity of written and spoken language. The usage of formal language were changed back to Standard Malay around the year 1998. Further elaboration can be obtained in Section 2.3.1 page 20.

6. Native vs Non-Native

   The nativeness of the respondents plays an important role especially when their native language is more varied than the language they are evaluating for a TTS experiment. For example, if the respondents first language is a tonal language, they would prefer the produced synthesised speech that consist of tonal quality despite the target language has a rather loose tonal rules or none at all. This is a frequent case for Malay respondents who are not native. To ensure there is no bias in evaluation, most studies conducted involved native speakers instead of L2 or L3 speakers.

1.5 Thesis Organisation

This thesis is organised as follows. The next chapter will describe the literature reviewed of the two main languages being studied this thesis: Malay and Iban. It will also describe the language typology and the divergent of the stock. Chapter 3 presents other research carried out on multilingual and polyglot speech synthesis, as well as the approach taken to handle multilingual/polyglot research. Chapter 4 outlines grapheme-to-phoneme conversion approaches used in the existing multilingual or polyglot research as well as monolingual TTS and then focuses on the grapheme-to-phoneme for Malay and its adaptation into Iban. Chapter 5 describes prosody assignment and manipulation for different TTS approaches and the implementation of the prosody assignment for Malay and Iban. Chapter 6 provides the in-depth review of the Iban polyglot TTS adaptation. Finally, Chapter 7 discusses the outcomes of the research and its evaluation in relation to the hypotheses.

1.6 Summary

This chapter provided a general overview of text-to-speech systems architecture, introducing the terms which will be used throughout the thesis and the objective and hypotheses of this research. The next chapter will provide an introduction to the two main languages in this thesis: Malay and Iban.

Chapter 2

Language and Melody in Speech

This chapter will give a brief review on language charcteristics before diving into the more intangible aspect of speech: melody. Speech melody reflects the rhythm in speech as how it is being looked at in this thesis. The thesis is about constructing a synthesiser using another language resource altogether and this would require compatibility over some common ground. The introduction to the focal language, Malay, and the language with under-resource language, Iban, language will be laid out.

2.1 Language Characterisation

The language characteristics play an important indicator to see whether two different languages share common characteristics that may be an important aspect in speech production. It is common that most languages share more than half of their phoneme inventory with one another, however would that make two languages close in characteristics in any way?

In linguistics, there are two language classifications in practice: historical (genealogical) and typological. The purpose of genealogical classification is to group languages to their relatedness. A typological classification can groups languages into types according to their structural characteristics. In principle, there is no limit to the variety of ways in which languages can be grouped typologically. One can distinguish languages between rich or poor phonemic inventories. One can divide the morphology, prefixing languages against suffixing languages and so on.

In developing speech technology, knowledge of which languages are typologically related can be helpful, and in this research, historical relatedness and genealogical characteristics are especially important for Malay-Iban relatedness.

2.1.1 Language Classification and Characteristics

Language characteristics help to determine the criteria that may be influential in producing a natural-sounding speech synthesiser. In multilingual and polyglot speech synthesis, it is important to identify the differences and similarities in the language criteria and how these can be represented in the parameters for a multilingual/polyglot TTS. Kirchhoff (2006) highlights the following criteria which may be used to define language characteristics in general:

1. *Linguistic Description and Classification*

   - *Language families*

     Language families are categorized with regard to historical and geographical groupings. For instance, Indo-European is the world's largest family in terms of number of speakers (Kirchhoff, 2006), but the characteristics of its member languages are very different. For instance, Spanish implements a phonetic spelling system while French does not.

   - *Language typology*

     Language typology refers to the classification of languages based on their structural characteristics. From a linguistic point of view, there are different aspects of typology:
     a) Phonetics, Phonology and Prosody
     b) Morphology
     c) Word Order

     In the main two languages studied in this thesis, both Malay and Iban have a rather loose word order. However, both are very similar except for lexical rearrangement and most sentences are close in sentence structure. Malay on the other hand has a more complex morphology than Iban. This is due to the adaptability of Malay in terms of loan words.

     While Iban mostly maintains and creates new words when necessary, Iban and Malay share very close phonetic distribution. This may lead to similar

phonology and typology. However, Iban is a stressed language, while Malay is not. Having said that, in Jako Iban,the speaker can easily mix Malay into Iban language when the word is not yet known like code-switching situations among Switzerland speakers.

2. *Language in context*

   Languages differ in the way they are used in actual communication. Divergence from standard pronunciation is influenced by dialects, idiolects and sociolects. For example, in a country like Switzerland, code-switching occurs very frequently from German to the other primary languages: French and Italian (Romsdorfer and Pfister, 2004). This criteria may be taken into consideration in multilingual/polyglot TTS systems.

   The Iban language is known as Jaku' Iban which means Conversational Ihttps://preview.overleaf.co Therefore, the language setting for Iban is more informal than formal. The use of Standard Malay recording, which is a formal language, may create a gap between the two.

3. *Writing systems*

   Kirchhoff (2006) classifies writing systems into certain categories and arranges them in a hierarchy as shown in Figure 2.1.

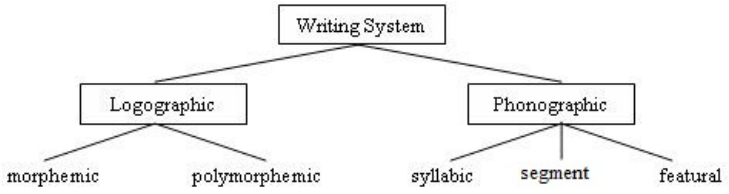

FIGURE 2.1: Classification of writing systems (Kirchhoff, 2006).

The logographic system involves graphemes which represent a word or a morpheme (a meaningful unit of language); an example of a language in this category is Chinese that has a character system, Hanzi (Kirchhoff, 2006). A morphemic system has a one-to-one correspondence between units of meaning and graphemes, while polymorphemic systems may have several units of meaning combined in one grapheme.

In an orthographic phonography system, graphemes represent sound units. In a syllabic system the graphemes stand for entire syllables. Japanese Kanji is an example of a syllabic system. In a segmental system, a grapheme roughly corresponds to a phoneme, while featural writing system uses elements smaller than the phone to correspond to phonetic or articulatory features. Most western alphabets (e.g. Greek, Roman, Cyrillic), semitic languages (e.g. Arabic, Hebrew, Akkadian), and non-Semitic languages (e.g. Farsi, Urdu and Hausa) are segmental. Korean is an example of a featural system.

As for the studied language, both Malay and Iban use romanised letters and therefore a segmental phonographic writing systems.

Most of these features may be important to categorise a language but they may not all be essential to the construction of a multilingual or polyglot speech synthesiser. While it is important to know the typology of the languages in question for example the set of phonemes, the formation of words and the syntax, the language family, context and writing system are likely to be of limited value.

This research is about manipulating data for a closely related language. As such, the information of the language background plays a very important role to ensuring that the best possible synthesised speech can be achieved. Therefore, these features: language families, language typology, language in context as well as the writing system may play a huge role in the construction of a TTS of this nature.

2.2 Melody in Speech

Music and speech belong to the same ontological root - sounds. Several studies have shown that language learning, acquisition, understanding and music therapy for dyslexic students can speed up the oral or speaking process (Mora, 2000; Schwantes, 2009; Gilbert, 2012; Fisher, 2001; Colwell and Murlless, 2002; Schunk, 1999; Kennedy and Scott, 2005). In fact, the melodic properties of speech actually may be the first information one grabs before one can speak the intended language (Odam, 1995; Crystal, 1986; Schwantes, 2009). This section will highlight these language features without going into the discrete features of speech prosody, or to the supra-segmental effect of the speech. Rather, this chapter redefines the term melody or rhythm of speech for different languages.

Linguistically, the prosody refers to the study of rhythm, intonation, stress and related speech attributes. Nooteboom (1997) defines prosody as properties of speech that cannot be derived from the segmental sequence of phonemes underlying human utterances. In speech research, prosody has always been associated with three perceptual features: length, pitch and loudness. These are most often depicted in numerical values that are represented by the duration, fundamental frequency and amplitude. On the perceptual level, these properties are the most important ones to perceived patterns of relative syllable prominences, coded in perceived melodical and rhythmical aspects of speech.

From a phonetic point of view, human speech is more than a characterized manifestation of sequences of phonemes, syllables or words. In normal speech, pitch, duration and the loudness of the speech fluctuate in some controlled non-random method which creates a pattern of melody. Some segments are produced to sound more prominent than others, to convey the intention of the sentences uttered. The prosodic features are not affected by normal orthographic or the conventional phonetic transcription.

The manner of speaking creates other properties of the speech sound. Properties of speech that accompany rather than form part of the consecutive segments of a word or a sentence are often called suprasegmental properties of speech (Nooteboom, 1997). This may also be perceived as timbre. For example, the speakers may speak softly or loudly or just normally. The speaker could also use a hoarse, breathy voice or have a baritone voice quality. The articulation may be produced carefully or slurringly, etc.

Typical prosodic features of speech are not reflected in normal orthography or in conventional segmental phonetic transcription.

Intonation in its strict interpretation is “the ensemble of pitch variations in the course of an utterance” (Nooteboom, 1997). This interpretation of intonation concentrates on those pitch variations that are related to perceived speech melodies, and thereby pay less attention to pitch variations that are related to the segmental structure of speech.

The early studies conducted by Hart et al. (1990), Fujisaki and Sudo (1971), Maeda (1976), O’Shaughnessy (1976), and Pierrehumbert (1980) and others tried to come out with a prosody model for the structure of intonation in terms of the actual course for their respective studies. What these approaches to prosody patterning have in common is that they strive for some kind of stylized approximation of the apparently unpredictable pitch

fluctuations that are found in natural speech, hence making the reproduction of such stylisation more tractable by data reduction (Nooteboom, 1997; Hirst, 2001; Silverman et al., 1992). Hart et al. (1990) has demonstrated that one can find a reliable basis for such stylization in the way pitch contours are perceived by native listeners. Hart et al. (1990) also showed that intonation can be described in terms of sequences of standard discrete pitch movements supposedly corresponding to voluntary action on the part of the speaker.

2.2.1 Involuntary Aspects of Speech

The prosodic values in speech are often related to involuntary or uncontrolled side-effects besides the rhythm of the language itself. It may be influenced by the stressing of the sentence context, the manner or the place of articulation, the speaker's articulation system or gender and many others. In order to describe the melody of the speech, these involuntary aspects of speech need to be identified first.

2.2.1.1 Speaker

Pitch is perceptually correlated with fundamental frequency (F0). The F0 is determined by the rate of vibration of the vocal cords located in the larynx. The range of F0 of individual speakers varies, and depends on the length and mass of the vocal cords. Therefore male and female, adults and children will have different speech ranges. These characteristics cannot drastically change.

Intensity is another measurable aspect of prosody. There is also a correlation between pitch and intensity: if two sounds have the same intensity and their frequencies lie between about 600 and 2000 Hertz, they will be perceived to be of about the same loudness. Sounds with different intensities may also be perceived as having the same loudness based on *equal loudness contours* as presented by Fletcher-Munson curves or the more recent ones by Robinson-Dadson. Looking past the differences between loudness, intensity and sound pressure, speakers may have thought the volume of their voice is the same but were perceived differently by the listeners.

2.2.1.2 Manner and Place of articulations (and suprasegmental effects)

Involuntary side effects occur in a particular production of speech sounds. For example, given other speech parameters are equal, high vowels like /i/ and /u/ have a higher pitch than low vowels like /a/ (Ladd and Silverman, 1984). In English, the duration of the vowel at the post-vocalic consonants (in a word) is shorter if the consonant is voiceless (Peterson and Lehiste, 1960). The effect is more prominent in phrase and clause boundaries (Klatt and Cooper, 1960). Some consonants are intrinsically longer in duration in different languages, for example the duration of /s/ and /ʃ/ is longer that of other fricatives. There are elaborate studies conducted as described by Klatt and Cooper (1960), Peterson and Barney (1952), and House and Fairbanks (1953) and many others, involving phonetic, syllables and stressed timing especially in English language studies.

It has been shown in Nooteboom (1997) and earlier studies that as vocal effort increases, vowel duration increases and consonant duration decreases. These differences are related to the wider opening of the mouth when speaking loudly compared to speaking normally.

Pauses play a very important role in speech perception. Speech pauses, are regularly used to demarcate major and minor phrases (Ladd and Campbell, 1991; Nooteboom, 1997). There are also pauses or silent intervals which were part of the production of some consecutive phones. For example, in human speech, it is natural to have a silent interval as part of the production of voiceless plosive/stop consonants.

2.2.1.3 Context and Emphasis

The style of speech will have a shift when the emphasis is not noticeable by the listener(s). This also forms the basis for the prosody study where the duration of the emphasized vowel is significantly lengthened (Klatt, 1975; Bolinger, 1972; Umeda, 1975). The lengthening can also be used to capture word frequency and discourse effects that are not otherwise incorporated in the rule system found by Bolinger (1972), Umeda (1975), and Carlson and Granström (1973). Klatt (1975) also proved that the duration of the vowel becomes longer at the phrase boundaries. This applies not only on a vowel, consonantal lengthening such as pre-stressed ‘s’ is longer than ‘s’ before an unstressed vowel (Crystal and House, 1988). Also, it is found by Klatt (1976), Campbell (1992), and Nooteboom (1997), that given other things being equal, lexically stressed syllables

are often considerably longer than lexically unstressed syllables, although this depends much on position within word and phrase. Perception of lexical stress depends to a large extent on the pattern of syllable duration.

2.2.2 Perceptual Equality

What this research is trying to identify is the characteristics in speech that may be used as the basis of finding the similarity in between languages that can be adopted from one to another. However, matching the intonations between the languages will carry together the involuntary aspect of speech that it is mostly speaker and language dependent. These aspects influence the stylization of the speech and some of them are important in order to determine the matching of intonation patterns.

Nooteboom (1997) had shown other studies that demonstrate a close-copy stylization which is a synthetic approximation of the natural course of pitch, meeting two criteria: it should be perceptually indistinguishable from the original and it should contain the smallest possible number of straight-line segments with which this perceptual equality can be achieved. The straight-line segments can easily help in joining a description of intonation in terms of neatly segmented discrete units (Hart et al., 1990; Hirst, 2001).

If one imitated the intonation of speech melody of an utterance, either with the same words or with different words, or with no words at all and by humming, one can obtain a pitch curve that is definitely not perceptually equal to the original. It is easy to hear many differences. However, native listeners can hear whether the imitation is successful in conveying the same melodic impression.

2.2.3 Perceptual Equivalence

Intonation is organised in terms of melodic patterns that are recognizable to a native speaker of the language. Hart et al. (1990) use the term perceptual equivalence which is where two different courses of F0 are perceptually equivalent when they are similar to such extent that one is used to judge a successful melodic imitation of the other.

According to Nooteboom (1997), perceptual equivalence implies that the same speech melody can be recognized in two realizations despite easily noticeable differences, in the same way that the same word can be recognized from different realisations.

The perceptual equivalence allows one to set up an inventory of standard pitch movements covering various sorts of generalisation for any intonation language. In Nooteboom (1997), once the inventory of pitch movements for a particular intonation is defined, it should be possible to generate sequences of such pitch movement. Perceptual equivalence can be evaluated by conducting perceptual tests of reiterant speech to the native speakers.

2.2.4 Perceptual Closeness

For the purpose of this thesis, perceptual closeness is defined as perceived similarity of speech sound produced by one language to the point where it can easily be mistaken as another language by a non-native speaker. And if one uses reiterant speech, a native listener may identify the speech sound as belonging to the original language. Perceptual closeness refers to the languages having a similar rhythm of speech.

Speech however is not rhythmical in the normal way music is. Music has a regular alternation of strong and weak elements in the stream of sound that the upcoming elements can be fairly precisely anticipated. Speech is rhythmical in a looser sense. Speech development in time is controlled by hierarchical mental pattern giving each syllable a certain strength that controls aspects of its production, among which is its duration.

2.3 Introduction to the Malay Language

This research focuses heavily on the application of one source language - Malay resources to be applied to Iban. The two languages are not two divergent dialects, but two totally different languages. However, the typology is close and the geographical positions of language usage are even closer. They also have a closely similar writing system and even a similar syllabification technique (refer Section 5.3.2 for a very brief description). It is thus necessary to have some background on the focal language.

Malay is the native tongue of Malaysia, Singapore, Brunei, Indonesia, Malagasy, selected Philippines Islands to name a few. It is as closely related to Minangkabau as Sundanese is akin to Javanese. The language belongs to the Malayo-Polynesian or Oceanic or Austronesian family, which covers an area from Formosan to New Zealand, from Madagascar to Easter Island, and includes the languages of the Philippines, the Malay Archipelago,

Micronesia, Melanesia excluding Papua, and Polynesia (Winstedt, 1927). The language classification is shown in the Figure 2.2. Samoa, Tahiti, and Tonga belong to the eastern most branch. Malay, Malagasy, Tagalog Bisaya and Bontok in the Philippines belong to the western branch. Similar is the case for all of the following languages: Batak and Menangkabau in Sumatra; Sundanese, Javanese, and Madurese; Balinese; the Dayak dialects of Borneo; and many other less known tongues (Winstedt, 1927). Iban is one of the languages from Dayak dialects.

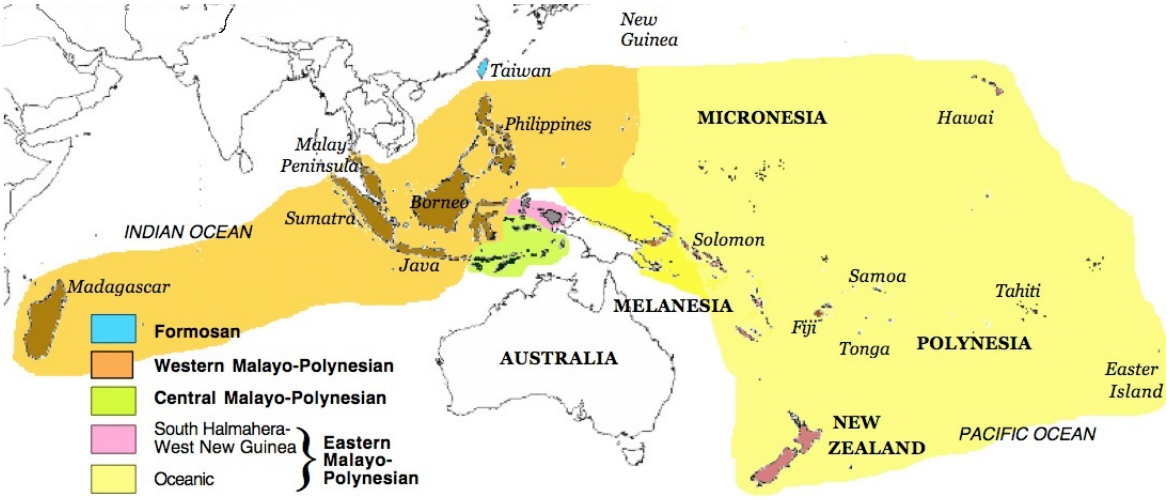

FIGURE 2.2: Malayo-Polynesian Language coverage

Austronesian is the largest language family in the world with about 1200 languages, representing one-fifth of the world's total languages. Its 350 million speakers are spread across an enormous territory ranging from Madagascar in the west to Easter Island in the east and from Hawaii in the north to New Zealand in the south, including peninsular and insular Southeast Asia, most of the islands of the central and south Pacific and Taiwan. While in the western regions of Austronesia some languages are spoken by millions, the many languages of the eastern regions are spoken by few people (one thousand or less per language on average).

2.3.1 The Writing System

There are different dialects influenced due to the typology of the language and the geographical position of the country. The centre and southern of Malaysia, northern, east coast, east of Malaysia, all have different variation dialects. The formal Malay taught in school is Standard Malay. Historically, the dialect commonly used for formal usage is the Johore-Riau dialect. This is what was referred to as Standard Malay. In

1984, a method to standardise the language and a way to follow the language rules in the spelling, vocabulary, terms, language styling and pronunciation (DBP, 2008) called Standard Language of Malay or *Bahasa Melayu Baku* was introduced. Standard Language is easier to teach in the school syllabus, however it is less natural to be used in a formal setting, especially when it involves spoken usage. In 1999, Standard Malay was reinstated as the formal language.

There are vast differences between the earlier form of Malay which is called old Malay and the form that is spoken today: Modern Malay as was described by Omar (1993). She classified Malay into Old Malay (from 7th century), Classical Malay (from 14th century) and Modern Malay (17th century onwards). The language has evolved dramatically in the 19th and 20th centuries. In Winstedt (1927), his discussion of Malay was based on *jawi* script. *Jawi* is an Arabic writing script used with some modification to the character forms to facilitate Malay words. The spelling of *jawi* gradually evolved without any written rules (Abdullah, 2010) and still does to this day. The transliteration of Malay occurred in the early 1600's but the first ever written romanisation of Malay (*rumi transcription*) was the wordlist done by Antonio Pigafetta entitled "The First Italian-Malay Vocabulary (1521)" (Bausani, 1960). According to Bausani (1960), one would find that the romanisation done by Antonio Pigafetta was based on the Italian language from Vicentine dialects. The writing system was not much different than the one used nowadays (written to spoken conversion). It is also said that his vocabulary list was the pioneer and the main reference to the early rumi spelling system which was applied now. Due to the inconsistent spelling used, multiple written systems were put into practice in the late 19th century. Among the early written system introduced were Straits Settlement Rumi (1878), Maxwell Rumi Spelling (1882) and Sweetenham Rumi Spelling (1881).

2.3.2 Phonemes Variations

The phone set used in this thesis is based on those of El-Imam and Don (2005), Ranaivo and Samsudin (2003) and Maris (1979) and they are improvised from Winstedt (1927). However the one introduced by Winstedt (1927) is more elaborate. Despite identifying the phoneme based on examples, Winstedt (1927) identified the distinctions between the phoneme used in Malay in comparison with the languages he was already fluent in and mastered prior to learning Malay.

Some of the phonemes are no longer used in Malay today but are still used in Iban. For example, there are three typess of "o" in Winstedt (1927) but only one currently used in Malay. However, two from Winstedt (1927)'s list are still being used in Iban. Similar with the grapheme "r". Iban has /r/ and /ʀ/, which according to Winstedt (1927): "always being reduced by non scholarly speaker". This sound however was enforced during the time of Za'ba. Za'ba was a distinguished Malay Language scholar in the early 20th century. In 1924, Za'ba founded the Pan-Malayan Malay Literary Society with the sole purpose of standardising Malay spelling and improving the Modern Malay Literature. In 1933, Za'ba revamped the Malay spelling and then, implemented the new writing system based on the R.J. Wilkinson spelling system that is known later as the Za'ba's Spelling System. In the introduced writing system, he identifies the Arabic alphabets as having different phonemes than those originally used in Malay. For example the spelling *maghrib* is pronounced as /maɣrib/ and this resulted the phoneme /ɣ/ being introduced to the Malay phone set. Other added phonemes were /ʕ/ in the word *ta'at*, /s^{l}/ in the word *solat*, /z^{l}/ in the word *zalim*, /x/ in the word *khabar* and /q/ in the word *quran*. These phonemes however were later reduced or simplified into similar latin phonemes in the current system. Therefore, *maghrib* is pronounced as /magrib/, the spelling of *ta'at* changed into *taat* and is pronounced as /ta-at/, *solat* is pronounced as /solat/ and *zalim* is pronounced as /zalim/.

2.3.3 Prosody

According to Winstedt (1927) there is no strong accent on any syllable in Malay words. And words like *perkataan*, *perbuatan*, *aluran*, *kedengaran* and *dikatakan* for example, are pronounced practically with the same stress on every syllable. Each word contains a prefix and a suffix except for the word *aluran* which only consists of suffixes. According to Winstedt (1927), ordinarily in the Malay word, the accent falls on the penultimate syllable except (1) when the penultimate is /ə/ in an open syllable and rarely in a closed, then the accent falls on the last syllable. Examples are as in the word *enam* and *tengah* (2) when a derivative is built up by prefixes from a monosyllabic root, the accent sometimes remains on that root, namely, on the last syllable (3) in the vocative, the stress is sometimes thrown on the last syllable. These has become the foundation of prosody studies in Malay. It was further supported by other studies later on. For example, Madzhi (1989) claimed to detect four degrees of word-stress, with primary

stress falling on the last syllable of isolated and complex words. Don et al. (2008) cited that most earlier studies also supported the similar stream of ideas: Verguin (1955) stated that the first (which is the penultimate) vowel was longer in duration, higher in pitch and greater in amplitude than the final vowel. Kähler (1956) found that the stress fixed on the penultimate when the root word is followed by an suffixes such as -kah, -lah or -pun. Alisjahbana (1957) distinguished dynamic stress, pitch stress and durational stress, claiming that word stress falls on the final syllable, except when it is a clitic pronoun such as -ku or -nya. Halim (1984) basing his study of 140 words, has found word stress on the penultimate in isolated words, and on the final syllable in context. He described a typical stressed syllable as longer and louder than an unstressed one, and having a pitch contour containing a peak of pitch, although the initial pitch could be higher or lower than the final pitch.

However these convictions that Malay has a stressed language characteristics were negated by Maris (1979) who claimed that word-stress in Malay is weak and not very prominent. It is also strongly negated by Don et al. (2008) saying that in their study on a wider context, they found that Malay does not have word stress at all. According to Don et al. (2008), research typically starts off with the assumption that Malay must have word stress like English, and that the task for the researcher is to find it and describe it. One of the strong reason stated by Don et al. (2008) was that the study of Malay prosody was begun based on English study and thus, many prefer to follow the main stream of the predefined framework. If one goes back to the conclusion on phonetics by Winstedt (1927) however, he said:

> "The Indonesian rule is that the accent falls on the penultimate whether of simple or of derivative words...In the Peninsula [Malaysia was originally Malay Peninsula] I confess I had supposed in common with Europeans who have lived there a quarter of a century that the Malay had generally gone back on the old Indonesian rule. But special observation for the purposes of this work has led me to revise my opinion, and to think that while practically there is hardly any accent at all in the words in question, still the Malay does say *perkataan*, *ingatan*, *kudanya*, *namanya*, and *jadikan* - though the suffix "kan" has not this shifting influence when the stem ends in a consonant, and *timbangkan*, *tambatkan* will be correct."

Whether or not Malay is a stressed language is a contentious issue with various researchers arguing for and against but in line with Don et al. (2008), this thesis treats Malay as a non-stressed language.

2.4 Introduction to the Iban Language

Iban is an isolect of the Malayic subgroup of the Autronesian language family. The language is spoken specifically by Iban[1] people in Borneo. They are the indigenous majority in Sarawak while they are a minority in Brunei (Sercombe, 1999). The Iban in Brunei have received little attention other than in demographic studies. The Iban in Sarawak, however have received rather more consideration from scholars in recent times.

The notion that Iban is a Malay dialect is not accepted by most Ibans, who themselves would prefer to be seen as a separate ethnic category. Based on Sercombe (1999)'s literature, there are views of Iban as dialect of Malay. However, it had sparked considerable controversies. Moreover, Iban and Malay are considered to be mutually incomprehensible by those who clearly identify themselves with one or the other grouping (Sercombe, 1999).

Yusof (2003) has proved that Malay and Iban have several features that are closely related to one another. Based on various studies on the phonology, morphology, lexicology and a few others, the relationship is very close, but not sufficient to be considered as a dialect. This again was proven by more recent studies.

In Sarawak, there are 63 indigenous languages spoken by the indigenous communities Yong et al. (2011), Ng et al. (2009), and Sercombe (1999). There are no sufficient resources of the language as well as no thorough studies conducted on these languages. Many researchers believe that the lack of Information Communication and Technologies (ICT) would resulting with difficulty in maintaining the language (Scannell, 2007; Ng et al., 2009; Saee et al., 2008). Upon realisation of the importance of the language survival, Sarawak Language and Technology (SaLT) was established. The studies of this thesis received tremendous help from SaLT who are still working to maintain their language. Due to the sparsity of the data, it is categorised as an under-resourced language.

Despite being categorised under very limited resource, the Iban language has undergone a very rigorous development recently in terms of natural language processing tools as

[1]The language and the people are referred by the same term

compared to other indigenous language in Sarawak. Some examples are construction of a domain ontology by Talita et al. (2010), Iban morphology analyser by Saee et al. (2012) and name entity recognition by Yong et al. (2011). An earlier study on Iban grammar was conducted by Omar (1981) although not for the purpose of creating a natural language processing tool.

2.4.1 The Writing System

Iban is primarily a spoken language called *Jaku' Iban* (conversational Iban). The language has a writing system only for the purpose of learning the language. The orthography was developed in 1900 (Howell and Bailey, 1900) and later standardised in 1956 (Scott, 1956). There has been a steady output of literature published in the language (Sercombe, 1999). Other than several dictionaries of Iban, there was recently a book written about the Iban old script which was based on syllables and characters.

The Iban alphabet was devised by Dunging anak Gunggu (Philip, 2007). He self-taught himself to read and write and then created the writing system in 1947. The Iban alphabet seems to have manifested characteristics of modern writing (Philip, 2007). There are 59 symbols in total which consist of syllables and letters.

2.5 Malay vs Iban Language Features

Ng et al. (2009) conducted an extensive study on the orthographic aspect of Sarawak's most indigenous language and compared to the two major languages used in the country: Malay and English. This included the languages: Iban, Bidayuh, Kelabit, Melanau, Sa'ban, and Penan. They studied the relationship that can be established via the portion of cognates in the Swadesh list of vocabulary words. In their study, it is concluded that the most indegenous language that is closest to Malay formal language is indeed Iban.

The study by Ng et al. (2009) is totally different than the one introduced in Ranaivo-Malançon (2006). Although the author studied the close language identification, Malay and Indonesian shared the early vocabulary and both were influenced by English and Dutch for Malay and Indonesian respectively. The same cannot be said for Iban.

However, due to the very insufficient data of Iban, many approaches of developing Iban language tools use Malay resources. For example, Iban's Automatic Speech Recognition (ASR) using Malay resources only uses 8 hours of data recording of Iban and it was found

that the word error rate (WER) improved by either adding 20 hours of Malay speech data or 4 hours of English data (Juan et al., 2014)[2]. The Iban grapheme-to-phoneme system which was developed earlier was developed from the pre-existing grapheme to phoneme system for Malay, although some of the rules are different, the grapheme-to-phoneme of Iban only needed two hours of manual post-editing (Juan and Besacier, 2013).

Yusof (2003) conducted an investigation of the similarity of Malay and Iban by investigating the phonological, grammatical and lexical aspects. Her study of previous literature as well as her initial assumption stated that Iban is a dialectal of Malay. The grammar and lexical units of both language have over time maintained certain linguistic element and lost others. She believed that both languages have similar morphology structure although she admits that there is a consistent reduction in Iban morphology when corresponding to prefixes in Malay. The morphological structure patterns was also consistent to what was presented in Saee et al. (2012) which showed that the Iban morphology analyser was more simplified than the Malay morphology analyser as presented by Ranaivo-Malançon (2004). Saee et al. (2012) also found that the vocabulary of Iban has similar pronunciation to Malay albeit some spelling differences as well as different word meanings for the same sounding words. However, due to the cognate percentage being less than 85% it is considered to belong to another language.

Yusof (2003) conducted a study to identify the relationship between Bahasa Melayu and Iban using a lexicostatistic approach. She found the proportion of cognate words to be 69%, as compared to other languages in the same family: Jawa, Tagalog, Acheh and Sunda. This again was proven by Ng et al. (2009) in which, the cognancy of Malay and Iban was 62.5% which is also the highest cognizant among other indigenous language in Sarawak that they had studied (Ng et al., 2009).

The similarity between the two languages described here was unique in terms of text processing, other than the ASR for Iban by Juan et al. (2014) and TTS Juan et al. (2011). This will be further explained in Chapter 6.

[2]Quality is much better when using Malay data

2.6 Malay Intonation Pattern in a Sentence

Despite the non-stressed and non-tonal language, there is almost a consistent pattern of Malay intonation which is used in general. Intonation in Malay is important to identify the subject and the predicate in a sentence. In English, for example, the subject and predicate can easily be identified by the verb separation, however it is not the case for Malay where the sentence can stand by itself without a verb. The changes of intonation pattern also influenced by the type of sentence: be it active or passive sentence, or be it a declarative or interrogative sentence.

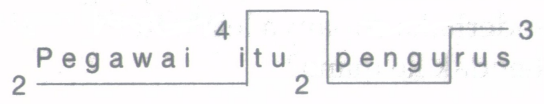

FIGURE 2.3: Intonation for: The officer is the manager

This section provides a general prosody overview by Karim et al. (1996) which is the main reference of the language syllabus in Malaysia. The numbers represent the strength of timbre (Karim et al., 1996). Number 1 representing low timbre while number 4 representing high timbre. Number 2 is the timbre that mark the vocalisation and maintaining the same timbre in the sentence production while number 3 provide the focalisation of the sentence (Karim et al., 1996).

FIGURE 2.4: Intonation for: The officer is the manager in interrogative active sentence

The example from Figure 2.3 cannot be presented in passive form with such limited words. However, the active sentence of interrogative sentence for active sentence is shown in the Figure 2.4 while the corresponding passive sentence is shown in Figure 2.5.

FIGURE 2.5: Intonation for: The officer is the manager in interrogative passive sentence

When the exclamatory sentence are used for giving instruction, there are two types of instructions being used. The first are used to be obeyed (Figure 2.6) and the second

are used in more formal and used towards higher ranking level (Figure 2.7) e.g. towards parents, colleagues, older recipients, etc.

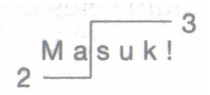

FIGURE 2.6: Intonation for: Come in!

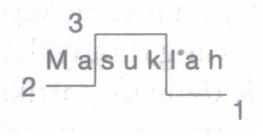

FIGURE 2.7: Intonation for: Come in! (a more subtle version)

Example of active and passive sentence for short sentence can also be viewed in Figure 2.8 and Figure 2.9.

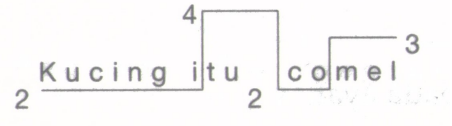

FIGURE 2.8: Intonation for active sentence: It is a cute cat

FIGURE 2.9: Intonation for passive sentence: It is a cute cat

Longer sentences still maintain the intonation pattern as described above. However, for longer sentences, it will be analysed into subject, predicate and description. The usual pattern of 2-4-2-3 is maintained and because the description is just an addition towards the subject and predicate and are treated as less important, the focalisation is put at the subject and predicate of the sentence only. However, if there is a need for emphasising the description for the subject and predicate, the timbre 2-3 can be put in the description phrase.

Despite being the 'bible' of Malay Grammar, Karim et al. (1996) received multiple criticism due to a few factors. Based on the compilation of comments conducted by Mohd Rasidi (2000), the grammar does not covering the breadth of Malay sentences and

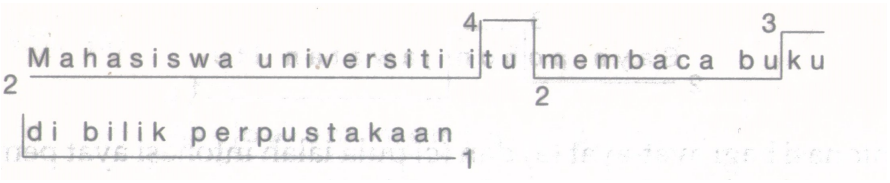

FIGURE 2.10: Intonation for: The first degree student reads a book at the library room

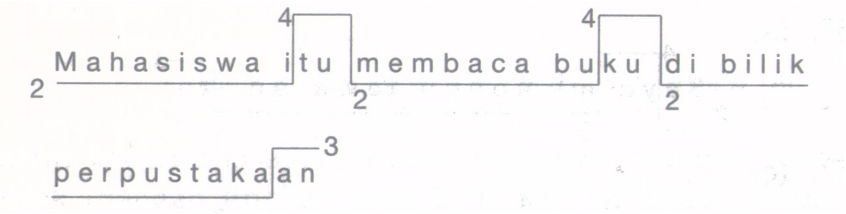

FIGURE 2.11: When emphasising on the description is required: The first degree student reads a book at the library room

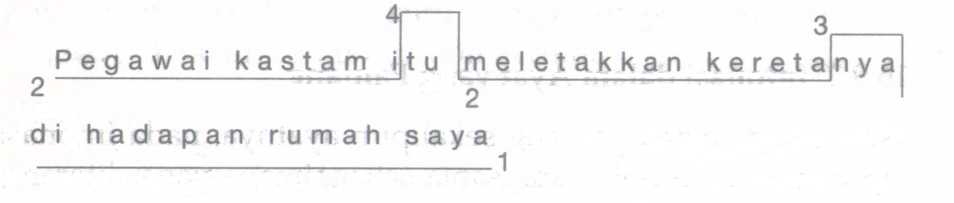

FIGURE 2.12: Intonation for: The immigration officer parked his/her car in front of my house

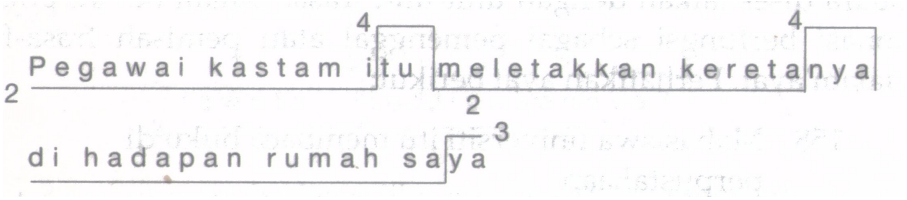

FIGURE 2.13: When emphasising on the description is required: The immigration officer parked his/her car in front of my house

is too moulded by the English generative sentence. According to Mohd Rasidi (2000), the examples also did not portray the regression of the language as of the current usage. Although the intonation patterns introduced by Karim et al. (1996) are correct, it is not exhaustive. Mohd Rasidi (2000) also stated other possible intonation patterns which one can applied in Malay but were not presented in Mohd Rasidi (2000).

Having said that, the presented intonation of Malay does reflect the Iban language. In the early stage of this research, an attempt to obtain prosody pattern of Malay and

Iban via reiterant study was what has been conducted by Klatt and Klatt (1990), Larkey (1983), and Levitt (1991). In the study all respondents were asked to identify which reiterant speech are correspond to Malay. They ticked 'yes' to all of the sample of reiterant speech. This may agree with Mohd Rasidi (2000) who believes that there are other intonation patterns which are not yet explored by Malay linguists.

However, if comparative sentences were to be annotated by INTSINT (International Transcription System for Intonation), both Malay and Iban intonation pattern would not reproduce the intonation set as described by Karim et al. (1996) above. This is due to one, as stated, there is no objective way to pronounce a sentence in Malay. And two, INTSINT itself has 8 variations of phonological representation of intonation and the automatic calculation would be base on the threshold value and since Malay is neither tonal nor stressed language, the intonation labelling would not be consistent from one speaker to another.

2.7 Summary

The literature on melody of speech was presented. The melody, which is also the rhythm or tempo in speech addresses the features other than the discrete prosody features. This may be interlaced with the prosody and linguistic features but without any signal processing or linguistic processing, this definition will not bring any strong meaning towards the pairing of languages based on the melody alone.

The two focal languages in this study were presented: the Malay and Iban. Despite being an under-resource language, Iban has been studied since the middle 19th century. An introduction to both languages were given together with the assumption and studies on the relatedness of the two.

This chapter will be followed by a literature review on multilingual and polyglot speech synthesis.

Chapter 3

Multilingual and Polyglot Speech Synthesis Review

This chapter will look at other approaches of multilingual and polyglot speech synthesis. It will also look at issues on polyglot speech synthesis with different language characteristics.

This chapter presents other approaches currently used in multilingual and polyglot speech synthesis. It will discuss the architectural distinction between multilingual and polyglot approaches and their respective linguistic representations, and also the different methods of language classification. The chapter then describes phoneme adaptation and prosody representation as applied in other TTS techniques and frameworks, and concludes by outlining various approaches to the rapid prototyping of TTS systems.

3.1 Multilingual and Polyglot Speech Synthesis Research and Commercial Product

Multilingual and polyglot TTS systems both handle multiple languages. A multilingual speech synthesiser has different algorithms, rules and speech data for different languages (Traber et al., 1999). A polyglot speech synthesiser has a primary language which is identified as the main language of the synthesiser. The main feature of polyglot speech synthesis is that any system using this framework will be able to synthesise multiple languages using the same set of recorded or trained voices.

Both approaches have advantages and disadvantages, and choosing between them depends on the goal of the developer, and whether time and resources are available to produce high quality synthesised speech as provided by multilingual TTS. However, often

time and resources are not easily available for under-resourced languages and therefore this research is about providing a framework for TTS development when resources are limited or the developer lacks sufficient linguistic information or knowledge of the target language.

This section starts with a brief introduction of a few multilingual or polyglot TTS systems. Then a more detailed discussion of multilingual and polyglot systems is presented with reference to the literature.

3.1.1 CHATR

CHATR is a generic speech synthesis system. It can be considered as a pioneer of high quality monolingual speech synthesis as well as multilingual TTS. This system developed at ATR offers multilingual synthesis for English and Japanese (with Korean and German closely following). Its main waveform synthesis technique uses non-uniform unit selection (Campbell, 1996) from speech databases using acoustic and prosodic features. It can build a voice from any phonetically labelled database. The system allows real-time text to speech functionality, as well as offering a development environment for investigating new speech synthesis techniques. The system is portable and has been tested on seven different common Unix platforms.

3.1.2 AT&T Bell Labs TTS

AT&T were among the earliest to develop multilingual speech synthesis. AT&T Multilingual TTS is a combination of multiple components from well developed TTS systems. The new AT&T Text-To-Speech (TTS) system for general U.S. English text is based on the components of the AT&T Flextalk TTS, the Festival System from the University of Edinburgh, and ATR's CHATR system. From Flextalk, it employs text normalization, letter-to-sound and prosody generation (Beutnagel et al., 1999). Festival provides a flexible and modular architecture for easy experimentation and competitive evaluation of different algorithms or modules. In addition, AT&T adopted CHATR's unit selection algorithms and modified them in an attempt to guarantee high intelligibility under all circumstances, and added the Harmonic plus Noise Model (HNM) backend for synthesizing the output speech.

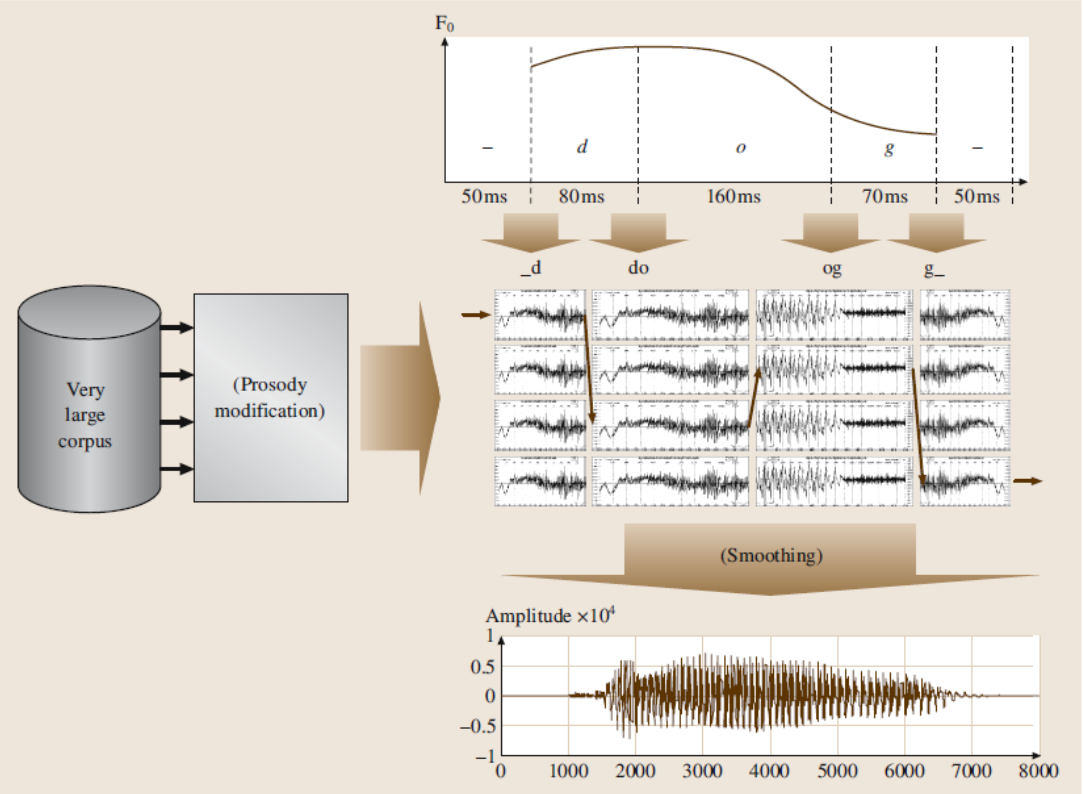

FIGURE 3.1: A schematic view of a unit selection-based speech synthesizer. The prosody modification and smoothing modules may not always be implemented. In fact, since this approach uses very large speech corpora, it is often possible to find speech units that naturally join smoothly while exhibiting prosodic features close to what is expected. Note that, unlike as suggested in this figure, unit selection-based synthesis systems do not systematically use diphone units. For a domain specific TTS, even words can be stored to ensure very high quality speech.

3.1.3 CLUSTERGEN

Kominek, 2009 created an incremental TTS for multilingual speech research. The CLUSTERGEN tools were developed as a tool in a statistical parametric speech synthesis approach. Kominek, 2009 focused on using the CART tree to have a complete lexicon. The main focus was to reduce the effort required to build TTS for new languages. New language here refers to having no existing and acceptable synthetic voices in the target language. The digital resources are also limited and therefore the phone set has to be explicitly designed during the voice building procedure and the creation of a pronunciation lexicon. The CLUSTERGEN synthesiser can be implemented within Festival/Festvox voice building environment (Black, 2006).

3.1.4 Festival

The Festival TTS system provides a general framework for the construction of a TTS. It specifies the stages in TTS development and also provides a range of alternative methods for use at each stage. The Festival framework is presented in Figure 3.2. The diagram shows a general method for creating a TTS regardless of the target language, as well as the input and output at each stage and the options and functions which can be used. Festival is a collective effort of years of research and prototypes. Festival is a multilingual TTS framework.

Pipeline Stage	Function	Description	Comment
Text		character string in ASCII or utf-8	"Hey, you!"
↓	Textify	convert text to token sequence	(name, whitespace, punc, prepunc)
Tokens		processable word-like elements	("Hey" "" "," "") ("you" "" "!" "")
↓	Token_POS	classify tokens into word types	converts numbers and
	Token	convert tokens to words	dates, e.g. 54th→fifty fourth
Words		lexical elements	hey, there
↓	POS	classify a word's part of speech	det, aux, content, ...
	Phrasify	split utterance into phrases	NB, B, BB
	Word	convert words to sound segments	lexicon and/or G2P rules
	Pauses	insert pauses at phrase breaks	beg/end/internal pauses
Phonetics		pronounceable elements	pau hh ey y uw pau
↓	Intonation	add Tobi accent features	H*, L*, L+H, etc.
	PostLex	vowel, cross-word reduction	segment seq can change
	Duration	predict segment durations	scales default duration
	Int_Targets	generate F0 interpolation points	finer grained than Tobi
Generation		speech production	
↓	WaveSynth	mechanism depends on the synthesis method (diphone, unitsel, hts, clustergen)	speech can be generated (lpc, hnm, hmm) or concatenated (unitsel)
Waveform		final result	

FIGURE 3.2: Festival pipeline with a short description of each stage, as presented by Kominek, 2009

3.1.5 Verbmobil

Verbmobil is the result of eight years of intensive research in a large speech-to-speech translation project. The system that was developed handles dialogues in three business-oriented domains, with translation between three languages: German, English, and Japanese. Verbmobil deals with spontaneous speech, which includes context repair speech, and uses deep semantic analysis and therefore can correctly recognise a speaker's slips and can correct a translation of what one tried to say rather than what one actually said.

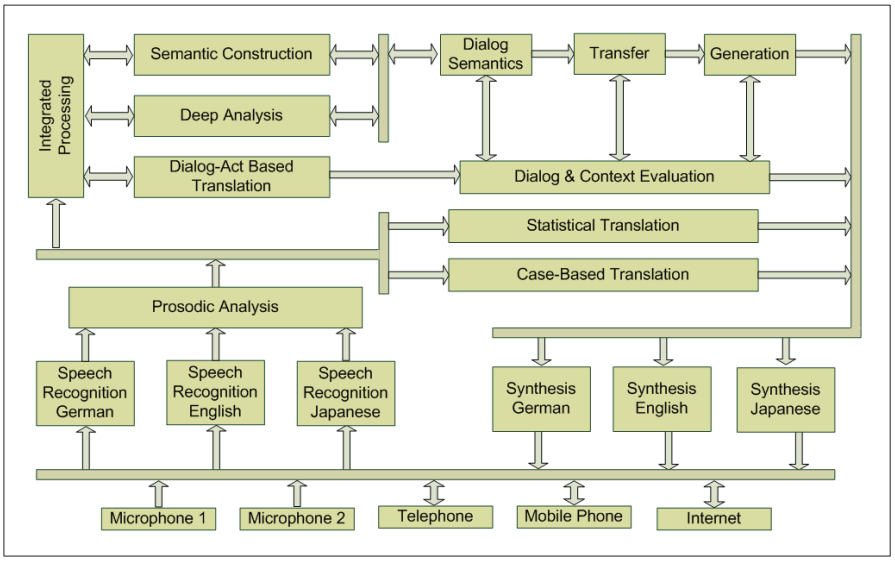

FIGURE 3.3: Complete Verbmobil Architecture (Wahlster, 2000). The synthesisers originated from CHATR as described in Campbell, 1996. Due to its superior naturalness, Verbmobil's German and English synthesiser uses the same architecture.

Inside the Verbmobil synthesiser, there are three different synthesisers for three different languages. All were based on the CHATR synthesiser (Stöber et al., 2000).

3.1.6 Loquendo Text-to-Speech

Nuance Communications market a commercial multilingual TTS available in thirty languages (and counting). This is one of the most popular commercial TTS products. It supports all major operating systems and speech-related standards and is available in an extensive choice of configurations to meet the requirements of any application. It can be used as a TTS package for existing language or can be modified at surface level. Each language is recorded for hours.

It uses unit selection techniques (Quazza et al., 2001). It has expressive TTS and can add animated phrases like "Welcome!" or "Amazing!" and sounds like coughing, laughter or crying. It also supports an expandable lexicon, where the user can enter new entries to define the pronunciation of acronyms, proper names, abbreviations, to name a few, according to the application context. It also allows high level prosody control, e.g. modifying speaking rate, pitch, pause frequency, and length. Other than that, Loquendo supports the Speech Synthesis Markup Language (SSML), which allows a varied input. It also supports mixed languages where voices can pronounce foreign-language words while maintaining their native accent.

3.2 Different Approaches in Multilingual and Polyglot Speech Synthesis

This section presents a general review as well as of the differences between multilingual and polyglot TTS approaches. For some, these terms may seem similar and so the differences might not seem great. However, high quality synthesised speech is easier to achieve with multilingual TTS while polyglot TTS requires more training for good TTS quality.

3.2.1 Multilingual Speech Synthesis

A multilingual synthesiser is more suited to applications for teaching and learning languages and when an accurate pronunciation of a language must be distinguished correctly from another language or when a foreign accent and dialect is not acceptable. It is also suitable for when the system to be developed does not have any issues in terms of availability of linguistic resources or resource storage size. This makes multilingual speech synthesis system a very reliable but expensive framework.

Generally multilingual TTS design closely follows monolingual TTS architecture. This can be seen with architecture as an example (Romsdorfer and Pfister, 2007). Figure 3.4 illustrates the PolySVOX system's flow from text analysis to phonological processing and onto prosody control prior to the synthesis module. Romsdorfer and Pfister, 2007 also highlight the list of language dependent rules and corpora which correspond directly to each stage of the multilingual TTS.

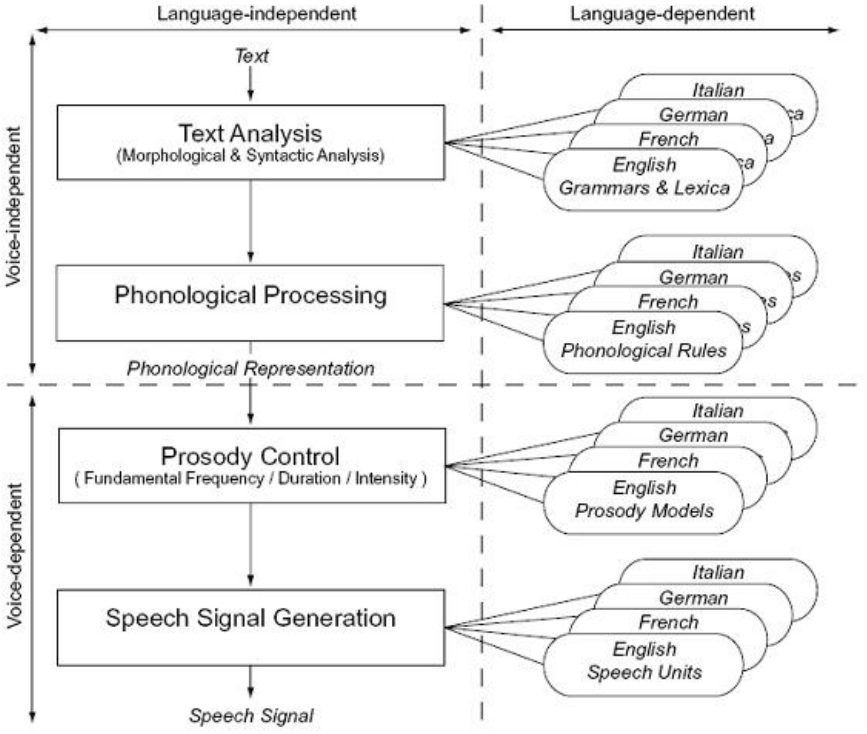

FIGURE 3.4: PolySVOX: An example of a multilingual TTS architecture (Romsdorfer and Pfister, 2007).

Another example is MBROLA (MBROLA-Group, 2005), which uses one speech synthesiser but has 72 diphone speech corpora from 37 languages, each with its own grapheme-to-phoneme transcription. In order to construct a TTS based on the MBROLA framework it is necessary to define the language-dependent components, i.e. the grapheme-to-phoneme conversion, the phonological rules, the language text analysis and preprocessing as well as prosody modelling. This demonstrates that some aspects of monolingual frameworks are essential to multilingual TTS architectures.

In order to build language-dependent components in multilingual speech, it is necessary to obtain information relating to the textual and linguistic aspects of the language and to present this information in the required format so that it can link with the other components. This data is best obtained from linguistic experts in the relevant target languages. This information collection requirement may seem difficult to fulfil but a far nearer to native-speaker quality can be achieved with this approach.

3.2.2 Polyglot Speech Synthesis

Polyglot speech synthesis is particularly suitable for mixed-lingual text (Romsdorfer and Pfister, 2007). For example, with occurrences of xenomorphs[1] it would not be practical to switch from one corpus to another. It is also useful for fast prototyping systems, as with SPICE (Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages) by Schultz et al. (2007), which requires only a very short voice recording for training in certain languages, depending on English as its base language (Kominek et al., 2007). As a result, the polyglot approach is also suitable for building a TTS for insufficiently resourced languages. Although it has many practical advantages, the output of a polyglot speech synthesiser will retain some traces of foreign accents or imprecise pronunciation.

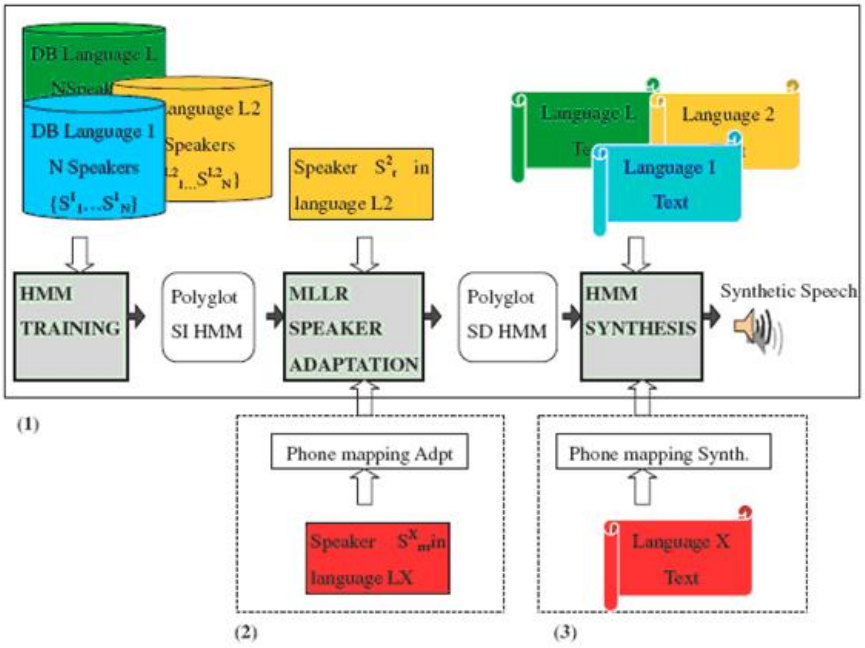

FIGURE 3.5: Example of Polyglot Architecture (Latorre et al., 2006).

Another example of a system developed in this framework is described by Latorre et al., 2006; its architecture is shown in Figure 3.5. There are two major phases in this synthesiser, namely the training phase and the synthesis phase. During the training phase, collections of speech in all the target languages are processed and the spectral features of the speech are extracted and stored using a method of training involving Hidden Markov

[1]Words that are built from combinations of two morphemes from different languages(Steigner and Schröder, 2007).

Models (HMM). In Figure 3.5, a speaker independent HMM is constructed after the first training instance. As a result of this, information about the speaker's vocal characteristics is not retained. To 'rebuild' a speaker's voice characteristics, speaker adaptation is required, which increases consistency in the synthesised speech quality.

In Kominek et al., 2007 the SPICE interface makes use of very short target language recordings, where the longest conversation lasts for 38 minutes and the shortest recording is 10 minutes. With this size of corpus the system needs to adapt phonemes from other languages for use in the target language.

Latorre et al., 2006 presented three sections that form a speaker adaptable polyglot speech synthesiser such as depicted in Figure 3.5. Section (1) shows the flow from recorded speech which undergoes HMM training resulting in the construction of speaker independent (SI) and later speaker dependent (SD) HMMs. Section (2) shows how the system adapts the phoneme mapping when the language to be synthesised is not included in the training data. The system makes use of a voice in the target language and adapts the target language phoneme into the available phoneme collection, and thus the polyglot system will construct a collection of SD-HMMs for the target language. This idea is an excellent approach to overcome situations when speech data is insufficient. However, such an architecture still requires voice recording in an appropriate environment and native speaker involvement in constructing the language resources.

Figure 3.6 shows that in the synthesis phase, the generation of speech still requires the text analysis component.

3.3 Literature Review on NLP Manipulation of Multilingual Processing

Prior section has summarised the approaches use to create a polyglot speech synthesis by Schultz et al. (2007), Latorre et al. (2006), and Kominek et al. (2007) that focussed on speech signal manipulation to handle insufficient speech data. This section focus on other researchers' approach to improve the multilingual TTS front-end component.

3.3.1 Text Pre-Processing and Analysis

Huang et al., 2001 define the first two processes in TTS as text analysis and phonetic analysis. In these processes, input text is normalised, checked for syntactic structure

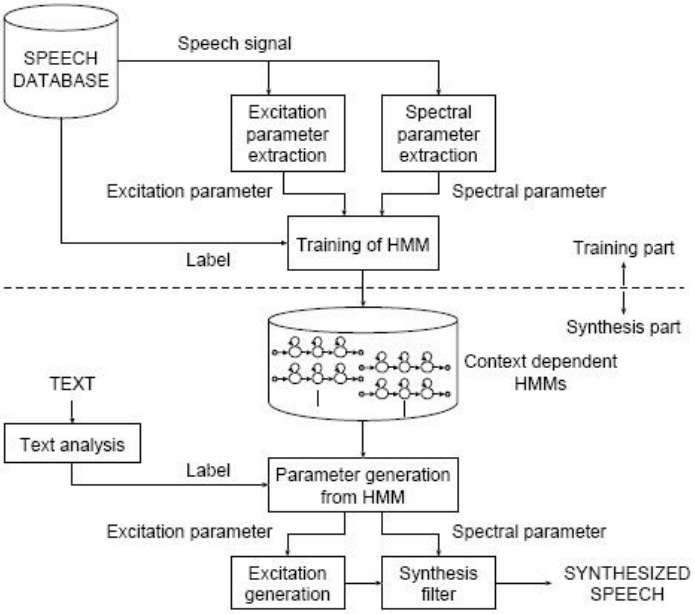

FIGURE 3.6: The distinction between the training and synthesis processes (Tokuda et al., 2002).

and transformed into its corresponding phonetic representation. In their text analysis, there is document structure detection to tag the input so that it can be processed into segments i.e. paragraphs, sentences, words and smaller units. Text normalisation performs the conversion from a variety of symbols, numbers and other non-orthographic text entities to a common orthographic transcription.

Sproat, 1996 describes the text analysis model for a multilingual TTS system developed for Bell Laboratories and based on Weighted Finite State Transducers (WFST). This model, together with Bell's text normalisation, forms the overall text analysis module for their TTS system. Text normalisation handles word segmentation, digit expansion, abbreviation expansion, the correct pronunciation of common words, and prosodic phrasing. After normalisation, WFST are constructed based on lexical, morphological, numeral expansions and phonological rules.

Another approach to text analysis for mixed-lingual text, presented by Romsdorfer and Pfister, 2007 in their system they called PolySVOX, required language-specific data and a language independent algorithm. They broke text analysis down into three main processes:

- language identification
- generation of phonetic transcription

- analysis of the syntactic structure of the text.

To handle mixed-lingual text, three types of foreign language inclusions were identified. These are:

- **mixed-lingual words** are produced by applying base language conjugation rules or compounding rules to a foreign stem. This type of inclusion mainly occurred with English or French stems in German text.
  Example: "$_{G}$Das ($_{E}$Musical) programm ($_{E}$New York's) wurde ($_{F}$en passant) ($_{E}$up)ge($_{E}$dat)et".
- **full foreign words embedded in a base language context** which can be inconsistent with the base language syntax.
  Example: "$_{G}$Wird das ($_{F}$Café) nicht von Ihren ($_{E}$Fans) belagert?".
- **foreign multi-word inclusions** which are correct according to both foreign and base language syntax.
  Example: "$_{G}$($_{E}$Lobbying)($_{F}$à discrétion) vor der Vergabe der Olympischen Spiele von 2012 in Singapur".

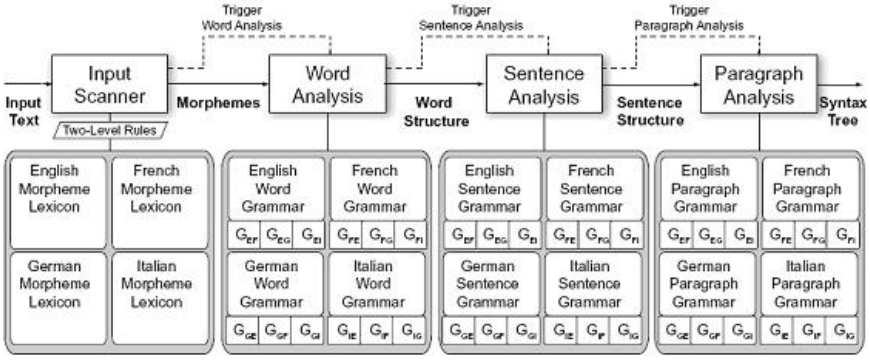

FIGURE 3.7: Architecture of morphological and syntactic analysis in the PolySVOX TTS synthesis system.

Text analysis processing in PolySVOX is illustrated in Figure 3.7. As shown in this figure, PolySVOX's text analysis approach is based on rule-based processing using a chart parser with word, sentence and paragraph grammars. The processing is done in sequence, from the smallest unit size (words) to the largest (paragraphs). The notation G_{ij} specifies an inclusion grammar that describes inclusions of language j in language i. The abbreviations E, F, G, and I refer to English, French, German and Italian respectively.

Each level (word, sentence and paragraph) is provided with a monolingual grammar alongside the inclusion grammar. When analysing mixed-lingual input text, monolingual analysis results are favoured over mixed-lingual ones.

Although PolySVOX is very focused and detailed, it is also highly specific to the synthesis of Swiss (French, German and Italian) speech and cannot easily be adapted to other languages.

3.3.2 Phonological Processing in Multilingual TTS

Romsdorfer and Pfister, 2004 discuss phonological processing, which is the module after text analysis in the PolySVOX system. At the initial state, a surface phonetic transformation is produced. The pre-defined set of phonological rules used in the PolySVOX TTS system is obtained.

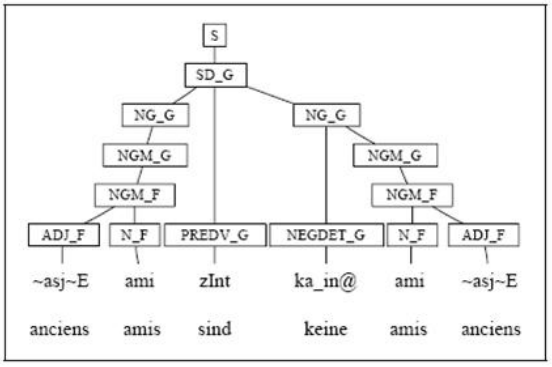

FIGURE 3.8: Syntax tree of the sentence 'Anciens Amis sind keine Amis anciens', including graphemic and phonetic terminals. The phonetic symbols largely follow the SAMPA definition. The suffixes _F and _G of the constituent identifiers indicate the languages French and German.

The output of this component is a complete phonetic representation of the sentence to be synthesised after the phonological transformation of the four languages (Romsdorfer and Pfister, 2004). In order to ensure the phonological component of the system is able to cope with different languages, flexible formalism rules are introduced, containing possible context restrictions for such phonological rules.

In total there are 74 rewrite rules to describe all pronunciation variations in the languages covered by the system. There are two stages in the experiment: the straightforward grapheme-to-phoneme conversions followed by the iterative insertion, deletion and

replacement of segments relative to the preceding iteration based on the pre-determined pronunciation variation rules.

Research into PolySVOX has pioneered a practical implementation of a flexible multilingual TTS and demonstrated what would be the best adaptation approach for such a system.

3.3.3 Prosody Modelling

Prosody in this TTS context refers to the pitch, duration and loudness of the overall speech. Although it is used interchangeably with intonation, prosody is both measurable and manipulable using prosody analysis and prosody assignment. In prosody analysis, information about speech prosody is extracted from speech. In prosody assignment, a value is estimated according to the prosody control definition as set in a particular system.

For example, the MBROLA synthesiser provides flexible manipulation of pitch and duration. As the recorded diphone waveform needs to be kept within a specific frequency range, MBROLA does not offer adjustment to its amplitude during the synthesising process.

In Hirst, 2001 an automatic prosodic analysis is presented using the MOMEL/INTSINT algorithm. There are four basic steps in MOMEL to normalise the speech signal, while INTSINT labels the waveform to a predefined intonation type. Based on the analysis of prosody, the contour can be predicted for the analysed language. Romsdorfer and Pfister, 2005 describe prosody estimation based on the changes among a few neighbouring syllables and a few prosodic parameters. Although their study was conducted on one language, it is believed that the approach is suitable for use with all languages.

Two case studies will be presented in this section. The first concerns Spanish speech synthesis using Case-Based Reasoning (CBR) as a prosody estimator, followed by a detailed description of prosody estimation implemented in PolySVOX.

3.3.3.1 Spanish Speech Synthesis using CBR as Prosody Estimator

Gonzalvo et al., 2007a describe prosody estimation using Case-Based Reasoning (CBR).

Starting with a HMM-based architecture similar to the polyglot speech synthesis architecture shown in Figure 3.6, they add a component to refine the excitation signal for speech generation.

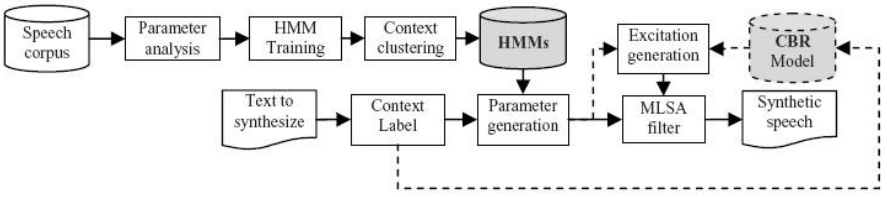

FIGURE 3.9: Spanish HMM-based Speech Synthesis by Gonzalvo et al., 2007b

Each HMM in this architecture represents a contextual phoneme. Similar phonemes are clustered based on contextual information and designated questions, such as: "is it to the right context an 'a' vowel" or "is left context an unvoiced consonant"? In this process, if a contextual phoneme does not have a HMM representation (e.g. if it is not available in the training data), a decision tree cluster will generate the unseen model.

Figure 3.9 shows the Spanish HMM architecture with the CBR approach represented by a dotted line. During synthesis, the target is to construct a list of phonemes to synthesise based on the input text. Chosen units are converted into a sequence of a HMM chain. In Gonzalvo et al., 2007b, spectrum and F0 parameters are generated from HMM models using dynamic features. The duration is estimated to maximise the probability of state durations. The excitation signal is generated from the F0 curves and the voicing information. Finally, the speech is constructed using spectrum parameters (Mel Log Spectrum Approximation) and the excitation signal.

The CBR strategy was originally designed for retrieving mean phoneme information related to F0, energy and duration. However, the work described in Gonzalvo et al., 2007b and Gonzalvo et al., 2007a focuses only on using CBR as an F0 estimator.

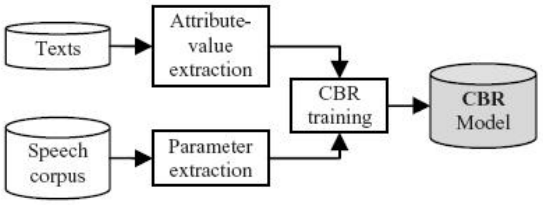

FIGURE 3.10: CBR Training workflow (Gonzalvo et al., 2007a).

In this architecture, text is first analysed by the SinLib Library, a Spanish text analysis tool (Gonzalvo et al., 2007b), which extracts the characteristics that will build prosody cases. When presented with a speech corpus, each file is analysed in order to convert it into new cases. The goal is to obtain a prosody estimation from the memory of cases that best match the problem. When a new text is entered, CBR will look for the most similar cases it has in memory from which to retrieve the prosody information.

This approach works well when there is one target language. However, it requires a lot data to cover the full breadth of prosody cases in the target language. The training also requires original voice recording. Therefore the approach is capable of producing a good model of prosody but isn't suitable if resources are limited.

3.4 Rapid Prototyping TTS

Schultz and Waibel, 1999 explored the effectiveness of porting multilingual speech recognition systems to new target languages with limited data. The key idea behind the multilingual speech recognition (SR) engine was to combine context-dependent acoustic models across languages to adapt the existing resources of Large Vocabulary Continuous Speech Recognition (LVCSR) to other languages using a method they call Polyphone Decision Tree Specialization.

The global phonetic inventory used was based on those of monolingual systems. Sounds represented by the same IPA character shared a common phoneme category. For the multilingual systems, language dependent acoustic models of five languages were combined: Croatian, Japanese, Korean, Spanish and Turkish.

Two methods of combining were used. The first is called ML-mix, where all models are shared across these five languages without preserving language information. The other is ML-tag, where the phoneme model shared across languages is created by attaching language descriptions to each of the phoneme categories. A context-dependent model is applied using decision tree clustering, which uses a set of linguistically-motivated questions about phonetic context. In ML-tag, additional questions about language and language group are added so that the question of phonetic context information is more important than language information. In pronunciation dictionary mapping, an appropriate mapping is required to describe the transition from global phoneme set to target

phoneme. The polyphone[2] decision tree method is implemented to make the process of finding the possible target language phoneme more accurate. This will reduce the error rate for the recognition. They also agrees with previous research that suggests that multilingual SR systems perform better when recognising the same language family rather than across family borders.

Black and Schultz, 2006 describe the requirements for multilingual speech synthesis using resources from SR, namely the definition of a global phoneme selection and a method of adapting SR data into TTS. The difference between SR and TTS speech resources is that TTS normally emphasises the phonetic balance of the speech corpus of one speaker while SR deals with phonetically balanced speech with a wide range of voices in training. In Black and Schultz, 2006, the speech database to be used in TTS is identified, a speaker clustering method is implemented, and the clustered speech data is labelled. Black and Schultz, 2006 used the CLUSTERGEN synthesiser, which is a statistical TTS.

In the experiment, two types of multilingual speech were evaluated: a clustered multilingual speaker without explicit language description and a clustered multilingual speaker with language features attached to the speech sample.

Schultz et al., 2007 and Kominek et al., 2007 describe the rapid prototyping approach to producing speech-to-speech translation using a system called SPICE (Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages). The system is developed in a web-based environment and requires only very small amounts of speech data. SPICE can be used for Bulgarian, English, German, Hindi, Konkani, Mandarin and Vietnamese.

The system flow uses the standard TTS framework i.e. text collection, audio collection, grapheme definition, phoneme selection, grapheme-to-phoneme rules, pronunciation lexicon and finally the speech synthesis module.

The grapheme-to-phoneme rules are the crossing point for the conversion from a monolingual to a multilingual system. Initially, a default phoneme needs to be assigned to each grapheme. At the second stage, the system learns from users as the pronunciation dictionary grows. The system then uses the Festival system for synthesis.

[2]A letter (or combination of letters) that has two or more pronunciations. E.g. <c>is a polyphone. It can be pronounced like /k/ in car and /c/ or /tʃ/ in cell.

Schultz and Waibel, 1999 conclude that the acoustic model combined with attached linguistic information will offer better recognition performance. However, according to Black and Schultz, 2006, the multilingual TTS with linguistic features attached performs no better than the one without. Although these are entirely contradictory conclusions, the applications involved in the comparison are targeting two different goals. Schultz and Waibel, 1999 aim to build a multilingual SR application to determine whether the technique they propose (Polyphone Decision Tree Specialization) helps to improve the application's ability to recognise multilingual speech. On the other hand, the SR speech data described in Black and Schultz, 2006 is clustered and undergoes signal modification to produce consistent speech as if it were coming from a single speaker.

Based on these four pieces of research it is clear that multilingual TTS quality depends heavily on the method of synthesis, in terms of how the speech is adapted and clustered. It also requires global phoneme definitions to account for all the sounds in all languages. Language peculiarities can then be attended to individually. Current research refers to these rapid prototyping approaches for comparison.

3.5 Summary

This chapter provided a brief summary of multilingual/polyglot speech synthesis as well as approaches to language adaptation in multilingual speech synthesis research. At the beginning of the chapter, a general discussion on multilingual and polyglot TTS approaches outlined the difference between the two, and thus their relation to this thesis's research title. The generic architecture of multilingual and polyglot frameworks were also presented.

One of the most important issues in a multilingual and polyglot TTS system is text preprocessing. Other research on polyglot synthesis may have a need to focus on the detailed description of text manipulation due to the nature of the language itself.

As the idea underpinning this research concerns adaptation from one resource to another, approaches to phoneme adaptation in other research (specifically speech recognition systems) were also explored. The chapter concluded with the work on rapid prototyping, fast TTS development and low-resource TTS systems. Most approaches use various statistical parametric approaches (cluster or HMM based) rather than concatenative approaches despite high quality speech for the unit selection case.

This chapter will be followed by the study of how phonemes can be reused specifically in phoneme substitution if the given target language needs to use the source language's full phonetic data.

Chapter 4

Phoneme Substitution in Letter-to-Phone Processing for Combinational Speech Resources

4.1 Adapting Phoneme Resources from Resource Languages

One of the limitations in producing speech synthesisers from another language is the lack of phonemes available to facilitate the target language. For instance there is the sound /β/ in Spanish but mainly perceived and produced like a /b/ by second or third language speakers. There is also the sound /ɾ/ in French but there is also /r/ in French. Therefore, even when one language's voice recordings can be used to produce another language's synthesiser, given that the foreignness of the synthesised speech is acceptable, the missing phonemes cannot be easily substituted with other phonemes for which (let's say) the manner or the place of articulation is near. This is one of the issues faced when some sounds are not used in another languages when resource sharing happens.

Making a new recording for new (or non-existing) speech resources is not straightforward. It requires a lot of trial-and-error, redoing of voice recordings, text refinement, maintaining the tone, speed and timbre in the voice recordings, obtaining a sufficiently good device for recording, using a good recording environment if an anechoic chamber is not accessible, labelling of the speech and many other factors which might be trivial but very crucial in producing a consistent voice. These processes are also influenced by a lot of parameters, for example the speaker's condition - health, age, fatigue level, mood and etc., device placement, background noise control, device noise, device consistency, labelling accuracy and many others. These parameters, combined with a thorough

procedure can create a lot of restrictions in creating acceptable data resources. Therefore, reusing an existing resource might be a preferable choice before a very elaborate recording and processing should be carried out.

This chapter will discuss a method which makes it possible to reuse existing data from another language. It will discuss the issues of phoneme substitution from a phoneme confusion perspective. The first section will give an overview of phoneme confusion, followed by the study on phoneme confusion done by multiple languages by different researchers. Then a study on phoneme confusion and an experiment on phoneme substitution will be presented.

4.2 Phoneme Confusion

In speech recognition, each phoneme is estimated using a collection of probabilities of what word might be formed by the phoneme recogniser. Multiple approaches have been used to handle these probabilities and possible words. It is tedious however to reverse engineer the technique used in phoneme recognition to identify the possible confusion of each phoneme since speech consists of signal recording and signal recording is too variable to be sufficiently consistent. During the earlier study of a speech recognition system, vector quantization was used to categorise different phonemes. It can also show the most probable phonemes and this indirectly shows the confusion of phonemes which might happen. In such a situation, a phoneme confusion matrix is created by aligning the hypothesis from the phoneme recogniser to the corresponding reference phoneme sequence from the forced alignment of a speech recognition system. The alignment will show the hypothesized phoneme actually realized at the position of the actual phoneme.

Instead of trying to identify the phoneme and therefore the word, this research attempts to identify the phoneme which can be perceived as another. Therefore to record the similarity or confusion between two phones, a more classical approach was used. Following previous literature on approaching the creation of phoneme substitution matrices, this section will explore phoneme confusion further.

4.2.1 Studies on Phoneme Confusions

In speech recognition, phonemes tend to be misinterpreted due to the confusion of the phoneme recogniser. Several studies have been conducted on human and machine perception, among them Miller and Nicely (1955), Fant et al. (1966), Lovitt and Allen (2006), Lovitt et al. (2007), Meyer et al. (2007), and Cutler et al. (2004).

Miller and Nicely (1955) used 16 consonants of evaluation by constructing logatomes or nonsense utterances with a CV syllable constructed where the V was always /ɑː/. Miller and Nicely (1955) devised confusion analysis to understand how humans confuse phonemes. Fant et al. (1966) used the similar syllable structure to Miller and Nicely (1955) where an English utterance test was constructed using 22 possible consonant phonemes at the initial position. In a Swedish test, 17 possible initial single consonants were used. Lovitt et al. (2007) in a different approach tried to identify where the causes of confusion started or happened in an automatic speech recognition system. Lovitt et al. (2007) extended the experiments in Lovitt and Allen (2006) which used only the CV structure by adding the VCV structure into the experiments. However, instead of human identification, Lovitt et al. (2007) used human mispronunciation, speech features confusion and phoneme recogniser confusion. Cutler et al. (2004) on the other hand expanded the study originated by Miller and Nicely (1955) by using 24 consonants over 15 vowels used in English among 16 native listeners, and 16 non-native (Dutch) listeners.

The confusion matrix study of Fant et al. (1966) listed the confusions that happened during the listening test in two conditions. In one, Fant et al. (1966) listed the confusions that happened when listeners were asked to hear a recording which underwent low-pass filtering at 2000Hz with a high quality filter. In the second, Fant et al. (1966) presented the confusions that happened when white noise was added to 13 signal-to-noise ratio (SNR) sounds. The sounds were played over high-quality loudspeakers to the listeners. Due to the effect of low-pass filtering on dentals and fricatives which resulted in those not being recognised at all, the results of added white noise were used as comparison. 13 dB noise is below the average speech level and therefore the effect of the noise was less drastic than the filtering (Fant et al., 1966). (For readability, the confusion matrices are given in Appendix A.4.) It is also important to state that only one subject was used in this study. The subject was a bilingual with equal command of English and Swedish

since childhood. The subject was given 10 randomised wordlists for each language for each phoneme.

What had been found in Miller and Nicely (1955) was further studied by Cutler et al. (2004), Meyer et al. (2007) and Lovitt et al. (2007). Cutler et al. (2004) conducted a study using CV and VC structures and compared the confusion between American-English and Dutch speakers. The main focus of the study was to provide a new data set of phonetic identifications given a different level of noise (calculated by SNR) by native and non-native listeners. Cutler et al. (2004) obtained 645 logatome syllables representing each of the different phoneme combinations. The noises were added from conversational speech which was also pre-recorded in a quiet room. Conversational speech was later added as a background noise to the recording. The recordings were mixed and added so that each logatome would have three different SNRs (0 dB, 8 dB and 16 dB). The results of the confusion matrices are seen in Appendix A.2 giving only the non-added noise confusion results for both native and non-native listeners. The paper has shown that the non-native listener performed below native phoneme-identification levels. However, Cutler et al. (2004) also concluded that the non-native listeners appeared to remain fairly constant (in producing the confusion phoneme) across SNRs within the tested range as compared to native speakers.

Meyer et al. (2007) presented the comparison of human and machine phoneme recognition. In the human speech recognition test, Meyer et al. (2007) used two kinds of signal. One was using noisy speech samples in which the sound to be evaluated was re-synthesised using MFCC. Another one used the original signal with added noise which was used to evaluate the loss of information caused by the process of re-synthesis. In their study, Meyer et al. (2007) used CVC or VCV structures and, like Miller and Nicely (1955), used nonsense utterances. For human speech recognition, five normal hearing listeners were requested to identify the two types of signals given. 150 utterances were given to be evaluated. The outcome from the study is given in Appendix A.3. According to Meyer et al. (2007) the choice of SNR when involving noise addition was based on presentation of only a few test lists to one human listener and proved to be reasonable for other test subjects as well. This was close to the SNR selected by Fant et al. (1966) who chose to include an SNR of 13 dB.

Lovitt et al. (2007) studied the confusion that occurred across three stages. Each confusion was categorised as the following: pronunciation confusion, frame confusion and phoneme confusion respectively. These were the three of the five stages in phoneme recognition. Pronunciation confusion refers to the mispronounced word. Frame confusion is the probability of error that the extracted features of the corresponding phonemes were not done correctly. Finally, the phoneme confusion is the mistaken identification by the phoneme recogniser itself. The purpose of the study by Lovitt et al. (2007) was to identify the confusion patterns to improve the performance of a recogniser by eliminating problematic phoneme distinctions. Lovitt et al. (2007) wanted the phoneme recognition to be re-analysed into a smaller subset of phonemes which could be considered as common confusion patterns so that the system should be able to provide the supposed result and not treat these selected phoneme group confusions as errors in phoneme identification. (The summary of phoneme confusions presented by Lovitt et al. (2007) is given in Appendix A.4.) Lovitt et al. (2007) also stated that the confusion (from the phoneme recogniser) may have lost its voicing and place of articulation features which resulted in the misidentification of phoneme. This also informed the direction of this research whereby, when the voicing and the place of articulation information were lost, the phoneme can still be determined.

4.2.2 The Study on Phoneme Confusion for Malay

For the phoneme confusion study, 17 respondents were involved in evaluating 255 sounds which consisted of CV, VC and CVC syllable structures. There were a few Malay phonemes missing in this confusion study. There were 39 identified phonemes and two unidentified ones based on Ranaivo and Samsudin (2003), 34 phonemes were based on MBROLA-Group (2005) and there were 38 based on Li et al. (2005). The phoneme lists were different due to the acceptance of declaring the borrowed phonemes from other languages such as Malay phonemes. For example, the Malay phonemes listed by MBROLA did not include /v/ as a phoneme even though there are Malay words using this phoneme. This is because the loan words with such phonemes usually undergo transformation. For example, violin is known as *biola* and is a loan word from Portuguese, *viola*; goddess, is known as *dewi* and is a loan word from Sanskrit, *devi*; fasting, is known as *puasa* pronounced as /puwasə/ and was a loan word from Sanskrit, *upavasa*. However, for loan words from English, there were two categories, unplanned adaptation and planned

adaptation (Ahmad et al., 2011; IPG, 2011). For planned adaptation, in occurrences of /v/, slight changes took place; governor is *gabenor* and private is *prebet*. For unplanned adaptation, the words did not undergo transformation when there was /v/. For example, television is *televisyen*, activity is *aktiviti* and university is *universiti*. Therefore in the Malay phoneme list, MBROLA-Group (2005) did not consider /v/ as a Malay phoneme.

Additionally, in pronunciation, there was a slight variation which was also not listed. For example, Clyness and Deterding (2011) stated that there is only one alveolar trill, "r", in Malay. However, during an observation of a speaker's recording, two "r"s were used: /r/ and /ɾ/. According to Clyness and Deterding (2011), the speaker used the formal style. The speaker may have phonological influences from standard Malay and English. Because there were no stringent rules as to when a certain sound should be tap or trill, or when some audible release and reduction of a phoneme was supposed to take place, it thus created an additional phone which is not considered in this confusion study.

For each listening test, each consonant was paired to a vowel. Three vowels were used for this study to get a better description of human perception. Therefore there were three instances for each generated consonant. Each consonant was paired to one closed vowel, one mid vowel and one open vowel. Each sound was not supposed to be meaningful in Malay.[1]

This thesis studies on Malay confusion matrix was based on 17 respondents. This number of respondents was very close to the study conducted by Cutler et al. (2004). This number number was different when compared to Miller and Nicely (1955) and Meyer et al. (2007) that both used 5 respondents. Therefore the approach of creating the confusion matrices was more closely similar to the Cutler et al. (2004)'s than Miller and Nicely (1955)'s. Lovitt et al. (2007) on contrast, uses a phoneme recogniser to identify the confusion. According to Lovitt et al. (2007), the phoneme recogniser was making similar errors to a human speaker made in speech production.

All respondents were encouraged to take as long as they wished to answer, and allowing submission part-by-part so they were not stressed during listening. However, they were requested to use the same equipment to ensure the consistency of the given feedback. Contrary to the studies conducted by Miller and Nicely (1955), Cutler et al. (2004),

[1]Five words coincidentally exist in Malay: *kek*, *gam*, *tak*, *Mac* and *di*

Meyer et al. (2007) and Lovitt et al. (2007) that used human speech recording, this study used synthesised speech. The generated sounds excluded the following consonants: /x/, /ʕ/ and /?/ due to limited occurrences in the Malay training data itself.

4.2.2.1 Phoneme Confusion Matrix for Consonants in syllable CV

The first was the study on phoneme confusion for consonants positioned at the beginning of the syllable CV. For the CV confusion study of Malay, it was comparable to another four confusion studies conducted by the research reviewed in Section 4.2.1. The confusion matrix for Malay was as presented in Table 4.1. Each phoneme was paired with three vowels in different occurrences. The vowels used were: /a/, /e/ or /ə/ and /i/.

As shown in Table 4.1, /p/ was highly confused with /b/ and /f/ among Malay listeners. In fact, it has a greater impression of being an /f/ than /p/ itself. In Miller and Nicely (1955) however, the phoneme /f/ was only minimally confused as /p/ compared to: /k/, /t/, and /θ/. Compared to Cutler et al. (2004)'s experiment for the consonant in CV structures, /p/ was also confused among listeners with /b/and /f/. But American English listeners also confused the /p/ to be /h/ which was also mistakenly identified as higher than the listeners' labelling of the /p/ as /p/ itself. As for Dutch listeners, /p/ was frequently heard as itself. It was also mostly confused with /h/, /b/, /f/ and /k/. Based on Meyer et al. (2007), /p/ was highly confused with /k/, /b/, and /v/, and according to Lovitt et al. (2007), /p/ was confused with /b/, /t/, /k/ and /f/.

The detailed comparison across different approaches and studies is presented in Table 4.2. The main focus of this comparison was to see the phonemes which were noticeably identified as another. The different degrees of misidentification were shown in colours. The red showed that the produced phonemes were confused by being the perceived phonemes more than the correctly identified phonemes except for Lovitt et al. (2007). The blue colour represented a different frequency number across the study.

TABLE 4.1: Phonemes confusion for onset consonants for syllable structure: CV

		Observed Phoneme Identified by Listeners																				
		b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
Speech Synthesiser's Phones Production	b	32		1	1						4				1	1				**10**	4		
	tʃ		47															**7**					
	d	1		46		1		1		1									1	1		1	1
	f				50										1		1	1	1				
	g		1	**6**		31		**7**	3											1			**5**
	h			1	3	3	24	1	**10**	1					**4**				1			**6**	
	dʒ		1					53															
	k				1		**4**		34	**12**		3											
	l									47		3				**4**							
	m	2								2	38										**11**	1	
	n									2	2	30	**12**								**8**		
	ŋ					2						**7**	34	**5**								**6**	
	ɲ									3	2		**13**	18							**10**	**8**	
	p	**12**			**21**				1		1				16	1				1	1		
	r	2		1	2										1	36					**10**		2
	s		1														50	3					
	ʃ		**6**														2	46					
	t		**4**	**6**				**9**	2								**5**		26	1			1
	v	1															1			39	**10**	3	
	w									2	1			2		2				1	41	**5**	
	j			2			1			**10**				2		1					2	36	
	z			3		2		**13**									2						34

Blue in the Malay study means the confusion was sufficiently misidentified and happened not due to one person's misperception. For Miller and Nicely (1955), the blue colour phonemes mean they were misidentified by more than ten times and greens showed that the misidentification happened ten or less but greater than or equal to four. This is due to the numbers involved in Miller and Nicely (1955)'s study being very high and when misidentification of less than four happened, it is believed that it was caused by isolated mistakes. For Cutler et al. (2004), blue referred to the number less than the number identified by the phoneme itself but higher than five, while green was for four or five frequencies of response only for the same reason as Miller and Nicely (1955). The same applied to Meyer et al. (2007). Lovitt et al. (2007), in contrast, did not use frequency of response. Therefore, the red in Lovitt et al. (2007) referred to the phoneme being misidentified throughout the three mentioned stages: human pronunciation confusion, frame speech features confusion and phoneme confusion. Blue indicates that the confusion happened in any of the two stages while green indicates that the misidentification happened only in one stage.

Looking generally at each phoneme in the study, the /p/ was always confused with /f/ across different research studies, only the frequencies of it occurring over other phonemes were different. Other than that, /p/ was also confused with /b/, except for Miller and Nicely (1955). /p/ was also constantly confused with /k/ except for the Malay study. However, in the VC study in the next section, the phoneme /s/ was also frequently confused as /k/. The /b/ was always confused as /v/.

From the list of confusions happening across different language settings and experiments, there was almost no clear correspondence across languages. For the Malay confusion study, it was expected that the recognition rate was higher for consonants in the CV structure and especially less confusing for plosive and sounds originating between dental and post-alveolar due to the place combined with the manner of articulation. This however was not proven. Based on general observation, it is believed that the voiceless phonemes tend to create more confusion than the voiced phonemes.

Cutler et al. (2004)'s respondents tend to misidentify plosive phonemes as /h/. It can be easily dismissed as a technical error. However, the sound produced after post-alveolar onwards (towards glottal) may have been confused as /h/ due to the aspiration effect. This was indirectly supported by Miller and Nicely (1955) quoted by Fant et al. (1966),

who stated that nearly all the confusions were in terms of different places of articulation within a subclass of constant manner of articulation. From Miller and Nicely (1955)'s 0 SNR feedback however, it was mostly true for the highest confusion in the list. For example, the phoneme /g/ was mainly confused as /d/ which was also a plosive but other confusions were from fricatives. For the phoneme /b/ however, the highest confusion was the phoneme /v/ which is a fricative but has a similar place of articulation. It sufficed to re-iterate that other studies used human voices with either added noise or re-sampled voiced.

According to Fant et al. (1966), when the sounds were re-filtered, the feedback showed the clear tendency that dental stops and fricatives were almost never recognised as such. This is also similar in the Malay study. When confusions happened during the dental of plosive, fricatives or nasal, the frequency of the confusions will be more on the post-retroflex sounds. The phoneme /t/ was confused as the affricate /dʒ/, /n/ as /ŋ/ and /z/ as /dʒ/. Although /dʒ/ is not even a plosive, the sound is closely similar to the sound /ɟ/ which does not exist in Malay. Hearing /z/ as an affricate was understandably due to the burst of sound before /z/. These similarities were believed to happened because the synthesised speech was generated using a Malay speech synthesiser using the HTS approach. The recording was done at the 44kHz but then during feature extraction, it was downsampled to 22kHz. This could lead to some respondents hearing the sound as if it had been filtered.

The confusions were less prominent for CV syllables compared to VC syllables. The confusions also occurred less often when paired with vowel /a/ and /i/ rather than /ə/ and /e/. This also explained why the previous studies always used /a/ or /ɐ/. Confusion studies on VC syllables did not have a lot of comparable studies. Therefore the comparison study will be done on the coda consonants for VC and CVC syllable for Malay studies.

4.2.2.2 Phoneme Confusion Matrix for Consonants in syllable VC

For the VC study, three vowels were paired to the Malay consonants. However, the pair of /ij/ was dropped because, it never occurs in Malay and following the standard Malay spelling, the /j/ sound is subtly assimilated into /i/. At the beginning of the experiment, it was believed that the VC syllable would produce more confusions than

the CV syllable. From the respondents' feedback, it was more accurate to conclude that the confusions were sparser than the syllable structure CV. Some confusion also had higher frequencies than the phoneme itself. This indicated that some phonemes could be easily confused with others when the phonemes were at the coda position. It was also found that the voiced and voiceless phonemes had more confusion phonemes than consonants. However, the voiceless phonemes' confusions were caused by an individual's perception rather than the perceptual confusion itself. This can be observed by the frequencies of occurrences for some confusion phonemes.

For plosive phonemes, /p/ was not able to be identified as itself more than half of the occurrences. It was confused with /f/, /k/, /b/ and /v/. This was similar to what has been presented by Lovitt et al. (2007) where the /p/ was also confused as /k/, /f/ and /b/ in Lovitt et al. (2007) study. The similarity existed for Malay onset consonants where the /f/ and /b/ were also listed as confusion phonemes. The /b/ in Malay had a high identification as itself, but also was confused as /m/, /p/ and /v/. This again was similar to Lovitt et al. (2007) - /v/ and /p/ and Malay onsets - /v/ and /m/. For phoneme /t/, it was confused as /d/, /b/ and /k/ but the only similarity with Malay onsets was /d/ while Lovitt et al. (2007) also listed /d/ and /k/ as its confusions.

Comparing the response with the syllable CVC, the phoneme /t/ was confused as /d/ and /p/. This is similar to the feedback presented by Lovitt et al. (2007). For the phoneme /d/, it was mainly confused with /m/, /n/, and /t/. All belonged to dental. The phoneme /k/ was confused with /g/ however the phoneme /g/, other than being confused with /k/, was also confused with /dʒ/.

The two affricates /tʃ/ and /dʒ/ were mostly confused with each other. The phoneme /tʃ/ wwas highly confused as /dʒ/. The confusion number was higher than the recognition of the phoneme itself. The phoneme /dʒ/ mostly was identified as itself. When confusions occurred, it was mainly perceived as /tʃ/. In real usage, these affricates rarely occur at the coda position in Malay words. When it did happen however, most of the time it was a /dʒ/. Examples of such usage are: *majlis*(ceremony), *majmuk*(plural), *buruj*(constellation), *hijrah*(migration), *koc*(train coach) and *Mac*(March).

TABLE 4.2: Phonemes confusion across different observations

Phoneme	Malay	Miller and Nicely (1955)	Cutler et al. (2004)	Meyer et al. (2007)	Lovitt et al. (2007)
p	f, b	k, t, θ, f	h, f, b, k, θ, t h, b, k, f, t, v	b, k, v, g, t, f	t, k, f, b
b	v, m, w	v, ð, f, θ, d	h, m, ð, θ, f, v h, f, m, p, k, w, l, v	v, g, p	v, p, θ, d
t	dʒ, d, s, tʃ	p, k	h, p, k, θ, f p, k, h, θ, f, b	d, b	d, p, k, r, q, tʃ, s
d	(none)	g, ʒ, z, ð	n, ð, b, θ, j, l ð, n, l, b, h, j, θ, m	g, b	t, θ, g, dʒ, ɾ
tʃ	ʃ	(not tested)	t t, dʒ	(none)	ʃ, dʒ, t, s
dʒ	(none)	(not tested)	ð, tʃ, d tʃ, ð, d, j, p	(not tested)	ʒ, tʃ, z, j, d, t
k	l, h	(not tested)	h, t, p p, h, t	g, v	t, p, g
g	dʒ, d, z	d, ʒ, z, ð	j, h, n j, b, h, k	k, v	k, d, t
m	w	n	v, l, n n, b, r, l	l, n, v	m̩, n
n	ŋ, w	m	m m, l	l, m	ɾ̃, n̩, ŋ, m
ŋ	n, j, ɲ	(not tested)	(not tested)	(not tested)	n, m
ɲ	ŋ, w, j	(not tested)	(not tested)	(not tested)	(not tested)
r	w	(not tested)	n, b, v b, w	(not tested)	ɚ, ɝ
f	(none)	θ, k, s, p	p, h, b, θ, ð p, b, k, h, θ, v, ð	v	s, θ, v, z
v	w	ð, b, z	b, ð, h, f, θ b, f, ð, p, h, w, θ	b, g	f, ð, z, b
θ	(not tested)	f, p, s, t, k	ð, f, p, h, b, t p, b, f, ð, h, t	(not tested)	ð, t, f, v, b
ð	(not tested)	v, z, g, b	θ, l, b, v, n, z, d b, θ, l	(not tested)	θ, d, v, f, b
s	(none)	θ,ʃ, f	θ, ð, f, z θ, z, f, ð	f, ʃ	ʃ, z, f
z	dʒ	ʒ, g, ð, d, v	ð, θ, v, w ð, θ, b, n	(not tested)	s ʒ, v
ʃ	tʃ	(none)	tʃ tʃ	(none)	tʃ, ʒ, s
ʒ	(not tested)	z	(not tested)	(not tested)	ʃ, z, dʒ, tʃ, s, ʁ, n̩
h	k, j, p	(not tested)	p, f, t, k p, k, f, b, θ, v, t	(not tested)	ɦ, q, f
j	l	(not tested)	d dʒ, n	(not tested)	(not tested)
l	r	(not tested)	m, b, n, ð m, b, p, w	n	l̩, oʊ, w
w	j	(not tested)	m dʒ, b	(not tested)	l, ɔ, uː

TABLE 4.3: Phonemes confusion for coda consonants for syllable structure: VC

		Observed Phoneme Identified by Listeners																					
		b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
Speech Synthesiser's Phones Production	b	26			2						**10**				**7**				1	**6**	2		
	tʃ		20			**4**	2	**23**										**4**	1				
	d	3		17		2		1	3	2	**9**	**7**	2		1				**5**			2	
	f	2			26		2	1		3		1				2		**5**	2	**8**	1		1
	g			2		26	1	**4**	**7**	3	1	3	1		2	1			3				
	h			1	2		19	**4**	**5**	2			1		1		**7**	3	**5**			**4**	
	dʒ		**7**			3		43				1											
	k		1			**9**	2	1	36	1		1			1				1			1	
	l								1	28					**5**	**7**			2			**11**	
	m	1									51	1			1								
	n									1	**6**	40	**5**		1							1	
	ŋ					2					**12**	6	32	1								1	
	ɲ							1				**13**	**17**	21								2	
	p	**7**			**12**	3			**8**	1	1		1		14				2	**5**			
	r			1	2				2	**12**		1				27	1		**6**			2	
	s											1					46	1					**6**
	ʃ		**6**					1	1								3	43					
	t	**4**	1	**6**	2			3	**4**	2					3	1			25			1	2
	v	2			**11**	1				1	**11**		2			2				21	3		
	w						2			3										3	44	2	
	j	1					3			2											1	29	
	z							**6**									**5**	3	1				39

Fricatives /f/ were confused as /v/ or /ʃ/. The confusions were quite scattered but two were the prominent ones. The /v/ confusions were also scattered but when confusions occurred, it was mainly detected as /f and m/. These were also true for the CVC syllable structure in the coda position. The /s/ was only confused as /z/ while /z/ was confused as /dʒ/ and /s/. The phoneme /ʃ/ was confused as /tʃ/ instead of /s/ in the initial assumption. The confusions of /h/ were very scattered. It was mistaken as /s/, /t/, /k/, /dʒ/ and /j/.

The bilabial nasal had no confusions even at the coda position. However, /n/ was sometimes confused as /m/ or /ŋ/. /ŋ/ was mistaken as /m/ quite frequently and sometimes as /n/ while /ɲ/ had many confusions with /n/ and /ŋ/. This may be due to /ɲ/ almost never occuring in coda position in Malay.

For /r/, it was mainly confused as /l/ and as /t/. Both are dental. For other glides: /w/ and /j/, no prominent confusion occurred.

4.2.2.3 Phonemes Confusion Matrix for Onset in syllable CVC

It was expected that the observation on the onset of a CVC syllable would show consistency in phoneme confusions given a better context (due to the adjacent consonants to the vowel). From Table 4.4, the distribution can be seen as less sparse than in Table 4.1. It is believed that this is due to the structure of the syllables, the respondents being surer of what they thought they heard and thus perceiving less ambiguity.

As with the CV structure, the phoneme /p/ was also mistaken as the phonemes /f/ and /b/. But the confusions were heavily focussed on /f/ and the same with /b/ where confusions were heavily focussed on /v/. However, a few phoneme confusions identified with /d/. For the phoneme /t/, the confusions were only with /tʃ/ and /d/ which showed noticeable reduction of confusions as compared to the syllable CV. The phoneme /d/ was not confused much other than /t/. The phoneme /k/ was confused as /tʃ/. The similarity between the two were that both have 'plosiveness' as manner of articulation. The phoneme /g/ also had multiple confusions like the CV structure. It was confused as /dʒ/, /d/ and /k/.

The affricate /tʃ/ was confused as /dʒ/; however /dʒ/ was not confused at all at the onset of syllable CVC.

For fricatives, when confusion happened for the phoneme /f/, it was perceived as /v/. The phoneme /v/ however, was not mistaken at all. A similar condition was found for the phonemes /s/ and /z/. The phoneme /ʃ/ was confused as /tʃ/, while the phoneme /h/ was mistaken as /p/.

The nasal confusions were less consistent for consonants at the onset of a CVC structure as compared to CV. As with the confusions in CV for /m/, it was also mistaken as /w/ but with lesser frequency. The phoneme /n/ was mistaken as /l/, /ŋ/ and /j/. The phoneme /ŋ/ was very sparsely distributed but has the consistency of being mistaken as the phonemes /n/ and /j/. Finally, /ɲ/ was confused as /ŋ/ and /j/.

The phoneme /r/ was sometimes mistaken as /v/. The phoneme /j/ was mistaken as /l/, however the phonemes /l/ and /w/ were not mistakenly perceived at all.

It can be concluded that at the onset position, when more phonemes were provided for the respondent to guess, the phoneme was less likely to be mistaken as another phoneme. However, it can also be observed that some phonemes can really be confused as something else and multiple phonemes were confused as its voiceless/voiced pair. It also happened due to the vowel used in the pair as well as the second consonant (its coda) usage.

The next section includes the comparable study conducted by Cutler et al. (2004).

4.2.2.4 Phonemes Confusion Matrix for Coda in syllable CVC

When constructing the CVC syllables for the confusion study, the focus was specifically on the coda of the syllable and the vowel usage. Before obtaining the results for phoneme confusions at the coda of the syllable CVC, it was assumed that the confusions would closely reflect the phoneme confusions at the coda for syllable VC. However it was later found that this was not exactly true.

TABLE 4.4: Phonemes confusion for onset consonants for syllable structure: CVC

		Observed Phoneme Identified by Listeners																					
		b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
Speech Synthesiser's Phones Production	b	29		**4**																**19**	2		
	tʃ		44					**9**										1					
	d	3		36		1		2		3		1	1			1			**4**	1			1
	f				46												3			**5**			
	g			**5**		33		**8**	**5**										2				1
	h		1		3		44		1						**4**							1	
	dʒ							54															
	k		**7**		1	1	2		43														
	l									50		1										3	
	m				1						45	1	1								**6**		
	n			1						**6**	2	35	**6**									**4**	
	ŋ									3	2	**5**	31	2							1	**10**	
	ɲ											3	**9**	37								**5**	
	p	**4**			**18**										28		1		3				
	r				3											45				**5**	1		
	s		1														53						
	ʃ		**6**					1									2	45					
	t		**6**	**5**				2									1		40				
	v									1									1	51	1		
	w															2				2	48	2	
	j									**9**						1	1					43	
	z							1															53

The summary of confusions between VC and CVC is presented in Table 4.6 in the respective columns. There was no indication that CVC or VC adds context to the articulation sequence that helps with the identifications. However, it can be seen that the confusions between CVC and VC for phoneme /p/, /b/, /t/ and /d/ are similar and also for /n/, /ŋ/, /ɲ/ and /r/. It can also be observed that the glides and liquids tend to be confused with each other. It should be emphasised that /r/ is not consistently a trill in Malay. When a very similar pronunciation is produced, Malay native speakers will easily accept it as an "r", despite it being produced as /ɾ/, /ɽ/ or sometimes to the extent of /ʁ/ (which might happened due to lack of practise of trill or tap during childhood). For synthesised speech, the sound may be produced as /r/ or /ɾ/ because it was what was being produced by the training voices.

Because /k/ and /g/ plosiveness originates from uvular and differs only in the voiced/voiceless categories, it was assumed that they will be confused with each other. However it was not true for /k/ which was where confusions happened; it was mainly perceived as /t/. /tʃ/ and /dʒ/ were again confused with each other although /tʃ/ confusions with /dʒ/ were higher than vice versa. Fricatives tended to be confused with its voiced or voiceless pair.

Despite being different, there were slight patterns of confusions that can be seen across languages. For fricative confusions, since the affricates existed in the languages (as listed in Table 4.6), respondents tended to confuse the fricatives as affricates or the corresponding plosive counterpart of the affricates besides the phoneme's own neighbour.

In the represented study, Malay has six phonemes: /a/, /e/, /ə/, /i/, /o/ and /ʊ/. However, there were more sounds due to style of talking, dialects and influence from first language. For example, if the "a" sounded like /a/, /ɐ/, /ae/ or /ɑ/, it would still be understandable and written as "a". There were also confusions for the phoneme /e/. The /e/ also usually produced as /ɛ/ or sometimes /ɜ/. This study needed to skip the vowel's confusion phoneme for Malay. This was due to the looseness of vowel pronunciation, and non-systematic writing system in Malay which made it problematic to distinguish such occurrences except for those respondents familiar with the IPA writing system. The feedback from the respondents on the vowels is presented in the Appendix at Section A.5.

TABLE 4.5: Phoneme confusion for coda consonants for syllable structure: CVC

		Observed Phoneme Identified by Listeners																					
		b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	ɲ	p	r	s	ʃ	t	v	w	j	z
Speech Synthesiser's Phones Production	b	6			**4**	1	2		1		**11**	**7**	2		**10**				**7**	2	1		
	tʃ		28			1		**18**										**5**	2				
	d	2		19	1		1	1	**4**		**5**	**4**	1		**4**	2	1		**8**	1			
	f				35		**6**		3						1		3		2	**4**			
	g	3		1	3	28			**10**			1	1		3				2	1		1	
	h				**6**	1	34		**5**	1						1			3			3	
	dʒ		**13**			2		37				1							1				
	k	1				3			40			1			2				**6**			1	
	l				1					13		**19**	**4**			**6**				1	2	**8**	
	m										49	**4**										1	
	n									2	2	35	**15**										
	ŋ				1						**17**	**11**	24		1								
	ɲ											**29**	**8**	14		1						2	
	p	**4**		1	**10**		2		**5**		1	1			25				**5**				
	r						3	1		**13**		1	2			27	2		1				**4**
	s															1	48						**5**
	ʃ		3														**4**	45					2
	t	3		**13**	3	2			3	1					**8**				21				
	v	3			**13**	2	2				**8**	1				1	**6**			**17**	1		
	w									1			1	1							41	**10**	
	j				1		2			**6**									1			44	
	z			1				2									**8**	4					39

TABLE 4.6: Phoneme confusion comparison for coda consonants for syllable structure: CVC

Phoneme	Malay VC	Malay CVC	Cutler et al. (2004) - English	Cutler et al. (2004) - Dutch
p	f, k, b, v	f, k, t, b	t, k, f, θ	b, t, k, θ, f, d
b	m, p, v	m, p, t, n, f	v, d, ð, k	d, v, t, θ, ð, p
t	d, b, k	d, p	k, p, θ	d, θ, k, ð, p
d	m, n, t	t, m, k, n, p	v, ʤ, n, ʒ, g	t, ð, θ, ʤ
k	g	t	t, p	t, g, p, θ, f
g	k, ʤ	k	d, b, ð, θ	d, t, ʤ, ð, k
tʃ	ʤ, ʃ, g	ʤ, ʃ	ʤ	ʤ, ʃ, ʒ
ʤ	tʃ	tʃ	ʒ, d	tʃ, ʒ, d, ð
m	(none)	n	ŋ, n, v	n, t, ŋ, dθ
n	m, ŋ	ŋ	m, ŋ, d	t, ŋ, m, d
ŋ	m, n	m, n	n, m, g, v	n, m, t, g, d
ɲ	ŋ, n	n, ŋ	(not tested)	(not tested)
r	l, t	l, z	(none)	t, d
f	v, ʃ	h, v	θ, p, t, k, ð	t, d, θ, p, k, ð
v	f, m	f, m, s	f, g, ð, d	d, t, f, ð, θ, b, l
θ	(not tested)	(not tested)	f, t, ð, p	t, f, ð, d, p
ð	(not tested)	(not tested)	d, v, ʤ, z, ʒ, g	d, t, v, θ, ʤ
s	z	z	f, θ	f, θ, ð, ʃ, z
z	ʤ, s	s, ʃ	v, d, ð, s, ʒ, ʤ	s, θ, ð, d, v, ʒ
ʃ	tʃ	s	tʃ	ʒ, s
ʒ	(not tested)	(not tested)	ʤ, ð, v	ʃ, ʤ, z, tʃ, ð
h	s, k, t, ʤ, j	f, k	(not tested)	(not tested)
j	(none)	l	(not tested)	(not tested)
l	j, r, p	n, j, r, ŋ	f, v	d, t, f, r
w	(none)	j	(not tested)	(not tested)

Since consistent confusions were difficult to obtain, a more direct approach to the confusions survey was conducted. Given a specific context, respondents were asked to listen and type back what they heard from the list of sounds.

4.3 Phoneme Substitution

Three different sets of surveys were conducted. For the first survey, intelligibility tests were carried out on non-modified and modified words. In the second and third surveys, the respondents were requested to listen to sets of sounds and were asked to type back what they heard. This is perception based on context. The sounds were synthesised by a HMM-based synthesiser, using the OALD pronunciation dictionary and a built-in Festival pronunciation dictionary. The list of words is given in Appendix **??**.

4.3.1 Intelligibility on Substituted Phoneme's Words

Formal studies were conducted on English words where mixes of valid English words with substituted phonemes were evaluated together with English words which had not undergone any changes. The words with change of phonemes were added into the pronunciation dictionary to ensure that the intended sounds were produced. Then, a call for respondents was made to evaluate the intelligibility of the sound. The surveys were run based on the assumption that the phonemes could be substituted with another phoneme in certain conditions so as to imitate the original word pronunciation. The intelligibility tests were conducted by letting the respondents run the survey at their own convenience and pace. All respondents conducted the survey using a pair of headphones.

Seventeen respondents participated in the survey. The feedback from the respondents was recorded based on the stated assumption. (The values in Table 4.7 represent frequencies.) From the data, there were 124 conditions where the modified words with one modified phoneme were perceived as the intended words. It is important to state that the modified words were not valid words, and therefore the respondents were forced to write the possible words. In 43 cases the modified words were identified as different words. The non-modified column shows that the frequencies of words that were not modified, representing the controlled experiment. This experiment was conducted to identify how the respondents perceived the synthesised speech in general. There were 111 correctly identified words and 50 incorrectly identified words (from the controlled sample).

From Table 4.7, the sensitivity of the overall feedback was 0.7164. The misclassification was 0.2835. To further analyse the results, a test of statistical significance was conducted using the chi square test. The expected frequencies are at Table 4.8.

Table 4.7: Respondents' identifications of the sampled data

	Modified Words		Total
	Yes	No	
Correctly Identified	124	111	235
Incorrectly Identified	43	50	93
Total	167	161	328

TABLE 4.8: Respondents' identifications of the expected data

	Modified Words		Total
	Yes	No	
Correctly Identified	119.6494	115.3506	235
Incorrectly Identified	47.35061	45.64939	93
Total	167	161	328

Based on this re-evaluation, the value of chi square, χ^2 was 1.1366. However for degrees of freedom (df) equal to 1 and p=0.05, χ^2 must equal or exceed 3.84 to be significant. Therefore in terms of intelligibility testing, perceiving the substituted phoneme as the intended word was possibly due to chance.

4.3.2 Perception based on Context

To further evaluate the possibility of that such a substitution can be perceived as the intended sound, a set of perceptual tests was conducted on onset and coda modifications in a context evaluation. In this evaluation, a string of three or four phonemes was arranged in sequence and each word had similar rhyme. One of the words would have a slight change: either the onset or coda was different from the others. Examples of an onset and a coda difference are as follows:

*green, groan, **c**rane, grain*

*clock, cloak, clo**g**, cluck*

These experiments were conducted on the assumption that it was easier for the respondents to confuse the sound of the different onset or coda due to the neighbouring words. The respondents were simply told to type back what they heard. There were 24 respondents in the onset study and 21 respondents in the coda study. All respondents were native English speakers.

4.3.2.1 Onset Evaluation

In the onset evaluation (Table 4.9), 138 respondents identified words that that were affected by the neighbouring words and 81 words were not. For the control set of sounds, in the total of 519 words paired into three to four words sequences, 395 words

were correctly identified by respondents and 124 were not. From Table 4.9, among the data (words) that underwent phoneme substitution, the sensitivity was 0.6301. The specificity of the study when there was no modification of the phoneme of the words was 0.7611. The overall sensitivity was 0.7222 and the misidentification was 0.2778.

TABLE 4.9: Respondents' identifications of the sampled data

		Different from Neighbour	Similar with Neighbour	Total
Respondent Feedback	As expected	138	395	533
	Not expected	81	124	205
Total		219	519	738

To further see the significant of the results, a test of statistical significance was also conducted. The chi square test was used again. In order to evaluate the frequency, the expected frequencies are presented in Table 4.10.

TABLE 4.10: Respondents' identifications of the expected data

		Different from Neighbour	Similar with Neighbour	Total
Respondent Feedback	As expected	158.1667	374.8333	533
	Not expected	60.83333	144.1667	205
Total		219	519	738

The value of χ^2 was 13.1627. For df=1 and p=0.05, χ^2 must equal or exceed 3.84 to be significant. Therefore it can be said that for phonemes substituted with matching phonemes in a specific context, the perception will be affected by the neighbouring words and is statistically significant.

4.3.2.2 Coda Evaluation

As with onset evaluation, the assumption was that when the coda of a word was substituted with a similar phoneme in a selected context, the perception will be affected by the neighbouring words. It was hypothesised that if such a condition happened, it should not happen due to chance. For synthesised speech, respondents identified 98 words that were affected by the neighbouring words and 76 words that were not. The control words (no modification made to those words) found that 353 words were correctly identified, and 113 were not.

TABLE 4.11: Respondents' identifications of the sampled data

		Different from Neighbour	Similar with Neighbour	Total
Respondent Feedback	As expected	98	353	451
	Not expected	76	113	189
Total		174	466	640

From Table 4.11, among the data (words) that underwent phoneme substitution, the sensitivity was 0.5632. The specificity of the study for control words was 0.7575. In total, the overall sensitivity was 0.7047 and misclassification was 0.2953.

To further see the significance of the test, a test of statistical significance was conducted. The expected frequencies are given Table 4.12.

TABLE 4.12: Respondents' identifications of the expected data

		Different from Neighbour	Similar with Neighbour	Total
Respondent Feedback	As expected	122.615625	328.384375	451
	Not expected	51.384375	137.615625	189
Total		174	466	640

The value of χ^2 was 22.9820. For df=1 and p=0.05, χ^2 must equal or exceed 3.84 to be significant. Therefore it can be said that for coda phonemes substituted with matching phonemes in a specific context, the perception will be affected by the neighbouring words and is statistically significant.

These experiments were also conducted using English synthesised speech. The feedback was expected to be also influenced by the machine generated speech and therefore the expected sensitivity and specificity were better than expected.

This showed that the confusion phones presented in Section 4.2.1 are applicable for use in phoneme substitution as long as they occur within the phoneme range listed in the confusion list. From the value of χ^2 for both onset and coda, it was believed that the coda might be better accepted as a substitution than as an onset where the changes (in the coda) were less frequently detected by the respondents.

4.4 Summary

The issues were investigated of reusing other resources to create another TTS albeit with the substantial chance of not having complete data, in particular the trained phonemes (of the diphones/phonemes) used in the resource language. The possibility of obtaining a substitute was investigated because it eliminated the need of new training or recording, as suggested by Kominek (2009). Without such, it is certain that synthesised speech will not sound native in the best possible situation and is distorted so as to be unintelligible. To avoid the worst case scenario, a study on possible substitutes was conducted on Malay and English. It was found that the respondents did best in perceiving synthesised speech according to what they believed was correct. Based on the findings from previous studies, intelligibility and perception tests based on simple listening and a contextual perception test were conducted. The results showed that intelligibility of substituted phonemes identified due to chance. For perception evaluation however, both onset and coda modifications can be perceived as intended as long as the context were given. The results obtained were tested using a chi square test, the intelligibility test not being found to be significant but the perception test being significant. It shows that phoneme substitution is possible if conducted with carefully selected substitutions given in a meaningful context.

Chapter 5

Prosody Processing for Combinational Speech Synthesis

This chapter discusses two prosody studies. The first is the evaluation of Malay prosody at the phoneme level and the second is the evaluation of Iban speech synthesis using Malay TTS. For the first study, the set of phonetically balanced phonemes of Malay speech was analysed and the phoneme values were compared to the Klatt duration threshold. This will then be compared to the value obtained by the Malay recording. The second study was based on the Iban synthesised speech using a Malay synthesiser. The study on the prosody contour between the two languages conveying the same speech in the respective languages will be presented. The purpose is to find similarities between the two as they share a similar language family root but are in different sub-classes.

5.1 Study on the Malay Phoneme

In this study, a recording of an hour of phonetically balanced text was analysed. The phonetically balanced text was taken from Tan (2008). The sound was evaluated independently. Each phoneme of the words was analysed and categorised and the duration and the fundamental frequency were extracted. For each phone studied, the corresponding classification applies:

- General Duration and Fundamental Frequency (F0) Value
- Duration and F0 Value based on Word Context
  - Duration of the phoneme at the word beginning
  - Duration of the phoneme at the word end
  - Duration of the phoneme at the middle word

  - F0 of the phoneme at the word beginning
  - F0 of the phoneme at the word end
  - F0 of the phoneme at the middle word

- Duration and F0 Value based on Syllable Context

  - Duration of the phoneme at the beginning of a syllable
  - Duration of the phoneme at the end of a syllable
  - Duration of the phoneme (vowel and diphthong) at the middle of a syllable
  - Duration of the phoneme (vowel and diphthong) at a syllable formed by a nucleus only
  - F0 of the phoneme at the beginning of a syllable
  - F0 of the phoneme at the end of a syllable
  - F0 of the phoneme (vowel and diphthong) at the middle of a syllable
  - F0 of the phoneme (vowel diphthong) at a syllable formed by a nucleus only

5.1.1 About the Data

This study was conducted on read texts. The speech was recorded in a recording room with a high quality microphone. The annotated text of the recording together with the recording were further processed so that the phoneme cluster of studies based on the phoneme's position in the word and in the syllable could be extracted. Only one speaker was used for the study so that a consistent style of speaking would be obtained.

The criteria of each phoneme is listed independently in the following figures.

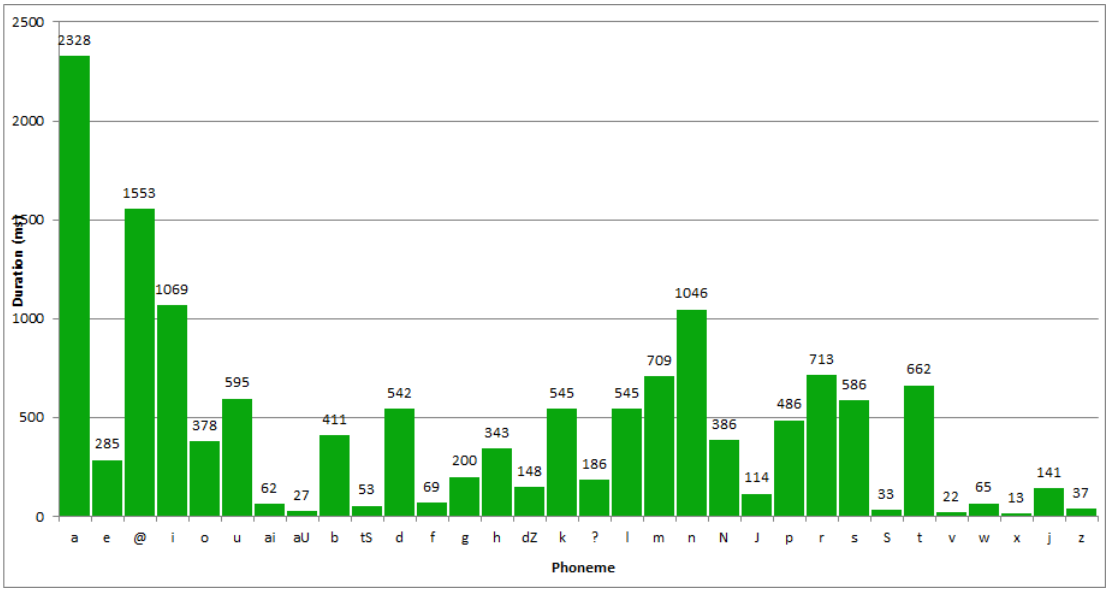

FIGURE 5.1: Individual Phoneme Frequencies

Figure 5.1 shows the frequencies of each phoneme occurrences in the sample size. This is approximately tallied to the distribution of the phoneme frequencies as presented by Khaw and Tan (2014) for phonetically balanced Malay phoneme frequencies.

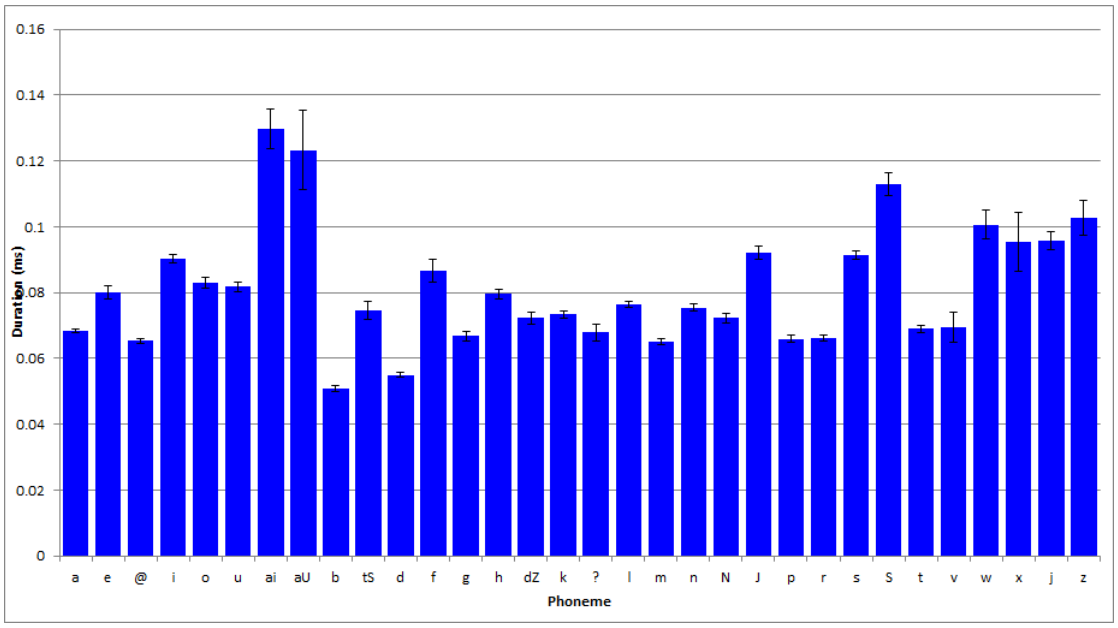

FIGURE 5.2: Mean duration values of overall phonemes with standard error

Figure 5.2 and Figure 5.3 show the overall duration for each phoneme irrespective of position, with standard error bars and standard deviation in the corresponding figures.

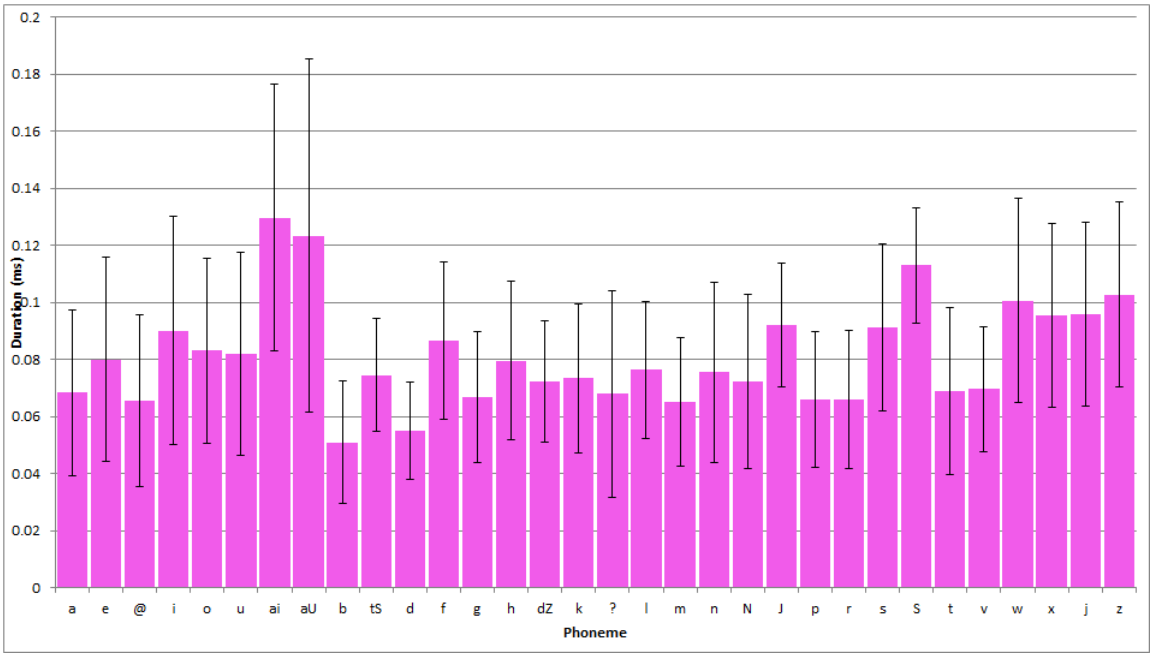

FIGURE 5.3: Mean duration values of overall phonemes with standard deviation

The F0 pattern for individual phonemes does not show any consistent pattern. Figures 5.4 and 5.5 show the summary of the mean of the F0 for the voiced phonemes with their respective standard error and standard deviation in the corresponding figures.

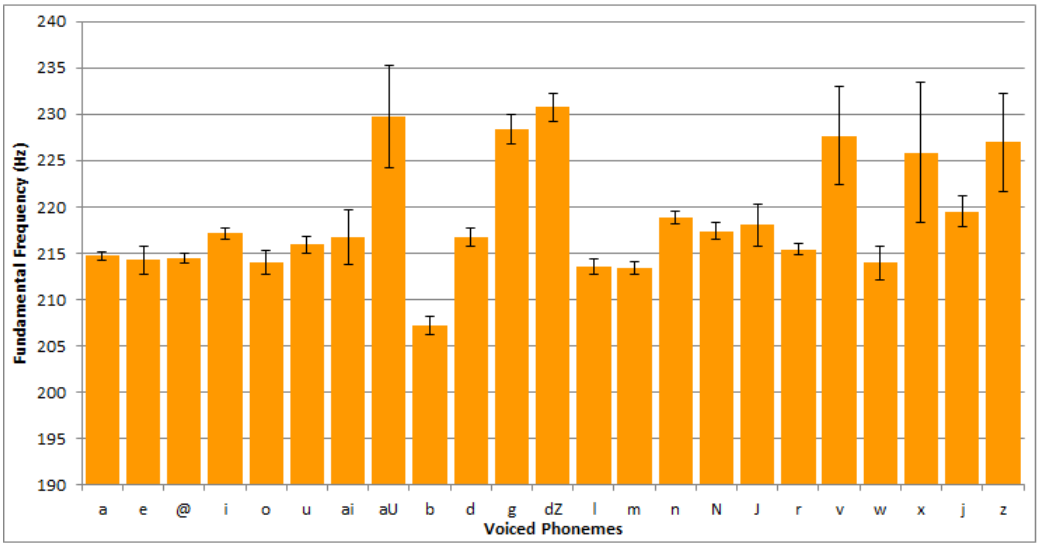

FIGURE 5.4: Mean F0 values of overall phonemes with standard error

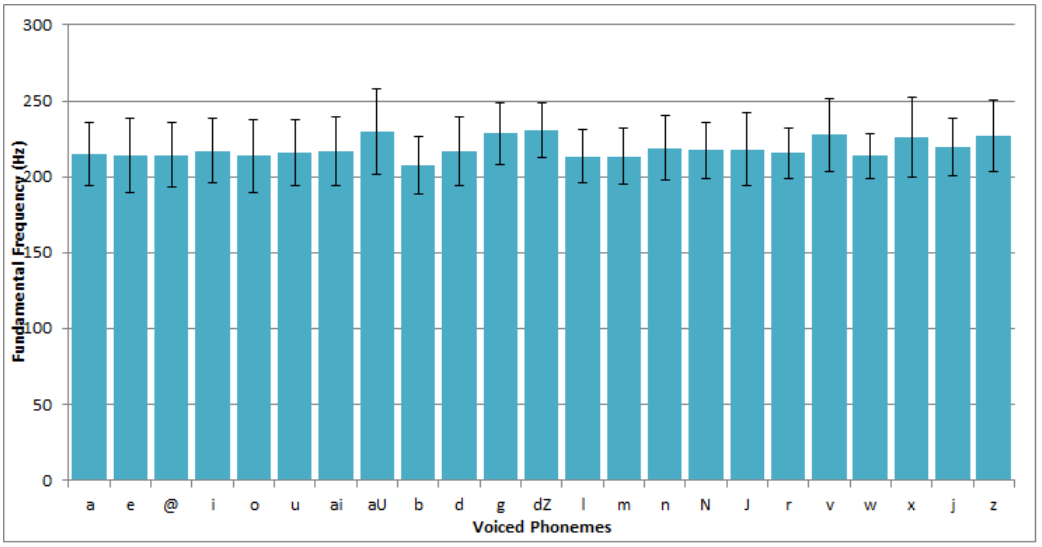

FIGURE 5.5: Mean F0 values of overall phonemes with standard deviation

5.1.2 Klatt Duration Model and Malay Duration Analysis

An extensive study on duration model has been provided by Klatt (1979). The Klatt's duration model has proven that segments have an intrinsic duration and an intrinsic compressibility, and that each factor has a multiplicative influence, with multiple factors combining to give an overall duration scaling.

5.1.2.1 Klatt Duration Model

Klatt's basis formula is:

$$\text{DUR} = \text{MINDUR} + (\text{INHDUR-MINDUR}) * \text{PRCNT}$$

in which
DUR = duration of the segment (calculated by the model)
MINDUR = minimum duration
INHDUR = inherent/intrinsic duration (built-in duration)
PRCNT = percentage of modification on top of the INHDUR over MINDUR difference value.

The duration of a segment is influenced by factors at many levels as described in Campbell (2000). At the lowest level, the category of segment and the category of neighbouring segments seem to affect its duration. The position of the segment in its syllable and the other constituents of the syllable are also important. Duration can also be affected by the position of the syllable in the prosodic foot and its position in the word. The position of the word or foot in the phrase can also have an effect, as can the selection of focus and pitch accents. Overall durations are also affected by speaking rate, which in itself can be affected by communicative context, style and emotion.

Some well known duration factors can be itemised as follows:

- Shortening of vowel durations before fortis consonants (pre-fortis clipping)
- Shortening of consonants in clusters
- Shortening in vowels in closed syllables
- Shortening of unstressed syllables
- Lengthening of accented syllables
- Foot-internal shortening
- Phrase-final lengthening

Due to the nature of Malay language - which is not a stressed language nor a tonal language, and the general rules of prominence are not fixed within a word or even a sentence, the pattern need to be studied in a slightly different manner Until such detailed study on the Malay language can be concluded. Due to this "looseness" of

the Malay language, a slightly different comparison of Klatt's standard duration model against the Malay recorded speech was conducted.

However, it may suffice to summarise the Klatt duration model rules as follows:

1. **Pause insertion rule**: insert a 200ms pause before each sentence internal main clause and at boundaries delimited by comma, but not before relative clauses.
2. **Clause-final lengthening**: the vowel or syllabic consonant in the syllable just before a pause is lengthened by PRCNT=1.4. Any consonants between this vowel and the pause are also lengthened by PRCNT=1.4.
3. **Non-phrase-final shortening**: syllabic segments are shortened by PRCNT=0.6 if not in a phrase-final syllable. A phrase-final postvocalic liquid or nasal is lengthened by PRCNT=1.4.
4. **Non-word-final shortening**: syllabic segments are shortened by PRCNT=0.85 if not in a word-final syllable.
5. **Polysyllabic shortening**: syllabic segments in a polysyllabic word are shortened by PRCNT=0.8.
6. **Non-initial consonant shortening**: consonants in non-word initial positions are shortened by PRCNT=0.85.
7. **Unstressed shortening**: unstressed segments are half again more compressible than stressed segments (i.e. MINDUR=MINDUR/2). Then both stressed and secondary-stressed segments are shortened by a factor depending on segment type: syllabic in word medial syllable PRCNT=0.5; syllabic in other positions PRCNT=0.7; prevocalic liquid or glide PRCNT=0.1; all others PRCNT=0.7.
8. **Lengthening for emphasis**: an emphasized vowel is lengthened by PRCNT=1.4.
9. **Postvocalic context of vowels**: the influence of a postvocalic consonant or sonorant-stop cluster on the duration of a vowel is given below. The consonant must be in the same morpheme as the vowel and be marked as unstressed. In a postvocalic sonorant-obstruent cluster, the obstruent determines the effect on the vowel and on the sonorant. Open syllable, word-final PRCNT=1.2; before a voiced fricative PRCNT=1.6; before a voiced plosive PRCNT=1.2; before an unstressed nasal PRCNT=0.85; before a voiceless plosive PRCNT=0.7; all others PRCNT=1.0. If non-phrase final, change PRCNT to 0.7+0.3*PRCNT.
10. **Shortening in clusters**: segments are shortened in consonant-consonant sequences (disregarding word boundaries, but not across phrase boundaries), and are

also modified in vowel-vowel sequences. Vowel followed by a vowel PRCNT=1.2; vowel precedes by a vowel PRCNT=0.7; consonant surrounded by consonants PRCNT=0.5; consonant preceded by consonant PRCNT=0.7; consonant followed by consonant PRCNT=0.7.

The duration as proposed in Klatt's duration model was compared against the Malay speech prosody extracted from natural speech.

5.1.2.2 Klatt Segment Duration and Malay Segment Duration Analysis Comparison

Table 5.1 and Table 5.2 show the phoneme inherent duration of Klatt's duration model, the minimum duration allowed in the Klatt synthesiser, the mean and median (and most of the time the mode value as well) of the duration extracted from natural speech of Malay text.

TABLE 5.1: Vowel and Diphthong Duration: Klatt vs Malay Natural Speech

Phoneme	**Klatt Inherent Duration**	**Klatt Minimum Duration**	**Beginning Duration**		**Middle Duration**		**Final Duration**		**Standalone Duration**	
			μ	M	μ	M	μ	M	μ	M
a	230	80	65	60	62	60	72	70	87	80
aː	240	100								
e	150	70	113	110	71	65	82	80	101	100
ə	120	60	64	60	52	50	68	60	85	70
əː	240	80								
i	135	40	98	105	73	70	90	80	136	130
iː	155	55								
o	240	130	97	110	75	70	89	90	83	80
oː	240	130								
ʊ	210	70	83	80	69	80	83	80	124	140
uː	230	150								
ai	250	150	370		141	130	119	113	140	130
ou	220	80								
aʊ					175	175	-	110	87	90

From both of the tables, the duration for Malay speech is mostly still within a close range of Klatt's synthesiser minimum duration, except for in a few cases, e.g. vowel /a/ and /o/. The purpose of showing the median value is only to double check that the mean value is not vastly different from the median point. The study also showed that

the median point almost always shares the same value with the mode of the phoneme duration.

This study was not able to conclusively demonstrate a similarity (or lack of) in duration pattern between the two languages. The tables show that more than half of the phoneme list met the range of duration proposed by Klatt's duration model. It loosely fits between the inherent and the minimum of the duration model. The study omitted other factors, such as the phrasal prosody effect, the syllable strengthening effect as well as the rhythmic effect which were the main reason for the Malay and Iban pairing which will be presented in a later section. However, for the purpose of this study, it suffices to use Klatt as a comparison benchmark.

Table 5.2: Consonant Duration: Klatt vs Malay Natural Speech

Phoneme	**Klatt Inherent Duration**	**Klatt Minimum Duration**	**Onset Duration**		**Coda Duration**	
			μ	M	μ	M
b	85	60	50	50	53	50
tʃ	70	50	70	70		
d	75	50	55	50	50	45
f	100	80	87	80	87	80
g	80	60	67	60	48	50
h	80	20	80	80	80	70
dʒ	70	50	71	70	99	75
k	80	60	73	70	47	50
l	80	40	74	70	87	80
m	70	60	61	60	75	70
n	60	50	66	60	79	70
ŋ	95	60	70	70	74	70
ɲ	95	60	70	70	74	70
p	90	50	66	60	73	70
r	80	30	62	60	73	70
s	105	60	90	90	98	90
ʃ	105	80	114	110	100	100
t	75	50	68	70	72	60
v	60	40	70	70		
w	80	60	109	110		
x			104	100	99	100
ʔ					70	60
ʕ			61	60		
j	80	40	97	100	57	50

The Klatt duration model consists of inherent duration and the duration was then manipulated based on the characteristics of the phoneme's segment at the phonetic context and phrasal position level. This manipulation was based on the ten rules describing the Klatt duration model (Klatt, 1987).

With an engine produced by Goh (2004), the speech synthesiser was constructed using English diphone and English prosody rules following the Klatt duration model. The outcome was fair where the quality of the synthesised speech was sufficiently intelligible (sounding like an American speaking Malay). Another pilot study was also conducted using Indonesian speech synthesis, using Indonesian prosodic rules with a tweak on grapheme-to-phoneme rules. The quality was good even though the synthesised speech still sounded Indonesian due to the penultimate stressing in Indonesian. Both of the studies used diphone synthesiser engines and were not formally tested with a sufficient number of respondents.

These pilot tests showed that even when very close language features were used in the synthesiser, the synthesised speech will still sound foreign. Klatt's duration model may be able to provide a range of acceptable duration times; however there are more varied parameters involved. The next section will review the pitch and intensity contour between two different languages but with similar rhythm: Malay and Iban.

5.2 Comparing Malay and Iban Speech Contour

Iban and Malay both belong to Malayo-Polynesian sub-language categories before they diverged into their corresponding branches. Other than having the same family root, Malay and Iban both have the following language family typology: Austonesian - Malayo-Polynesian - West/Nuclear Malayo-Polynesian - Malayic. The divergence occurs at the following branch where Iban is categorised under Malayic-Dayak and Malay is under Malayan. Being in the same sub-language may not indicate a lot, however one may say that some of the languages have similar rhyme, although one language may sound closer to another. For example, when a non-native Tagalog speaker hears a Tagalog conversation, they might mistake the language as Malay or Thais. This might be due to the melody of Tagalog itself.

As with Iban and Malay, there was no comparison study so that one can say one language is closer to another. One very prominent reason would be that Iban is a stressed

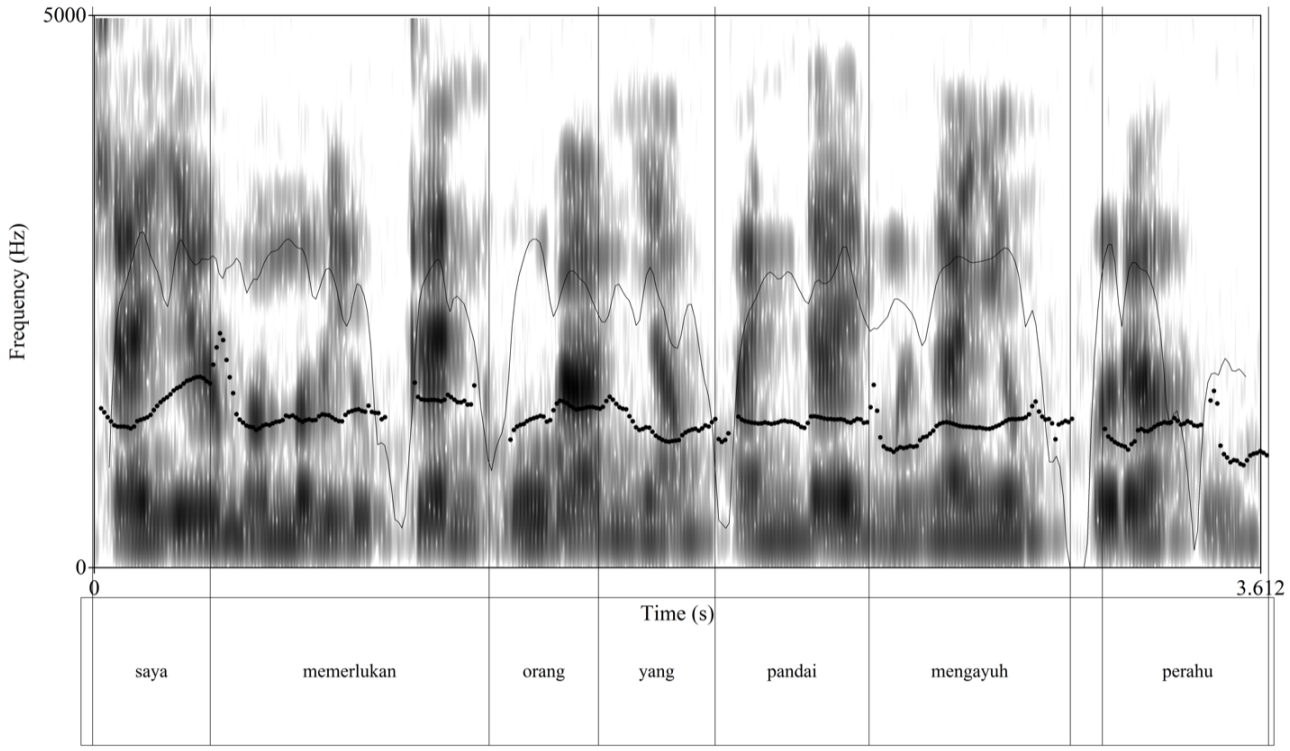

FIGURE 5.6: Native Malay saying the translation of: *"aku beguna ka orang ke nemu ngemudi ka perau"* in Malay or "I need someone who can row a boat". The corresponding Iban is shown in Figure 5.10.

language, while Malay is not. However, being from the same root and spoken geographically close, it is understandable that the languages can and do reflect one another to a certain degree.

This similarity may not be tangible. However, some patterns can be found in the translated version of the Iban text which was used for evaluation of similarity in the following section. Figure 5.6 shows the pattern of pitch and intensity contour as well as the annotation of the speech by a native Malay speaker of the translated version of the Iban sentence. The corresponding Iban speech can be viewed in Figure 5.10.

The recording of Malay speech was not done in an ideal recording environment. It was done in a open ventilated room with a built in microphone to a laptop as compared to the low level recording room with a microphone attached to a computer for the Iban speaker. The corresponding text for Figure 5.7 is shown in Figure 5.12. From both of the sets, the contour of the intensity and the pitch are close. They both have steady pitch contour while the intensity contour has falling and rising according to the word

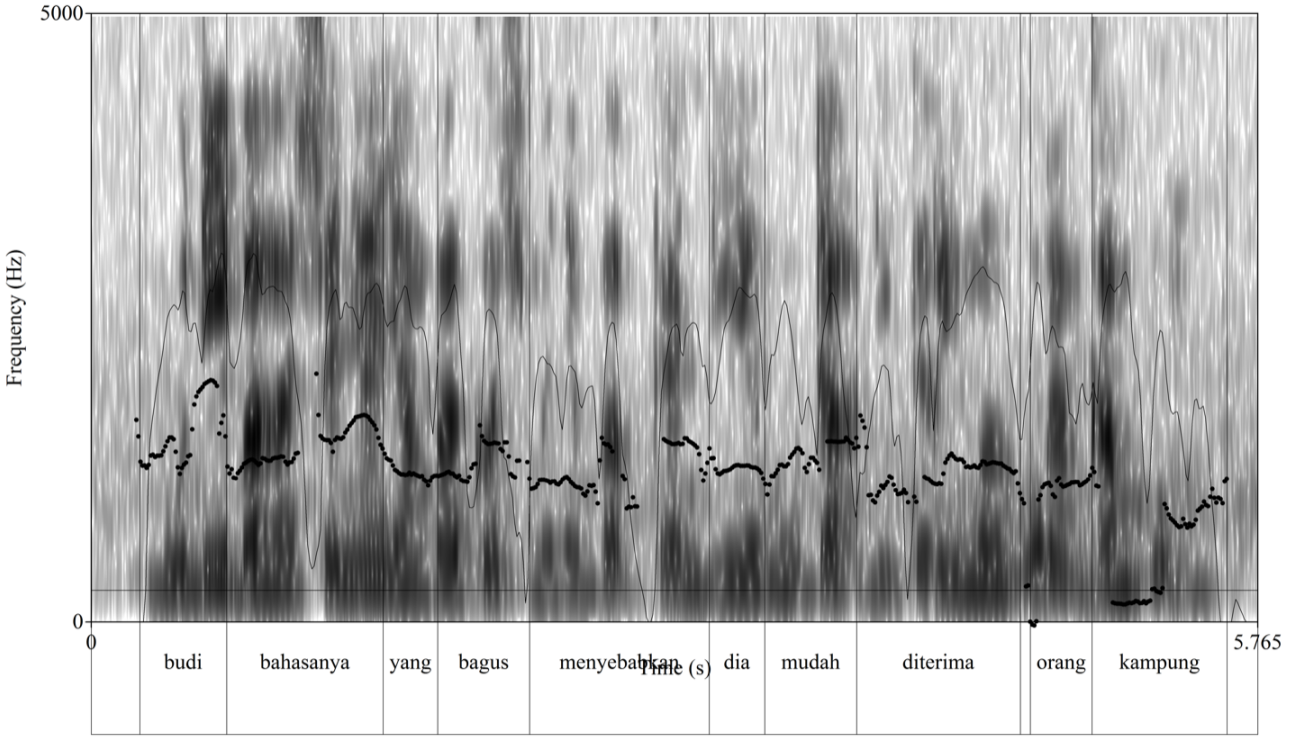

FIGURE 5.7: Native Malay saying the translation of: *"pendiau enggau pemanah iya lalu lengkas diterima bala maioh"* in Malay or "His wonderful manners made him immediately liked by the locals". The corresponding Iban is shown in Figure 5.12.

lexicon. The pattern however is very different for the pitch contour between the third set: Figure 5.8 and Figure 5.14. This is the contour for an exclamation sentence of which there are a few ways of delivering the sentence.

Despite having differing sentence length, both languages showed close duration agreement. Again, as stated at the beginning of this section, this may not be significant; however, this requires a more detailed study.

5.3 Study of Similarity

The previous section has shown the speech features which are visible in the spectrogram and the comparison with the corresponding Iban text. While the method was not conclusive, it makes it possible to view the pattern of rising and falling of the pitch contour and the intensity contour while also showing the general duration of the speech time. In order to show the similarity between the two languages, this section provides a

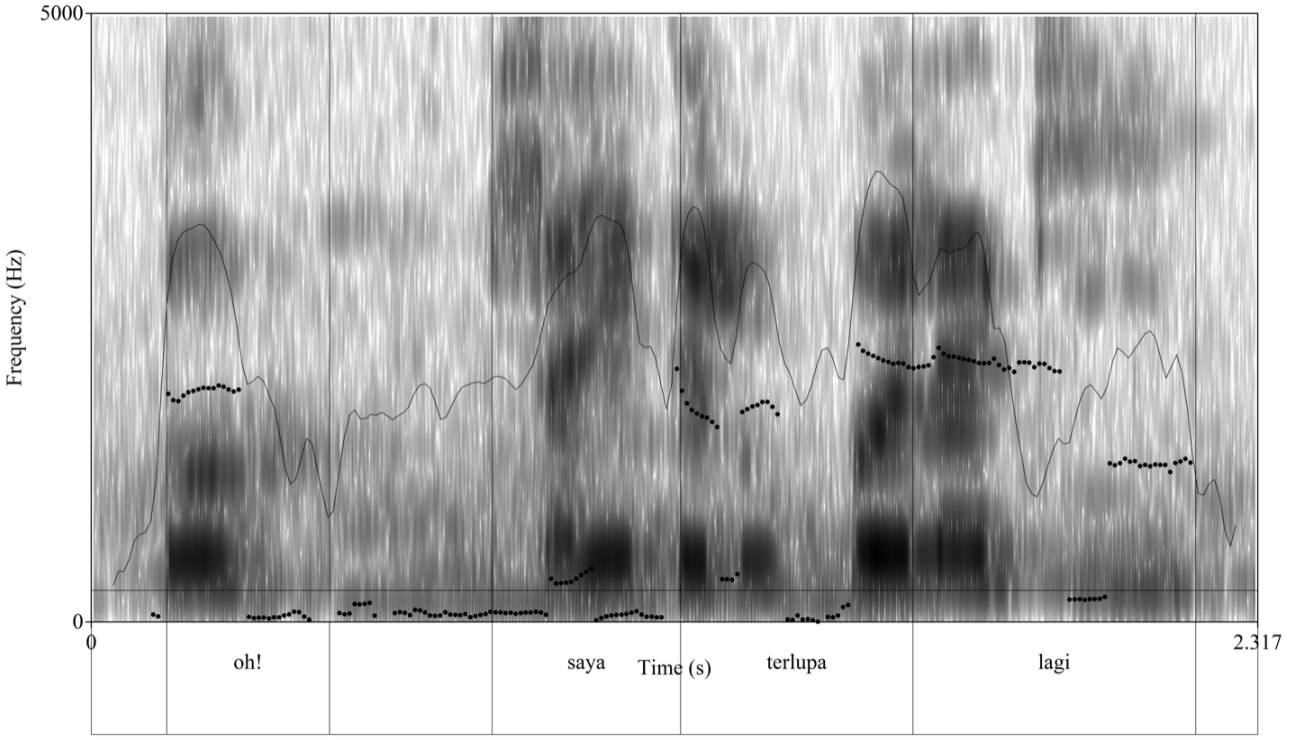

FIGURE 5.8: Native Malay saying the translation of: *"oh, aku enda ingat baru ga!"* in Malay or "Oh, I forgot again!". The corresponding Iban is shown in Figure 5.14.

comparison between two similar languages, using a different prosody model, one trained by Malay speech and another one is the native Iban speaker itself.

5.3.1 Prosody Comparison Study

A set of Iban sentences were synthesised using the HMM-based Malay synthesiser. No tuning was done to the original TTS or the speech data other than adding the Iban words into the pronunciation dictionary. Since the Malay TTS is constantly undergoing improvement, it relied on the pronunciation dictionary to provide an accurate pronunciation. This is suitable for the purpose of this study: to determine that by using the statistical method (HMM-based synthesiser) of the Malay model, an acceptable quality of Iban speech synthesiser can be produced.

This analysis of this experiment is similar to the one conducted by Dusterhoff and Black (1997) which was also duplicated and augmented by Santen and Hirschberg (1994) and Silverman et al. (1992). In Silverman et al. (1992), the focus of comparison was on

the pitch accent sentences. Santen and Hirschberg (1994) added more components in their comparisons. The parameters were slightly different than the those used by Dong et al. (2007). The following were observed by the three researchers when analysing the prosody of the syllables.

The following prosody criteria were observed by Silverman et al. (1992), Santen and Hirschberg (1994) and Dusterhoff and Black (1997) in their corresponding prosody studies.

- number of syllables within the phrase
- stressed syllables within the phrase
- accented syllables proceeding the syllable, within the phrase
- accented syllables succeeding the syllable, within the phrase
- distance (in syllable units) from the previous syllable to the next events
- number of non-major phrase break since the last major break
- onset length of the syllable
- rhyme length of the syllable

These criteria were observed by Santen and Hirschberg (1994) and Dusterhoff and Black (1997) as well.

- percent of the syllable which is unvoiced
- position of the syllable within a word
- two syllable window on either side
- two neighbouring syllable accentedness
- two neighbouring syllable lexical stress
- two neighbouring syllable onset type
- two neighbouring syllable coda type

To find more information about the difference between two contours and the closeness between two contours, Dusterhoff and Black (1997) also observed these feature.

- tilt event type
- syllable break values.

Mandarin is a tonal language, in which each character carries a tone. Based on Dong et al. (2007), tones exhibit as patterns of pitch contour from the acoustic point of view and rhythm exists as prosodic unit groups. Prior research by Dong et al. (2007) found that

the existence of a prosodic word, which is a phenomenon that speech units are usually grouped into small prosodic units normally consisting of two to three syllables. At the acoustic level, the prosodic word boundary is usually presented as duration, pitch change, and energy change. Based on their experience, tone and prosodic word groups affect the naturalness of Chinese speech very much. Therefore, the defined parameters should address these two important aspects. Originally, a set of 40 features were observed and then simplified into 12 clusters of features. Dong et al. (2007), compared the acoustic divergence and the similarity between two corpora. Since the goal is for automatic language learning, the observation was carried out on the syllable features themselves, and not directly involving the neighbouring syllables. The syllable features observed by Dong et al. (2007) are as follows:

- duration of the syllable
- duration of initial part of the syllable and final part of the syllable
- pitch mean, pitch range
- pitch start, pitch middle, pitch end
- energy centre (position that divides energy into half)
- RMS energy
- start energy and end energy (50ms of the beginning and end of the syllable)

5.3.2 Constructing Iban TTS Speech without Speech Data

The experiments described by the researchers in Section 5.3.1 were conducted for stressed and tonal languages. For such, the pitch, the contour and other features were directly seen by the pitch change, duration and the energy changed as the features mentioned. However, the rhythm exists as prosodic unit groups as described by Dong et al. (2007). Therefore it may be difficult to dissect the rhythm from the stressed or tonal criteria.

When constructing Iban words, advice was obtained from the researchers from Universiti Malaysia Sarawak (UNIMAS). Iban is a stressed language, and its typology is a quantity insensitive stress (Gordon, 2002; Baughman, 2012). Although the stress does not always determine a different meaning, like the English word <minute>has a different stress (and phones) when it is a noun /'minit/ than when it is an adjective /maɪ'nyut/.

When constructing Iban speech using a Malay TTS, there are a few comparisons conducted to test the similarity on the phonetic level. Both languages do not have the same whole phoneset. The following is the Iban phoneset as compared with Malay:

- phonemes used by both languages: /p/, /b/, /m/, /w/, /t/, /d/, /n/, /tʃ/, /dʒ/, /s/, /l/, /r/, /ɲ/, /j/, /k/, /g/, /ŋ/, /h/, /ʔ/, /a/, /e/, /ə/, /i/, /u/, /ai/, /au/
- phonemes used by Iban only: /o/, /ui/, /ia/, /ea/, /ua/, /oa/, /iu/, /iə/, /uə/, /oə/
- phonemes used by Malay only: /ɔ/, /f/, /v/, /z/, /x/, /ɣ/, /ʃ/, /ʕ/

However, as mentioned before, the Malay TTS is still being revised and is an ongoing development. Some of the phonemes cannot be recorded by the system. The phones which can be studied were those presented in Figure 5.1.

An experiment was conducted on syllabification where a collection of Iban sentences were run through Malay syllabification. The Malay syllabification worked well for Malay words but did not work for all recently identified loan words. The program also worked from the grapheme level. When the Iban words were put into the syllabification system, the syllabification system could syllabify all Iban words correctly.

Finally, a set of Iban texts were synthesised using the Malay TTS. To ensure the use of the Malay phone set and Malay prosody, no additional training of the Iban recorded speech was added into the synthesiser. The only Iban related item was the pronunciation dictionary which consisted of only the words and their corresponding phonetic transcription. It is also important to highlight that the Malay is not a stressed language.

5.3.2.1 TTS Data and the experiment

For training and testing the data, the original recording included 16 hours of recording from 16 people. The recording was conducted at the Univerisiti Sains Malaysia, Penang, Malaysia with mixtures of gender, age and ethnicity. However, when running the TTS, the synthesised speech did not reach an acceptable condition. Several hours of speech recording were added with additional data provided by the Nanyang Technological University, Singapore. The synthesised speech reached an acceptable quality when the recording sample reached 130 hours of voice recording.

A few Malay phonemes were not covered by the Malay HTS system due to training data limitation. The phonemes were: /x/, /ɣ/ and /ʕ/. The phonemes /ɣ/ and /ʕ/ are used only in Arabic loan words and, in normal conversational Malay, they had already been simplified into /g/ and /ʔ/ respectively. However the phoneme /x/ was frequently used but for this synthesiser, it was simplified as /k/ instead.

The Malay TTS uses a pronunciation dictionary and a HMM synthesiser. Therefore, to produce the Iban synthesiser, an Iban pronunciation dictionary needed to be created. It was assumed that the synthesiser could produce Iban speech, however it was also believed that the quality of the speech would be inferior as compared to the Malay synthesised speech due to the non-existent Iban speech data (in the training set) and thus the unavailable Iban prosody information provided inside the TTS. However, it is also believed that since the rhythm of speech between the two languages is not so much different, if is plausible that the TTS quality may be intelligible and possibly acceptable to the respondents.

5.3.2.2 Measure of Accuracy

It is difficult to follow closely the method used by Dusterhoff and Black (1997), Black and Hunt (1996) and Ross and Ostendorf (1999) due to the type of language used in these experiments not being the same while the others studied stressed languages. In order to have some measure of accuracy, the root mean squared error (RMSE) between the generated contour (using Malay TTS) and the original recorded Iban speech was calculated. The correlation between the generated speech and the original recording was also used. Since the pitch is generally dependent on the gender of the speaker and both the synthesised speech and the recorded speech were of a female voice, it is expected that the F0 deviation might be larger. It is supposed that the RMSE indicates the characteristic divergence between the two sound waves for the given features while correlation indicates the similarity across two sound waves. Since Malay is not a stressed language and Iban is, the parameters observed also included energy. Observations also included the first formant (F1), second (F2) and third (F3) which, for at least for the first formant, the correlation was expected to be very high since the shape of the contour will not differ much. The observation was also carried out to see if there were any further conclusions that could be made.

5.3.2.3 RMSE and Correlation

Thirty sentences were recorded and synthesised for the use of these experiments. The synthesised speech and pre-recorded speech were compared by using the following parameters: duration, F0 mean, minimum F0 of the segment, maximum F0 for the segment, mean of the energy, minimum energy in the segment, maximum energy of the segment,

and F1, F2 and F3 of the segment. Since the nature of the Malay and Iban languages are similar in perception but are not the same in term of language characteristics, the analysis was conducted using all - phoneme, syllable and word segments.

5.3.2.4 Phonemes

Analysis on the recorded and synthesised speech based on the syllable segments was carried out. The result is as shown in Table 5.3. It was expected that all RMSE will be high because the sound waves were produced by different speakers, in different styles and one of them is a statistical control. However, it was expected that, at least, the correlation for the formant (the first two) and the energy will be similar.

TABLE 5.3: RMSE and Correlation for Phoneme

	Duration	**F0 mean**	**F0 min**	**F0 max**	**Energy mean**	**Energy min**	**Energy max**
RMSE	54.71	37.92	35.91	51.25	6.54	11.00	6.24
Corr	0.4919	0.4395	0.4183	0.4462	0.6218	0.6445	0.4983

	F1	**F2**	**F3**
RMSE	189.70	298.42	303.34
Corr	0.6784	0.6944	0.4328

For phoneme segments, the duration and F0 showed that there was only a slight correlation between the recorded Iban speech and the synthesised Iban text. From Table 5.3, the RMSE values show a large deviation between the recorded speech and the synthesised speech for the duration, the fundamental frequencies values and the formant values. However, the energies showed that the differences between the two were not as large, and one can see that the energy mean and energy minimum have better correlations than the comparative durations and fundamental frequencies. The correlation for the first formant and second formant were even better. This is expected since the formant shape on the spectrogram was very close from one to another when representing the same sound despite the speaker characteristics and language. For voiceless consonants, the fundamental frequencies were assigned 0. The RMSE and correlation measurement were further observed for longer segments: its syllable.

5.3.2.5 Syllables

The same sound files were further analysed based on its syllable segments. Comparing between the two: Tables 5.3 and refSyllRMSECorr, the RMSE values increased for

some features but decreased for others. The RMSE increased for duration, maximum F0 range, minimum energy and the third formant. The RMSE decreased for F0 mean and minimum F0 range, the mean energy and the maximum energy used and the first and second formants. Both sound waves were generated by female voices. Female voices are known to have a higher frequency range than male voices. This difference showed in both language tones: Malay and Iban had close speech rhythm for both the pitch and energy. However, the pattern was not fixed since the pitch and the energy for both synthesiser and recorded speech correlation values were not strongly correlated.

TABLE 5.4: RMSE and Correlation for Syllable

	Duration	**F0 mean**	**F0 min**	**F0 max**	**Energy mean**	**Energy min**	**Energy max**
RMSE	71.64	35.77	33.34	59.05	3.84	12.49	4.23
Corr	0.738138	0.353044	0.48169	0.245726	0.3073	0.72054	0.165326

	F1	**F2**	**F3**
RMSE	128.50	213.05	231.4
Corr	0.751509	0.791706	0.458977

From Table 5.4, the correlation values also did not get any better than the phoneme's correlation except for the correlations for the first formant and second formant. The minimum energy used was close for the two sound waves. The duration showed high correlation despite having higher RMSE than did phonemes. This was expected since the error will increase with the increase of the size of the segments. The F1 and F2 for the syllable also showed close correlation because the syllable's formant could better represent the formants than when it is in a phoneme segment where the consonants' segments were compared individually between the recorded and synthesised speech. When the comparison covered the whole syllable, the modification on the adjacent consonants of the vowel could better reflect the contour of the formants.

5.3.2.6 Words

From the word perspective, the correlation generally showed fewer patterns than syllables except for a slight rise in the correlation of the duration. The RMSE showed the similarity was better for the mean of the F0 and the mean of the energy for compared words. This was also the case for all the three formants. This showed that pattern similarity variations occur in phonemes and syllables more than in words. This also means that words may not be the best segment to be used to evaluate the difference between

the prosody used in Malay after being implemented in Iban to the Iban recorded speech itself.

Table 5.5: RMSE and Correlation for Word

	Duration	**F0 mean**	**F0 min**	**F0 max**	**Energy mean**	**Energy min**	**Energy max**
RMSE	121.86	33.28	35.07	64.49	3.12	14.49	3.60
Corr	0.754149	0.445538	0.470262	0.287832	0.44598	0.601278	0.296505

	F1	**F2**	**F3**
RMSE	104.99	169.03	181.51
Corr	0.58161	0.761607	0.40411

5.3.3 Perceptual Evaluation

As measured by accuracy, words showed reduced of RMSE values for some features. From syllable correlation and RMSE, the correlation showed the duration, energy and the formants were closely similar to the recorded speech. We find it important to evaluate the perceptual acceptance among native speakers.

In this perceptual test, expert respondents were asked to rate the quality of Iban polyglot speech synthesis. A five scale rating has been given in which they range from very good to very poor. The rating range is as follows:
5-very good
4-good
3-fair
2-poor
1-very poor

The following windows showed the best rated sentence (rated 5), fair rated sentence (rated 3) and poorly rated sentence (rated 2). Each frame will show the synthesised speech signal and their corresponding annotations and followed by the recorded speech signal and annotations. The thin line in the spectrogram represents the F0 contour and the dotted line represents the energy contour. Respondents have to give a rating in a scale of one to five. The following sentence showed one of the highest ratings. For example, in the following two windows, comparative recording of an Iban native speaker speech analysis is compared to the Iban polyglot speech synthesiser. Figure 5.10 showed

the comparative recording for the same sentence by a native speaker. For the entire Iban recording, the speaker was a female native speaker while the Malay synthesiser generated a female voice as well.

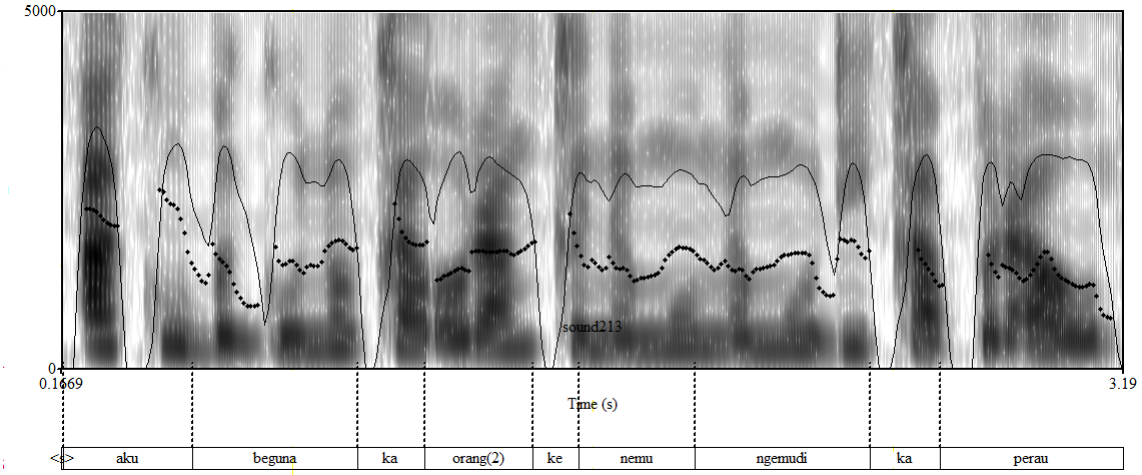

FIGURE 5.9: Malay TTS synthesising Iban text: *"aku beguna ka orang ke nemu ngemudi ka perau"* or "I need someone who can row a boat"

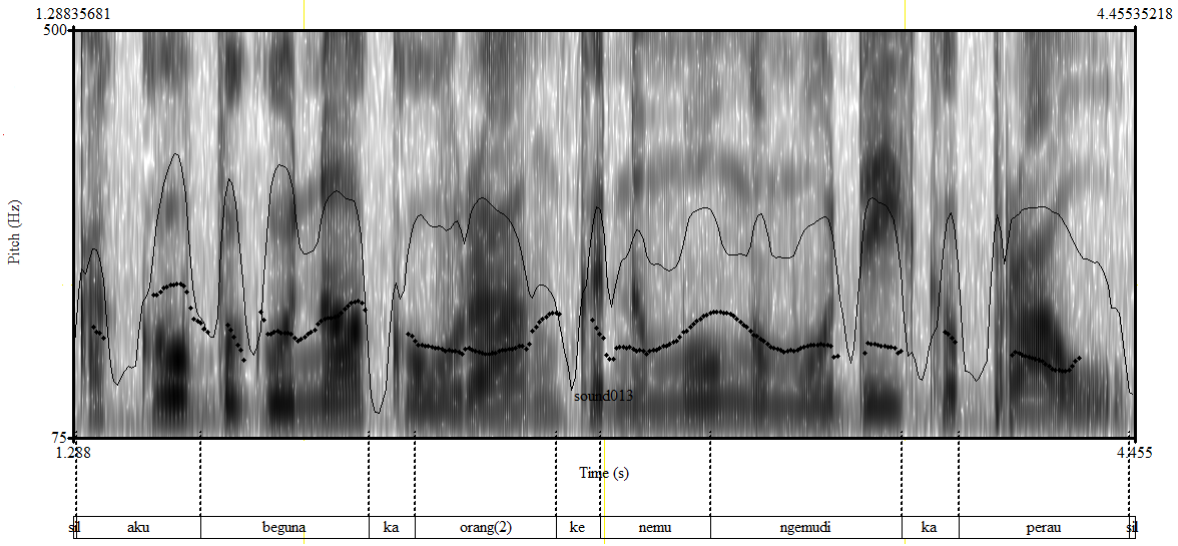

FIGURE 5.10: Native speaker of Iban saying: *"aku beguna ka orang ke nemu ngemudi ka perau"*

For Figure 5.9 and Figure 5.10, the pitch for the synthesised speech fluctuated across the sentence while the recorded speech had a steady contour. The energy used was about the same in terms of contour while the duration of the segment looks very closely similar. The speech segment (syllable) features can be itemised as Table **??**.

The label *Dur* refers to the duration and *Ene* to the energy. The RMSE and the correlation of this sentence is provided in Table 5.7. There was no definite value which

TABLE 5.6: Comparison between the synthesised and recorded speech syllables features. The left is the synthesised speech and the right is the recorded speech

Syll	Dur	F0	Ene	F1	F2	F3
a	161	254	79	1059	1766	2896
ku	211	252	76	512	1277	3000
b@	141	189	77	340	1556	2918
gu	161	184	76	328	1335	2835
no	171	210	77	344	1308	2804
ka	191	235	75	778	1547	2878
u	100	187	77	481	1098	2860
raNG	211	212	77	549	1530	2532
k@	131	236	71	744	1838	3105
n@	151	196	76	380	1759	2855
mu	181	200	77	326	910	3011
NG@	141	196	75	383	1257	2882
mu	221	198	77	315	1060	2945
di	141	204	74	278	2519	3250
ka	201	194	74	874	1653	2867
p@	141	207	73	444	1657	3015
raw	382	185	77	518	1405	2678

Syll	Dur	F0	Ene	F1	F2	F3
a	70	186	72	871	1544	2263
ku	301	219	78	597	1371	2730
b@	80	195	75	515	1434	2831
gu	190	182	79	474	1178	2670
no	231	197	77	585	1422	2938
k@	140	180	70	598	1178	2552
u	120	168	76	471	919	2573
raNG	301	174	76	666	1390	2495
k@	130	202	73	761	1910	2977
n@	130	165	73	446	1708	2993
mu	201	181	75	412	966	2856
NG@	150	195	75	434	1019	2851
mu	221	168	75	408	1076	2733
di	201	168	76	435	2353	3027
k@	170	180	71	680	1609	2728
p@	120	200	71	386	1590	2636
raw	391	153	75	620	1255	2682

can be used as a reference for what is the best other than the one stated in Huang et al. (2001) which stated, in a controlled duration and phoneme identity, the measurement for male speech over a long sentence with a RMSE of the pitch of 15Hz or less and correlation of the pitch of 0.8 or above indicated quality that may be close to perceptually identical to the natural reference utterance. However, it was also stated that such exactness is useful only during training and testing and cannot be expected during training on entirely new utterances from random text. This experiment on the other hand used the Malay speech resources to produce a different language. However, by comparing to the mean results of Dusterhoff and Black (1997), the correlations obtained were quite close. For the correlation of boundary models, Dusterhoff and Black (1997) obtained 0.778 for the duration, 0.530 for the F0 and 0.408 for the energy.

TABLE 5.7: RMSE and Correlation for "*aku beguna ka orang ke nemu ngemudi ka perau*".

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	49.49747	31.33876	26.07568	44.18211	2.818009	11.99755	3.038963
Corr	0.807267	0.546372	0.532278	0.437538	0.336118	0.754804	0.190286

	F1	F2	F3
RMSE	132.4262	157.6954	246.6741
Corr	0.853744	0.936078	0.466062

Figure 5.11 and 5.12 showed the synthesised and recorded speech signals in which the synthesised speech was rated as fair (scale 3). From the speech signal features, the RMSE and correlation have a pattern which is not very different than the one rated highly as shown in Table 5.7. However, the reason for the fair rating was because of the mispronunciation of one of the keywords of the sentence which resulted in one part of the sentence being understood while the second part was not. The other reason raised was that the phrase breaks were positioned at the slightly odd places resulting in some of the respondents finding it necessary to listen to the sentence a couple of times before it could be understood.

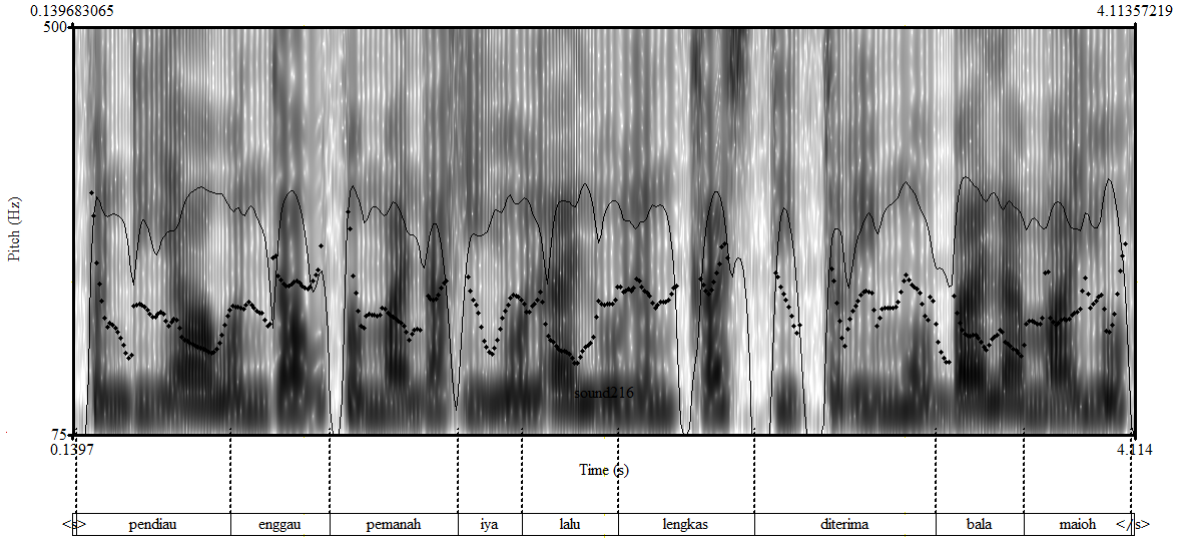

FIGURE 5.11: Malay TTS synthesising Iban text: *"pendiau enggau pemanah iya lalu lengkas diterima bala maioh"* or "His wonderful manners made him immediately liked by the locals"

Comparing between the pitch contour and the energy, Table 5.11 shows more fluctuating energy than the recorded speech in Table 5.12. This may not be affecting much (like the one showed in Table 5.9 as compared to Table 5.10). However, it is believed that the low rating were due to the incorrect position of the phrasal break. In the recorded speech, with the sentence: "*Pendiau enggau pemanah iya lalu lengkas diterima bala maioh*" ("His wonderful manners made him immediately liked by the locals"), the phrasal break occurred after *iya* (him), a short break occurred after *lengkas* (fast) and *diterima* (accepted) while the synthesised speech pauses were at *enggau* with a short break, *lengkas* with a phrasal break and *diterima* with a short break.

From Table 5.9, the mean duration's correlation is slightly lower than that in Table 5.7. Other than that, the correlations between the two are fair.

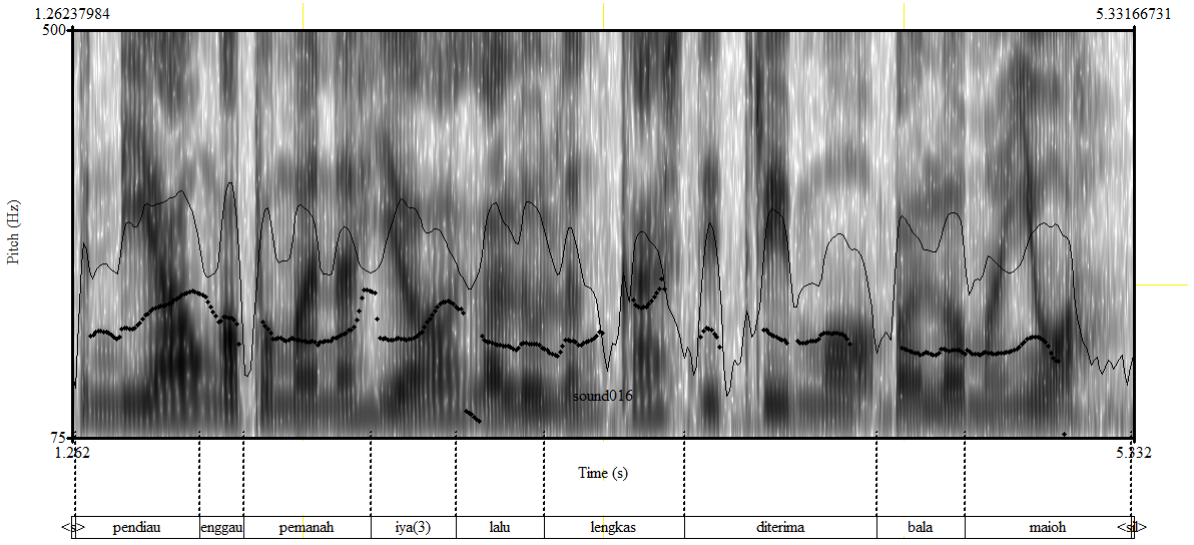

FIGURE 5.12: Native speaker of Iban saying: *"pendiau enggau pemanah iya lalu lengkas diterima bala maioh"*.

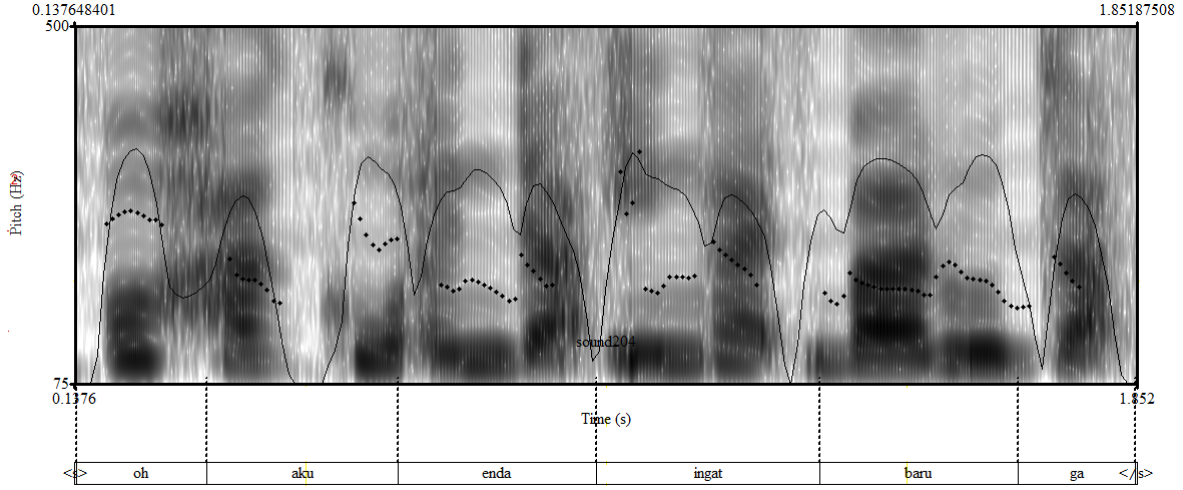

FIGURE 5.13: Malay TTS synthesising Iban text: *"oh aku enda ingat baru ga"* or "Oh, I forgot again"

Figure 5.13 showed a sentence rated as poor. This is an exclamative sentence. There are two phrases: *enda ingat* (forgot) and *baru ga* (again). Each word's pronunciation is correct, however, the lack of emotion made the sentence was less likeable and slightly difficult to understand and it needed to be listened to again for several times. (The lowest rating received were scale two, which means part of the sentence can be understood and another part is not.)

One clear distinction of the recorded speech over synthesised speech other than the phrasal break is pitch contour. This difference can be seen from Table 5.11. The correlation mean of the F0 is negative. However, the RMSE for duration and the first

Table 5.8: Comparison between synthesised and recorded speech syllables features for an example of fair rated speech synthesis

Syll	Dur	F0	Ene	F1	F2	F3
p@n	180	208	76	710	1901	3063
di	100	191	74	319	2461	3188
jaw	301	180	78	560	1588	2920
@NG	120	207	77	355	1211	2910
gaw	251	232	75	528	1371	2731
p@	110	249	76	526	1479	2904
ma	180	196	77	543	1376	2741
nah	190	203	73	630	1820	2836
j@	130	194	78	392	1400	2732
la	170	189	77	600	1688	2736
lu	190	185	79	413	1581	2825
l@NG	220	222	76	337	1505	2923
kas	291	244	74	983	2115	3273
di	160	213	73	531	2401	3279
t@	150	222	69	736	2114	3214
ri	130	204	75	378	2163	2891
mo	241	213	78	326	1014	2909
ba	150	177	79	581	1526	2727
low	180	170	78	621	1731	2720
ma	187	199	77	588	1424	2575
joh	214	217	77	436	1523	2944

Syll	Dur	F0	Ene	F1	F2	F3
p@n	160	183	70	428	1561	2761
di	70	183	74	373	2206	3004
jaw	261	210	79	682	1745	2799
@NG	60	219	71	441	921	2463
gaw	110	195	78	592	1163	2635
p@	110	191	74	676	1695	2950
ma	180	176	76	609	1426	2813
nah	200	189	73	571	1858	2831
j@	331	194	76	600	1835	2883
la	170	186	75	578	1569	2519
lu	170	170	77	487	1629	2782
l@NG	220	172	72	475	1537	2790
kas	321	221	71	846	2032	2799
di	150	181	71	540	1923	2679
t@	120	200	62	859	2088	3230
ri	120	180	76	431	2493	3139
mo	351	177	72	595	1155	3008
ba	140	164	74	541	1330	2570
low	200	164	75	533	1481	2851
ma	150	162	71	492	1518	2458
joh	491	170	72	866	1829	3262

Table 5.9: RMSE and Correlation for "*pendiau enggau pemanah iya lalu lengkas diterima bala maioh*".

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	87.2817	28.85761	20.82009	62.54446	3.83592	10.4129	3.854496
Corr	0.536293	0.405713	0.303489	0.051389	0.642099	0.789808	0.448449

	F1	F2	F3
RMSE	157.3637	234.8424	244.7875
Corr	0.505963	0.799532	0.394725

formant for the poor rated speech is better than the corresponding value rated fair in Table 5.9. A similar result can be seen for the corresponding correlation values. This showed the RMSE and correlation values may not be able to conclusively determine the perceptually good and poor synthesised speech, but it may be able to indicate the similarity between the features of the two languages. This will be discussed in Section 5.3.4.

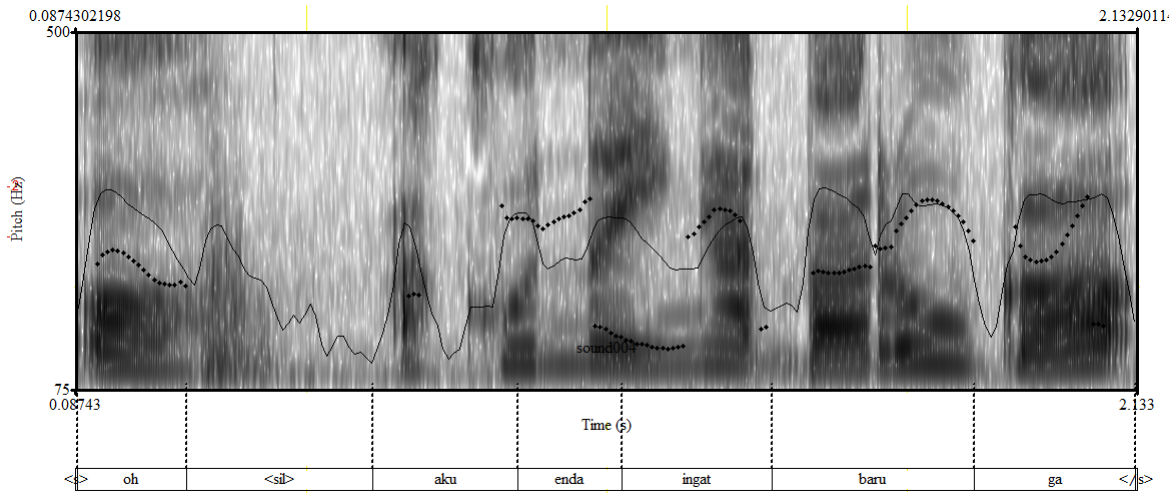

FIGURE 5.14: Native Iban speaker saying: *"oh, aku enda ingat baru ga!"*.

TABLE 5.10: Comparison between the synthesised and recorded speech syllables features for an example of poor rated speech synthesis

Syll	Dur	F0	Ene	F1	F2	F3
oh	211	272	77	792	1422	2992
a	111	197	73	859	1571	2630
ku	201	244	76	587	1351	3068
@n	161	190	77	440	1770	2866
daKK	161	196	74	647	1799	2712
i	90	236	79	355	2552	3227
NGat	271	207	74	518	1674	2923
ba	141	187	79	626	1549	2769
ru	181	191	78	422	1418	2856
gaKK	191	191	71	704	1712	2850

Syll	Dur	F0	Ene	F1	F2	F3
oh	211	219	75	574	1109	2739
a	70	185	69	1053	1717	2867
ku	211	247	68	733	1546	2859
@n	100	274	71	380	1519	2913
daKK	100	289	73	513	1601	2625
i	70	256	72	486	2674	3316
NGat	221	268	70	638	1768	2841
ba	161	215	75	616	1389	2438
ru	231	272	75	461	1166	2598
gaKK	311	249	75	686	1307	2591

TABLE 5.11: RMSE and Correlation for "*oh, aku enda ingat baru ga!*".

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	54.24297	57.87659	60.06996	63.35456	4.764452	13.28909	5.167204
Corr	0.716011	-0.19736	-0.10766	0.027308	0.206531	0.182089	0.048193

	F1	F2	F3
RMSE	127.0976	231.5953	207.7229
Corr	0.72523	0.874439	0.652061

5.3.4 Similar but not the Same

The previous section has shown segment comparison between the prosody of Malay and the prosody of Iban. While running this experiment, the dispute about the stressed of Malay language became clearer. It is obvious that Malay does not need stressing rules

in speech. Stressing is not so important as compared to the phrasal break. Stressing can be totally ignored if one wants to, but one needs to maintain the phrasal clause and phrasal break.

However, from the perceptual evaluation point of view, there was no noticeable phrasal clause. The synthesised sentences also sounded robotic when the sentences were long. Similar issues occurred for exclamatory and interrogative sentences. This showed that the prosody learned from the Malay training could be used to synthesise Iban speech.

The initial assumption when the Malay speech resources were used in the training was that the synthesised speech would sound foreign. Theoretically, this is true for non-related languages. But this experiment and perception test showed that some languages have similar prosody so that they can be used from one to another, given enough training provided by the focal language. This method however requires comparison between the desired and target language before the plugging in of the focal language can be done. As for this thesis, the comparison was based of the typology of the language family, the grapheme-to-phoneme similarity, the syllabification similarity and the applicability of the substituted grapheme.

There are other languages that sound like Malay, but using a totally different vocabulary set. Tagalog has very similar rhythm; however, it is a stressed language with a lot of glotalization. As with the Indonesian language, the default stressed position is at the penultimate syllable of a word. It could also occur in the final syllable. By understanding this similar root but different language aspect, brings to light the dispute linguists have about the focal language - Malay. Malay, Indonesian, Brunei Malay, Tagalog, Iban and others come from the same root, however, they all have stressed syllables except for Malay. Respondents agreed that the speech synthesis sounded natural although some words may sound drastically different due to tone mismatch and phrasal break misplacement. For exclamatory and interrogative sentences, the synthesiser was unable to produce the speech correctly even for the Malay language.

Table 5.12, Table 5.13 and Table 5.14 show the average high rating's (sentences rated as 4 and 5), RMSE and correlation. The fair rating is for the sentences with average scale 3 and the poor rating is for the sentences with an average scale rating of 2. (No scale rating of 1 (very poor) was given to any sentences.) By obtaining the high rated rating, it was hoped that a pattern of the speech features would be better observed.

Table 5.12: RMSE and correlation for sentences rated with the scale 4 and 5 using word boundary

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	125.51	32.05	34.15	63.69	3.06	14.43	3.55
Corr	0.7743	0.4935	0.5235	0.3020	0.4984	0.6775	0.3905

	F1	F2	F3
RMSE	101.91	166.87	181.13
Corr	0.6852	0.7544	0.4596

Table 5.13: RMSE and correlation for sentences rated with the scale 4 and 5 using syllable boundary

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	72.1301	35.3402	33.0837	58.7019	3.8286	12.4422	4.2179
Corr	0.740063	0.372814	0.511743	0.270942	0.280626	0.76185	0.147283

	F1	F2	F3
RMSE	126.7701	213.9239	231.5717
Corr	0.782227	0.783302	0.439705

Table 5.14: RMSE and correlation for sentences rated with the scale 4 and 5 using phoneme boundary

	Duration	F0 mean	F0 min	F0 max	Energy mean	Energy min	Energy max
RMSE	56.5181	37.1634	34.9415	51.0322	6.5322	11.0966	6.2451
Corr	0.4797	0.4350	0.4158	0.4420	0.6311	0.6549	0.5183

	F1	F2	F3
RMSE	187.4618	298.5892	299.7163
Corr	0.6738	0.6860	0.4213

As explained by Huang et al. (2001), there is no definite value of RMSE and correlation that defines what would be the best values. However, the best comparison that has been found (Dusterhoff and Black, 1997) showed that the value of RMSE and correlation are very similar. As cited, for boundary rating, Dusterhoff and Black (1997) obtained 0.778 for the duration, 0.530 for the F0 and 0.408 for the energy. Compared to the word, syllable and phoneme, the correlation of the duration is about the same for word boundary and slightly different for syllable but very different for phoneme. The F0 mean are all worse than the one recorded by Dusterhoff and Black (1997). The energy value between this experiment and Dusterhoff and Black (1997) is very close for word boundary but very different for the syllable and higher for phoneme boundary.

These values are believed to be affected by the type of language studied. Malay is not a stressed or tonal language as compared to Dusterhoff and Black (1997)'s English language study. Since English is a stressed language, there is less variation of duration and pitch which can be recorded (for the same word). The speech needs to be in a certain range and thus would result in a closer correlation even if the RMSE is not correlated.

A very consistent patterned language was represented in Dong et al. (2007) as the study was on Mandarin, and therefore, most of the segments (syllables in the study) had a similar contour and therefore high correlations and very low RMSE. The results of RMSE for prosody parameter prediction by Dong et al. (2007) are as follows: duration is 45ms, average F0 is 33.19Hz, and average energy is 697.5. A study by Dong et al. (2007) also showed a very close correlation between the predicted or calculated speech and human speech. The results of the correlation are as follows: 0.701 (duration), 0.829 (F0) and 0.681 (energy). This showed that the language has almost consistent pronunciation and has features which result in the parameters being locked in a certain range of values in order to produce the correct pronunciation.

When compared to the results obtained by Dong et al. (2007), the results obtained for the Iban-Malay study may not seem significant. However, it is suffice to note that the experiments conducted by Dong et al. (2007) were comparisons between a language learner's recording and the synthesised speech from a teacher's speech (called prosody model). Thus, in the experiments, the students were required to listen to the prosody model speech and then they had to reproduce the sound as closely as possible to the one they listened to. This again created a consistent voice range by the listener to correctly imitate the voice of the teacher's prosody model speech.

The main idea of having a different level of boundary analysis is, it is hoped that more similar features between Malay and Iban can be identified and the features made possibly clearer with a different level of boundary analysis. The high correlation for duration at the word and syllable boundary may indicate that the duration is similar between the two languages. The consistently low F0 mean may indicate that both languages may not have stringent rules for F0 contour. The fluctuated energy mean also might not mean much, however, if one cautiously observed the RMSE values for word, syllable and phoneme, they consistently produced very low energy differences. This may lead to the similar observation on the expressive speech in Malay whereby it is difficult to obtain

the expressiveness of the Malay speaker to collect sufficient samples for such studies except when using actors. A declarative sentence of the Malay language may also heard as monotonous speech when listened by non-native speaker. However no formal study can be cited at the time of writing this thesis.

5.4 Conclusion

This chapter focused on Malay-Iban prosody, mainly the duration and the pitch pattern in normal speech. An experiment was conducted to obtain Malay prosody characteristics at the phoneme level to be compared with the well-known duration model - Klatt. In this observation, it was realised that Malay prosody is easily adaptable and therefore the Klatt basic principle of duration assignment can be used. However, the same cannot be said about the fundamental frequency model. Based on the observation in comparison with Klatt's duration model, the applicability of Malay using an English duration model when the languages are not related at all may indicate that the language can have a rather loose duration model and therefore not tie to any particular language type. This may also be the contributing reason for the argument among linguists about Malay as a stressed language.

The follow up experiment was on constructing Iban generated speech by the use of a Malay speech synthesiser with the output being compared to a native Iban speaker's speech. The speech was synthesised using Malay speech trained data. The sentences without doubt were not expected to be perfect, considering that no tweaking or language adaptation was applied to the synthesised speech other than adding Iban words into the pronunciation dictionary. However, the quality of the synthesised speech was surprisingly as good as Malay speech except that the quality was worse when the sentences were longer and thus contained more lexemes. It is believed that the best possible improvement can be achieved by adding relatively short Iban speech recordings as implemented by Khaw and Tan (2014) when adapting Kelantanese dialect from Malay TTS. Although in the case of Khaw and Tan (2014), the dialect they proposed (Kelantanese) came from Malay. However, Iban and Kelantanese both have similar speech rhythms to Malay.

Chapter 6

Adapting Iban into Malay HMM-based Synthesis

6.1 Malay TTS using HMM-based Synthesis

Adapting a Hidden Markov Model (HMM) into speech synthesis is a popular option in the field of speech research nowadays due to its quality and its flexibility. The quality is said to be as good as a unit selection system but with less memory consumption, and more agile in terms of changing speaker identities, creating expressive speech, changing speaking styles and definitely more flexible for use in multilingual speech environments (Tokuda et al., 2002; Romsdorfer and Pfister, 2005; Latorre et al., 2006; Gonzalvo et al., 2007b; Schultz et al., 2007; Tokuda et al., 2013). However, as with the unit selection approach, HMM-based speech synthesis requires sufficient data to be able to achieve high synthesised speech quality. HMM-based synthesis on the other hand only requires 2MB for a standard HMM for the Speech Synthesis (HTS) toolkit without the use of any compression technique (Tokuda et al., 2013) which also makes it a desirable option for use in portable/mobile technology.

6.1.1 Pre-processing Data

Before going into the development of HMM for Malay TTS, a few pre-processing tools unrelated to HMM synthesis development need to be discussed. The first and foremost is the grapheme-to-phoneme converter.

TABLE 6.1: Grapheme-to-phoneme mapping for Malay in general

Grapheme	Phoneme	Grapheme	Phoneme	Grapheme	Phoneme
a	/a/	e	/e/ or /ə/	i	/i/
o	/ɔ/	u	/ʊ/		
ai	/ai/	au	/aʊ/	oi	/ɔɪ/
b	/b/	c	/tʃ/	d	/d/
f	/f/	g	/g/	h	/h/
j	/dʒ/	k	/k/	kh	/x/
l	/l/	m	/m/	n	/n/
ng	/ŋ/	ny	/ɲ/	p	/p/
q	/q/	r	/r/	s	/s/
sh / sy	/ʃ/	t	/t/	v	/v/
w	/w/	x	/ks/	y	/j/
z	/z/				

6.1.1.1 Malay Grapheme-to-Phoneme

The Malay grapheme-to-phoneme (G2P) rules can be viewed as straightforward. Malay is almost an orthographic phonography language in which there is almost a one-to-one correspondence from grapheme-to-phoneme. However, there are some consistent changes as well as irregular conversions.

6.1.1.1.1 One-to-one correspondent

Table 6.1 shows the one-to-one correspondence of Malay orthography and its phonemes. The longer graphemes should precede the shorter graphemes in these one-to-one rules. This mapping however can be overwritten by the rules stated in Section 6.1.1.1.2 and Section 6.1.1.1.3.

6.1.1.1.2 Persistent G2P rules

Schwa rule The rules apply to the grapheme 'a' at the end of the word which is always replaced with /ə/.
Example: *masa* is pronounced as /ma-sə/

Glide insertion Glide insertion for two different situations is necessary in Malay G2P. This is only necessary for two consecutive vowels: (u+a) and (i+a).

Example: *cuai* is pronounced as /tʃʊ-waɪ/
tiang is pronounced as /ti-jaŋ/

Pharyngeal insertion In Standard Malay, specific consecutive vowels occurrences will result in pharyngeal insertion. The consecutive vowels that can result in the insertions are: (a+a), (a+i), (a+u) and (o+a).
Example: *taat* is pronounced as /ta-ʕat/
dai is pronounced as /da-ʕi/
taun is pronounced as /ta-ʕun/
doa is pronounced as /do-ʕa/

Glottalisation Phoneme /k/ will have a different behaviour when it is located at the end of a syllable. At the coda position, /k/ will be pronounced as /ʔ/.
Example: *kakak* is pronounced as /ka-kaʔ
kekwa is pronounced as /keʔ-wa

Phoneme deletion When the phoneme /r/ is located at the end of the word, it will be a silent /r/.
Example: *sukar* is pronounced as /su-ka/
seluar/ is pronounced as /se-lu-wa/.

Nucleus replacement in the final syllable When the word final syllable (or even if the word consists of one syllable), which consists of onset, nucleus and coda, whereby the vowel is 'u' or 'i', a replacement will occur.
Example: *puluh* is pronounced as /pu-lɔ/
bilik is pronounced as /bi-leʔ/

6.1.1.1.3 Irregular G2P

Other than the stated rules, there are conditions where the pronunciation differs or it does not comply to the standard pronunciation method. Such irregular conversions require the word itself to be stored in the irregular pronunciation data. Such examples are:

Pronunciation of [e] **or** [ə] There is no distinction in the written form for both of the pronunciations above. As graphemes, they are both mapped from 'e'. The native speaker has knowledge based on the vocabulary built over the years. However, it is found that the phoneme /ə/ occurs five times more frequently than /e/ based on the study of the phoneme prosody in Section 5.1. Therefore, the default pronunciation is /ə/ while the list of word uses of /e/ will be stored in irregular pronunciation dictionary.

Alphabet with diphthong sequence but pronounced individually In the Malay sound system, the sequences (a+i) and (a+u) are almost always classified as diphthong. However, there are certain words which are not treated as a sequence of diphthongs.
For example:
bau (smell) is pronounced as /ba-wʊ/ unlike the word *pau* which is pronounced as /paw/ or /paʊ/.
dai (preacher) is pronounced as /da-ʔi/ unlike the word *pai* (pie) which is pronounced as /paɪ/.

Sense disambiguation There are a few pairs of words in Malay that have the same spelling but different pronunciations. The pronunciation of the words depends heavily on the sense of the sentence. For example:
perang can be pronounced as /pə-raŋ/ (war) or /pe-raŋ/ (both variations of the colour brown)
sepak can be pronounced as /se-paʔ/ (to kick) or /sə-paʔ/ (to slap).

This issue has been addressed in English-Malay machine translation research, however word sense disambiguation studies were conducted for English as the source language to be translated (Lim, 2006; Tat et al., 2001). At the moment this thesis is written, no study on handling Malay word disambiguations for TTS usage are known.

Influences from the original language For every loan word, the assimilation would already have occurred before the written form is formed. For example, the word *struktur* from structure, *puasa* from *upavasa* (Sanskrit) and *biola* from *viola* (Portuguese) have all undergone adaptation. However, there are words that have already been assimilated into Malay writing, but the pronunciation still follows the origins of the word such as *agenda* which was supposed to be pronounced as /a-gen-da/ but is instead pronounced

as /a-tʃen-da/, and *media* was supposed to be pronounced as /mə-di-jə/ or /me-di-jə/ but instead is pronounced as /mi-di-ja/.

Therefore, to facilitate the pronunciation of these irregular words, the words need to be stored in an irregular pronunciation dictionary.

6.1.1.2 Automatic Labelling Tool

The Malay speech synthesis uses a HMM-based synthesiser. The speech was first trained, the spectral features were extracted and the parameters represented by the HMM states (cf Figure 3.6 p. 40). When speech generation occurs, HMM predicts these speech parameters from the given text by concatenating spectral and excitation parameters. These parameters contain information on the voicing, fundamental frequency and spectral envelope represented by mel-cepstral coefficients. Therefore, for the speech waveforms to be reconstructed from the sequence of these acoustic parameters, a sufficiently detailed label was necessary to provide the information required so that the best HMM states can be selected.

The HMM required the label to be in this format before it can be synthesised:

```
sil^b-@+r=a#2_5/A:1_3/B:i/C:1_11/D:25
b^@-r+a=s#3_4/A:2_2/B:i/C:1_11/D:25
@^r-a+s=aj#4_3/A:2_2/B:i/C:1_11/D:25
r^a-s+aj=b#5_2/A:3_1/B:i/C:1_11/D:25
a^s-aj+b=@#6_1/A:3_1/B:e/C:1_11/D:25
s^aj-b+@=d#1_6/A:1_3/B:b/C:2_10/D:25
aj^b-@+d=i#2_5/A:1_3/B:i/C:2_10/D:25
```

Each alphanumeric character carries meaning except for A, B, C and D. The template of the label file (.lab) was revised in by Oura (2011). The template proposed 53 features to be recorded. In the Malay TTS, only 13 features were recorded. The following is the meaning of each character sequence. Non-alphanumeric characters represent delimiters.

1. the identity of the phoneme before the previous phoneme
2. the previous phoneme's identity
3. the current phoneme's identity

4. the next phoneme's identity
5. the phoneme after the next phoneme's identity
6. the syllable count for the current phoneme (forward)
7. the syllable count for the current phoneme (backward)
8. position of the current phoneme's identity in the current word (forward)
9. position of the current phoneme's identity in the current word (backward)
10. position of the current phoneme in the word (b:beginning, i:intermediate, e:end)
11. position of the word of the current phoneme in the sentence (forward)
12. position of the word of the current phoneme in the sentence (backward)
13. total number of syllables for the whole sentence

It is believed that this information is sufficient for a context dependent HMM to provide the distinctive features required to reconstruct the speech.

To create this label file, it would be possible for it to be produced by hand although it would be too prone to human error, not to mention too time consuming. Therefore, an automatic force alignment algorithm was applied using the automatic speech recogniser, Sphinx3, together with its associated tools, Sphinxbase and HMM toolkit (HTK).

6.1.2 Training and Data

HMM-based speech synthesis consists of two main parts: the training part and the generation part (as shown in Figure 3.6, page 40). The training part performs the extraction of the acoustic parameters. There are two kinds of parameters: excitation and spectral parameters. If one considers human speech production, the source of the sound is represented by the airflow and the voicing information of the vocal cords in the excitation model. Then the resonance that carries the information of the acoustic speech pressure wave, with the spectral envelop of glottal flow, vocal tract resonance and lip effect, forms the spectral parameters. The training goal is to assign these extracted feature vectors to HMM states.

6.1.2.1 Corpus Acquisition

For HMM-based speech synthesis, not only is recorded speech required but also prepared text data. The speech requires text annotation and for the purposes of training there is a need for a pronunciation dictionary for the system.

At the point of the construction of Iban speech from Malay TTS, there were 130 hours of recorded speech from 199 speakers collected for the Malay TTS. This included news reader from television broadcast news and direct recordings in the recording room. During the recording, it was tried to achieve a balanced number of speakers from different ethnic groups as well as a balanced number of male and female speakers.

The speech recording comes with its corresponding text. It is used in the training of the HMMs. There is also the construction of a pronunciation dictionary. The words were obtained from Kamus Fajar of Fajar Publication and then were added by using a web crawler. The word list also included all the affixed verbs as well as nouns. There is also the possibility for Malay text to include English words since not all English loan words have been assimilated into Malay. Thus, more than 5,000 English words entries were selected by running the text and identifying non-existent words from the constructed pronunciation dictionary and then these English words were added to the pronunciation dictionary. In the end, more than 76,000 entries were available in the pronunciation dictionary. The earlier construction of the corpus is presented in Tan et al. (2009).

6.1.2.2 Training

After acquiring the speech corpus, the phones in the utterances were aligned to create speaker dependent acoustic models. Since manual alignment of utterances is expensive and time consuming, automatic alignment was applied by force aligning the utterances using an automatic speech recogniser, Sphinx 3 from CMU. The aligned speech was then used to train an acoustic model for the HMM speech synthesis system. The training process followed closely the HMM-based speech synthesis (HTS) user manual with Malay data.

The HMM-based speech synthesis system had three significant differences from HMM-based speech recognition. Two kinds of parameters were extracted from the speech recording for the training for the TTS. One was the information about the spectral parameters - mel frequency cepstral coefficient (MFCC) - in the case of the Malay TTS, and the other one was excitation parameters. This was different from ASR because only MFCC is required. MFCC can represent the information of voiced sound but the unvoiced sound cannot be represented by the spectral parameter. Therefore in TTS, the excitation parameters are needed to represent the unvoiced region. HMM

uses multi-space probability distribution to model the F0 sequence which consists of a continuous distribution for voiced frames and a discrete distribution for unvoiced frames. The duration modelling uses a semi-Markov structure where the temporal structure is calculated by a Gaussian distribution.

The extracted information was stored in context dependent HMMs and state duration models. However, the context dependency also kept linguistic context information such as phoneme information, lexical stress, pitch accent, tone, part of speech and many others. It also kept the information at the syllable level, word level, phrase level and sentence level. While it was ideal to have all available characteristics recorded, the synthesiser could still provide sufficiently good quality speech by only giving optimal information. For Malay speech synthesis, only 13 linguistic features were recorded, the stored information of the current phoneme covers up to the pentaphone of the current phoneme.

6.1.3 Synthesising

Speech waveforms were generated based on information about the sentence that had been converted into label files as described in Section 6.1.1.2. The linguistic information, the duration of each state and the spectral parameter used the label file of the text input to construct the sentence. Based on the label sequence, the HMMs of the context dependence were concatenated to form the input sentence. The duration of each state was determined to maximise the probability of the state duration probability distribution. Then the spectral and excitation parameters were determined using the HMM speech generation algorithm of the HTS. Finally the synthesised speech was generated using a speech synthesis filter.

6.2 Adapting Iban into Malay HTS with very Minimum Data

The Malay TTS using HMM depended heavily on the associated tools as well as the data. The data collection was a very tedious process. For example, the speech recording was still an ongoing process at the time that this thesis was written. The massive data collection was necessary to obtain better quality speech as well as the flexibility of adapting different speakers as the front voice if desired. The pronunciation dictionary

was considered complete; however revisions were still carried out to ensure the accuracy of pronunciation.

Although the quality may not surpass the unit selection speech synthesis approach, the flexibility of reusing the TTS for another language is potentially valuable, especially when resources are limited. Before the adaptation of Iban into Malay TTS was conducted, a few preliminary studies were conducted by producing synthesised speech of Malay using English, Indonesian and Spanish data. Despite having a good phoneme coverage, the English producing Malay system still sounded foreign and spoke too fast. The Spanish, although having very similar basic grapheme-to-phoneme rules as well as syllabification rules, did not have schwa, /ə/ in Malay. That itself made the synthesised speech sound odd. Most Malay listeners did not have problems with Spanish stressing. As for Indonesian, the pronunciations were almost all correct but a few frequently used phonemes were missing, like /v/, /ʃ/, /q/, /ʕ/, /x/ and /f/. However, the listeners realised the intonation used was Indonesian. This may have been because of the consistent stressing of the penultimate syllables.

Then the Iban synthesised speech was constructed using Malay data. Other than that the quality of speech was not acceptable and the noticeably wrong pronunciation of glottal at the word end, no comments on the foreignness of sounds were received. This resulted in an attempt to create Iban synthesised speech using a Malay synthesiser and Malay data with a very small amount of Iban information included.

6.2.1 Iban-Malay Similarity and Dissimilarity

Iban and Malay are very similar if one looks at the surface of the phoneme set, the grapheme-to-phoneme conversion and even the words. One language could be considered a dialect of the other. However, it has been proven that it was not the case based on the cognate of the two languages. They however, could be very similar in other ways non-related to vocabulary aspects of the languages.

6.2.1.1 Iban Grapheme-to-Phoneme

Based on discussion with an expert, the languages have very similar grapheme-to-phoneme rules in case where the grapheme exists. The one-to-one mapping of Iban is shown in Figure 6.2.

TABLE 6.2: Grapheme-to-phoneme mapping for Iban in general

Grapheme	Phoneme	Grapheme	Phoneme	Grapheme	Phoneme
a	/a/	e	/e/ or /ə/	i	/i/
o	/ɔ/	u	/ʊ/	o	/ɤ/
ai	/ai/	au	/aʊ/	ui	/ʊi/
ia	/ia/	ea	/ea/	ua	/ʊa/
oa	/ɔa/	iu	/iʊ/	ie	/iə/
ue	/ʊə/	oe	/ɔə/		
b	/b/	c	/tʃ/	d	/d/
g	/g/	h	/h/	j	/dʒ/
k	/k/	l	/l/	m	/m/
n	/n/	ng	/ŋ/	ny	/ɲ/
p	/p/	r	/r/	s	/s/
t	/t/	w	/w/	y	/j/
z	/z/				

As with Malay, the phoneme 'e' can be pronounced as /e/ or /ə/, and in addition, the Iban 'o' can be pronounced as /ɔ/ or /ɤ/. Iban however, has more extensive use of diphthongs than Malay. Only the diphthongs /aɪ/ and /aʊ/ are common in both languages.

Iban pronunciation rules also have some similarity to Malay. However, for some rules, there are words that would be an exception to such grapheme-to-phoneme rules and therefore those exceptions should be listed in the irregular words list.

Glide insertion Sometimes insertions happen, and sometimes they are replaced with diphthongs. For example, the word *dua* is pronounced as /dʊ-wa/ but the word *kuap* is pronounced as /kʊap/.

Glottalisation Similar to glide insertion, different words will have it but other words with the similar sequence would not. For example, the word *menua* is pronounced as /me-noa/ and *iya* is pronounced as /i-ja/ without glottalisation, but the word *sida* is pronounced as /si-daʔ/.

Nucleus replacement The word *puluh* exists in both Malay and Iban. In Iban, it is pronounced as /pʊ-luəh/ while in Malay, it is pronounced as /pʊ-lɔ/. However *ngirup* is pronounced as /ŋi-rɔp/ while the closest word (in term of spelling) for that word

in Malay would be *hirup* is pronounced as /hi-rʊp/ which is quite close to the Iban grapheme-to-phoneme conversion.

Pronunciation of two orthographic vowels In Iban, the graphemes 'e' and 'o' each have two possible sounds. 'e' can be pronounced as /e/ or /ə/. 'o' can be pronounced as /ɔ/ or /ɤ/. The grapheme 'o' can also br pronounced as /ʊ/ as in *orang*.

Therefore an extensive pronunciation dictionary is required in order to provide accurate grapheme-to-phoneme results. It is believed that, similar to the variety of Malay rules, it may be best to keep a pronunciation dictionary for Iban and then, due to the variety of pronunciation differences, only use the grapheme-to-phoneme rules for words not existing in the pronunciation dictionary.

6.2.1.2 Iban Syllabification

Other than trying to match the grapheme-to-phoneme correspondence between the two languages, similarity studies of the syllabifications of the two language were carried out. Syllabification only works at the grapheme level and therefore the rules stated in the previous section was apply before syllabification.

For Malay syllabification, the list of irregular pronunciations, as well as words with selected persistent grapheme-to-phoneme changes are stored in a pronunciation dictionary. A similar method was performed for Iban where the irregular pronunciations were stored in the irregular grapheme-to-phoneme dictionary. However, obtaining the complete list for Iban irregular G2P was not possible due to insufficient recorded data. Therefore the pronunciation that was stored in the pronunciation dictionaries was the following:

- graphemes which do not follow one-to-one mapping of the Iban grapheme-to-phoneme rules. For example: *orang* which is pronounced as /ʊ-raŋ/ and *buk* which is pronounced as /boəʔ/.
- all words with diphthongs.
- glottalised lexicon. It is inconclusive whether glottalisation occurs in the language more than non-glottalisation. Therefore, glottalisations were listed in the irregular grapheme-to-phoneme dictionary.

Using the surface syllabification rules, all the Iban words can be synthesised as in Malay using Appendix C. This may be due to the similarity of the orthographic systems. Both languages have almost a direct one-to-one mapping from orthography to phoneme other than the given irregular graphemes-to-phoneme conversion. In fact, given all the irregular words available in the dictionary, Iban words were more consistent to syllabify. The Malay language, which has a more complicated morphological structure, would need additional rules to compensate for irregular syllabification when involving complex morphological transformation as described by Ranaivo-Malançon (2004)[1].

6.2.2 Adapting Iban into a Malay TTS

Other than unavailable speech resources and text resources, under-resourced languages also require standardisation in terms of pronunciation. As for Malay, no matter what the political situation and changes that occur, Standard Malay is always the foundation of the study in the Malay language. Other studies will evolve from it, for example the Standard Language of Malay and the different dialects of Malay. It would be ideal to have a standard reference so pronunciation would be consistent. Jaku' Iban which means conversational Iban is not a formal language but it is a formally taught language and therefore would already have a standard syllabus in teaching and learning in schools.

For creating Iban polyglot speech synthesis, grapheme-to-phoneme conversion was done by adding Iban words and lexemes into the pronunciation dictionary. Iban phrases were also included in phrase dictionary for pause insertion. Then the labelling was done using the Malay labelling tools as described in Section 6.1.1.2. As described, only the current phoneme and the pentaphone information were stored and therefore the Iban synthesised speech used Malay resources after labelling.

It was expected the quality would not be good, but due to the good quality of the Malay synthesiser, the quality was expected not to be very poor either. In the following subsections, the work on the evaluations will be described and the feedback of the respondents will be presented.

[1]It is necessary to note that the Iban text used in syllabification testing was not from the phonological balanced list, but from the Jaku' Iban word list provided by the SaLT team, Universiti Malaysia Sarawak.

6.2.2.1 The Experiment Design

Putting restrictions om Iban native speakers to be listeners may not be productive since generally, the language itself is used in a bilingual setting at home. People in Borneo have a variety of languages. There is a possibility for them being able to speak Iban even when they are for example, a Kadazan or Dusun people just because they are living inside Iban communities. By finding respondents that already exposed to Iban at an early age, it is hoped that more accurate results would be obtained.

Respondents were asked to evaluate two languages: Iban and Malay. Originally, the same amount of sentences to be evaluated was given in this final experiment. However, since the purpose was only to obtain the respondents level of acceptance towards speech synthesis quality in general, the number of Malay sentences for evaluation was reduced to three.

6.2.2.2 The Questionnaire

Thirty Iban sentences were given in the questionnaire. The respondents were required to answer four questions for each sound. First, they were asked to type back what they thought they heard. Then, on the scale of five, they needed to identify the level of effort required to understand the synthesised speech. On the third question, they were required to identify the level of naturalness/likeability on the scale of five. Finally, they were requested to give their insight on the overall quality of the sentence speech quality.

It is necessary to obtain all these answers from respondents. By having the respondents type back what they had heard, not only it would be possible to determine whether they really grasped the sentence or not, but how accurate their perceived speech was on the said words. The level of effort ranged from '*No effort require to understand*' to '*No meaning understood*' were divided into five scales rating. The rating was used to identify the level of effort of reproducing the sentences.

It would be beneficial to get the respondents acceptance towards synthesised speech. Therefore the third question asked was: '*What is your opinion/feeling when you listen to the synthesised speech*'. The answer was also a scale rating from 1 to 5. However the labels provided for the respondents were directed towards the closeness to the native speaker's speech.

The fourth question is about the overall quality rating of the synthesised speech. Although it might be redundant with the prior questions, this answer also showed the acceptance of the respondents towards speech synthesise in general. Considering that this may be exclusively for this Iban adapted synthesiser, the respondents' feedback on the Malay speech synthesiser were observed as well.

6.2.2.3 The Respondents

Fifteen respondents participated in the survey. The respondents age ranges are mainly below 35 with 3 respondents below 20, 2 respondents between the age of 21 to 25, 5 respondents between the age of 26-30, 3 respondents from the age of 31 to 35, 1 respondent between the age of 36 to 40 and 1 respondent age 41 or more. Most respondents were students in various education institutes in Malaysia. 5 were doing diplomas and degrees, 6 were doing post graduate studies (Master or PhD) and the remainder were either a tutor or a lecturer at the moment the experiment were conducted.

6.2.2.4 Listeners Configurations

To avoid unnecessary distraction during listening, respondents were asked to use a pair of headphones instead of computer speakers. Respondents were also requested not to rush into completing the experiments. Should the need to stop occur, respondents were asked to submit whatever questions they have completed and then continue from where they have stopped. When such situation occurred, results were only recorded once all responses have been received.

6.2.2.5 The Speech Data

The synthesised speech was generated using the HMM-based synthesiser for Malay. The list of sentences was extracted from a compilation of Jaku' Iban used by the SaLT team. This compilation was generated from the available text obtained from all resources, specifically from the Tun Jugah Foundation. The selected text and thus, the speech synthesis covers all phonemes used in the Iban language.

6.2.2.6 Rating Scale Lists

The summary of the respondents evaluation can be observed in Section 6.2.3. It is however necessary to list the scales used for the different ratings. The intelligibility rating can be summarised as follows:

- 5 – completely correct
- 4 – one or two (for long sentence) missing words or mistakes
- 3 – half of the sentence is correct
- 2 – at least meaningful consecutive words (or one correct word for short sentence)
- 1 – only one correct word/not answered/totally incorrect sentence

The effort rating scale was also classified into a five points scale. The following is the description for the representation of the scale:

- 5 – No effort required
- 4 – Attention necessary, little effort required
- 3 – Moderate effort required
- 2 – Considerable effort required
- 1 – No meaning understood

Respondents were also asked to evaluate a naturalness/likeability rating. Likeability rating is more subjective but limited within the five scale:

- 5 – Like a native speaker
- 4 – Practically normal
- 3 – An (or some) anomalies in intonation
- 2 – Sounds robotic
- 1 – Annoying to be heard

By giving this five points scale, respondents were asked about what were their perceptions towards the synthesised speech. This question is unrelated to the prior questions. The questions asked for the respondent's acceptance rather than their opinion on the naturalness level of the synthesised speech.

Finally, the respondents need to rate the overall quality of each synthesised speech. The rating scale is as follows:

- 5 – Very good
- 4 – Good
- 3 – Fair
- 2 – Poor
- 1 – Very poor

6.2.3 General Respondents Rating

The results based on general response is presented in the following graph:

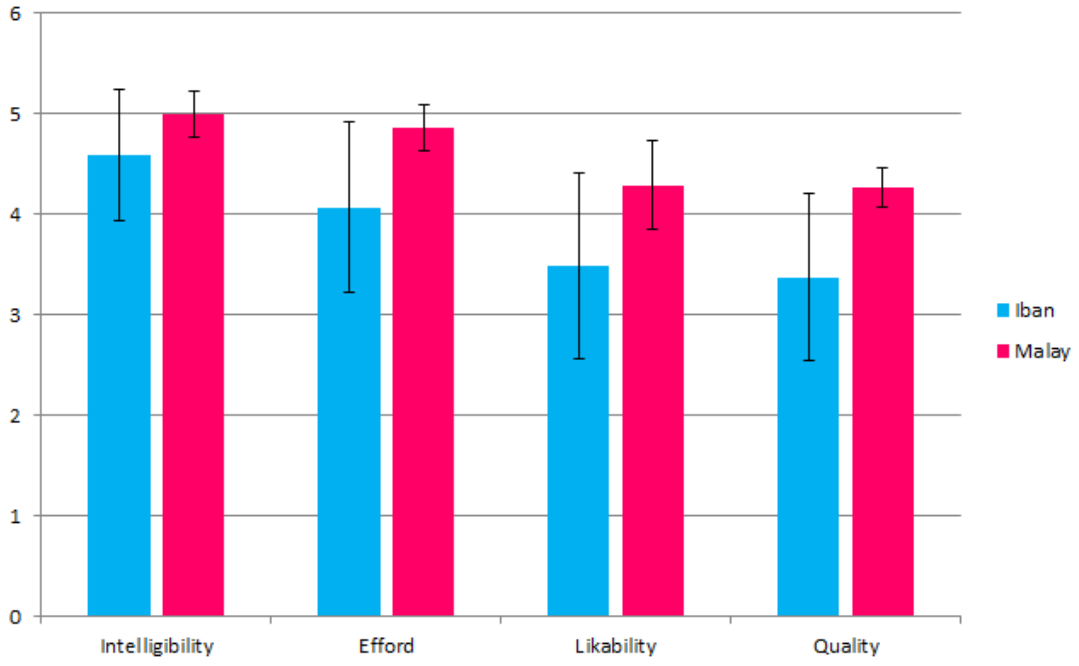

FIGURE 6.1: Respondents feedback on Iban polyglot synthesiser and Malay synthesiser.

The graph shows respondents rating for Iban polyglot synthesiser in term of intelligibility, effort required to understand the synthesised speech, likability and quality rating. Also shown is the rating for the Malay synthesiser by the same respondents to show their general rating towards the original synthesiser's language.

For Iban intelligibility mean rating, 4.5908 out of the scale of 5 was obtained with 0.6575 standard deviation of the sample and 0.1201 standard error of sample mean.

It is necessary to emphasise that the Malay rating may not be suitable to be used as the formal study on the respondents acceptance rating for Malay since only three synthesised sentences were given and thus, the more accurate rating that should be used must be obtained from a more formal study, for example in Khaw and Tan (2014). However, the

Malay speech samples ratings were included to be an indicator or marker in term of the respondents' acceptance of synthetic speech in general.

Respondents' rate for the effort required to understand Iban polyglot synthesised speech as 4.0700 in which standard deviation is 0.8531 and standard error is 0.1559.

Respondents' rating for the intelligibility is quite high but the mean for the effort required to understand is rated 0.5 scales lower than intelligibility rating. Since the background of the respondents varied with some of them never undergo such experiments and survey before, some may feel not confident with their given answers. Other than that, respondents found that when the synthesised speech is harder than the others, extreme effort was required resulting in the respondents to be able to grasp the correct words stated by the synthesiser but with tremendous effort. More observation will be discussed in Section 6.3 related to individual survey rating.

Respondents rate the likeability rating as 3.4873 with 0.9276 is standard deviation and 0.1694 is standard error while their opinion on the mean overall quality rating is 3.3745 with 0.8364 is standard deviation and 0.1527 is standard error.

Likeability rating would be influenced by the general acceptance factor – how close to human would one expect the synthesised speech to be, or how different the dialect used than the respondent's own standard of Iban nativity. And there are also factors contributing to the respondents' mood when listening to the individual sounds. If these characteristics are to be separated, the respondents would be expected to give a very deep evaluation rather than general perception of the synthesised speech. Therefore, the respondents were requested to choose the most suitable likeability rating they can provides from the five options. More observations on likeability will be discussed in Section 6.2.6.

The respondents' rating on quality is between good and fair. It is possibly best to compare the respondents rating to the rating of the focal language of the synthesiser: the Malay quality rating. The rating for Malay for the benchmark comparison is good. In general, the respondents themselves find the Malay synthesiser as 'good' rather than 'very good' despite Malay being the source language of the synthesiser.

6.2.4 Experts Rating

Five experts participated in the survey. By expert it is meant those who are already conducting or undergoing Natural Language Processing related experiments and they are/were a research students (MSc/PhD) and thus very familiar with research and survey sets of long questions. When their responses were extracted, the expert's mean rating only varied between 4 to 5. This might mean that the experts found that the synthesised speech as good enough in general. They might also be sympathetic and appreciative towards the effort of creating Iban synthesiser. Dissecting the experts response was not meant to show a higher rating (although this is actually the case) but in comparison with Malay TTS benchmark study, the respondents do indeed rate both of the synthesiser very good and good when compared to the general respondents rating.

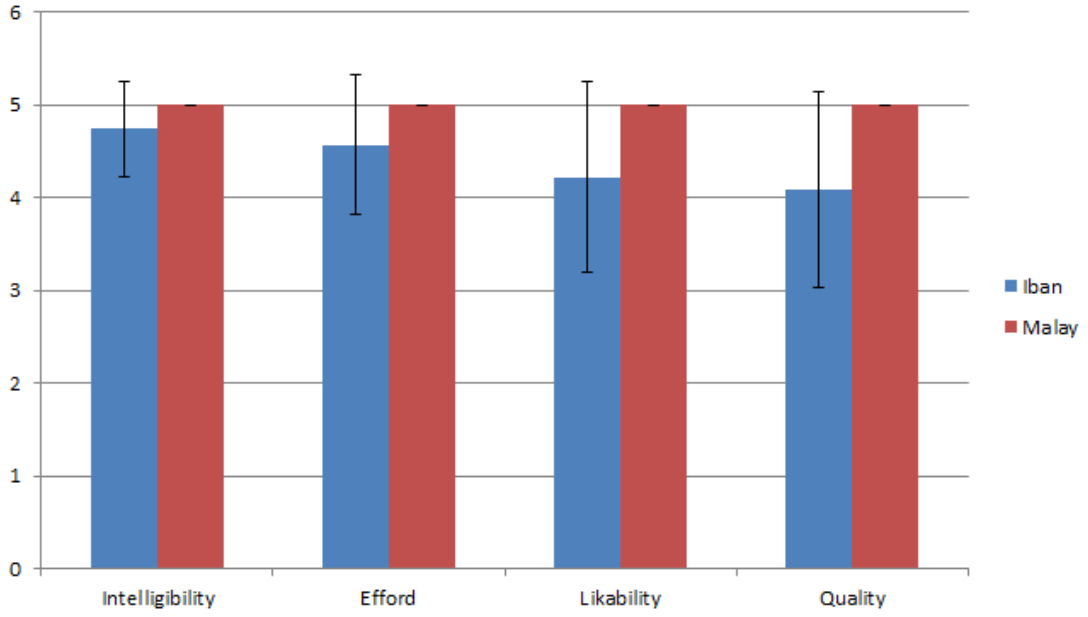

FIGURE 6.2: Experts feedback on Iban polyglot synthesiser and Malay synthesiser.

6.2.5 Individual Questions Rating

To further understand the rating provided by the respondents, respondents feedback based on individual sounds were studied. The list of sentences is given in Appendix D.

In Figure 6.3, the x-axis shows the labelled sound and the number of word counts in each sentence is given in parenthesis. The word counts may not be as important as

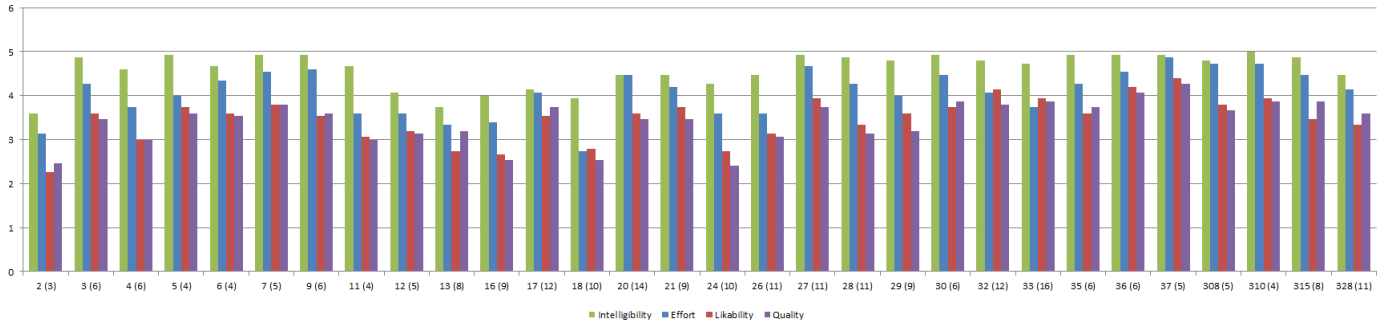

FIGURE 6.3: Overall Respondents Feedback based on Individual Sounds

the syllables in term of intelligibility but understanding the meaning will depend on the words and the therefore the word counts are provided.

Initially, it was thought that the longer sentence is, the more intelligibility or written text errors will be encountered. However, it is not necessarily the case. From the above graph, the sound 2 only consists of three words, but the intelligibility was 'just fine' as compared to the next sentence which consists of six words. In fact, the sentence with the longest and complicated (sound 33) were rated very highly by the respondents.

It was also thought that the better the intelligibility level is, the less effort would be required to understand. It was also assumed that the longer the sentence is, the more effort should be required to understand the text. These however were not the case.

For both intelligibility and the effort rating, it is found that sentences with the issues of: glide insertion, glottal insertion strength, mismatched diphthong, mismatched vowel and sentences with expression are affecting overall rating of the sentence. Each of these issues will be described in Subsection 6.2.6. The length of the sentence did not seem to influence the ratings directly.

Based on overall rating, respondents likeability are mostly rated between 'Normal' and 'An (or some) Anomalies in Intonation' with the rating slightly inclined towards anomalies existed in the perceived speech. This may not tallied with the respondents' high rating for intelligibility and effort required to understand. Likeability rating is not directly related with the understanding of the sentence, but rather the comparative evaluation to how native speakers speak the same sentences. The final option in the scale was added simply to estimate the respondents perceptions and acceptance – because it was found

that despite the sentence being very robotic sounding, the expert respondents were not distracted as much. Hence the option ‘ Annoying to be heard’.

Quality is a very subjective opinion. Some respondents may rate the quality poorly because some anomalies in the sentence. Other respondents may rate the quality well despite having a (or some) words misidentified. It is found that there is no consistent correspondence between quality with intelligibility, effort or even likeability. Expert respondents would tend to rate the quality as good or very good when compare against other respondents may be due to their knowledge that comparatively, the quality of the Iban polyglot synthesised speech were sufficiently good for them to be used in possible applications or tools. The individual questions rating by the respondents is given in Figure 6.4.

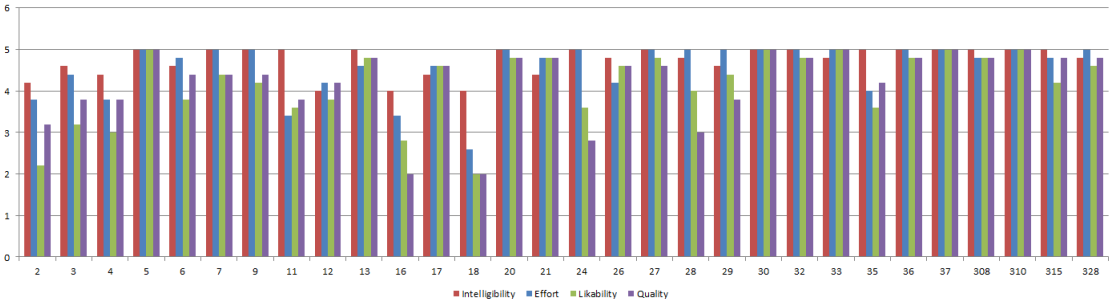

FIGURE 6.4: Overall Expert Feedback based on Individual Sounds

For more accurate tabulations on all the presented graph and data, Appendix E is attached.

6.2.6 Factors Influencing Respondents Rating

Based on the observation from respondents feedback, some limitations of the Iban polyglot synthesis has been identified which influenced the intelligibility and effort ratings.

6.2.6.1 Glides Insertion

Glide insertion rules also exist in Malay. However, it is found that, in the Iban polyglot synthesiser, sometimes the glide insertion may appear (perceived) too strongly. Expert respondents stated that those occurrences of glides insertion between ‘a’+‘e’ in bemain and ‘i’+‘o’ in maioh were too strong. From the responses, it can be concluded that there is a possibility that more than one level of strength when glides insertion is required exists in Iban.

TABLE 6.3: Iban substitution diphthongs based on perception of native speaker recording

Diphthong	Example	Supposed Pronunciation	Substitution	Pronunciation in TTS
/ea/	rumah	/rumeah/	/a/	/rumah/
/ia/	kiak	/kiaʔ/	/ja/	/kjaʔ/
/ie/	bilik	/bilieʔ/	/e/	/bileʔ/
/iu/	niup	/niup/	/i+u/	/niup/
/oa/	menua	/menoa/	/o+a/	/menoa/
/oə/	buk	/boəʔ/	/ɔ/	/bɔʔ/
/ua/	kuap	/kuap/	/wa/	/kwap/
/uə/	pun	/puən/	/wə/	/pwən/
/ui/	ukui	/ukui/	/ʊj/	/ʊkʊj/

6.2.6.2 Glottal Insertion

Glottal insertion rules also exist in Malay. In Iban however, a different degree of strength may be used in different occurrences. It is also suffice to state that, a couple respondents pointed out that there are a variation of pronunciation at the final glottal insertion depending on the region of where the speaker originated from.

Examples of glottal insertion are, the word *katak* (frog) in Malay is pronounced as /kataʔ/ and the word *iti* (a collective noun) in Iban is pronounced as /iteʔ/. However the word *badu* and *nginti* cannot be pronounced as /badʊʔ/ and /ŋinteʔ/ but should be pronounced as /badʊk/ and /ŋintek/. This implies that similar to the *Glides Insertion*, Iban may have two or more types of strength for glottal insertion.

6.2.6.3 Mismatched Diphthong

Iban is a language that uses a lot of diphthongs. However not all them available in Malay. The list of the unavailable diphthongs are shown in Table 6.3 in the left-most column.

Due to the high number of diphthongs in Iban as compared to Malay, non-existing phonetic similarity phonemes need to be replaced with another phoneme or sequence of phonemes. The table above shows the list of diphthongs that do not exist in Iban being replaced by the best available sound(s) in Malay. These replacements were based on the native speaker's opinion.

During these selection process, the list of original diphthongs was shown with a few words that associate with the respective diphthongs and the native speaker were given all the available vowel and diphthongs used in Malay. If native speakers found that there was no match between them, they would suggest the best replacement that they could think of.

The limitation of such a process is the native speakers will only think closely in respect to the given word list. Therefore it was later found in the survey that some diphthong replacements were not consistently a good fit for the targeted diphthongs.

6.2.6.4 Mismatched Vowel

Iban phonemes follow the old Austronesian language. In short, Malay has a wider consonant sound that can be used to represent the desired sound. However, in term of vowels, there is a variation for the alphabet 'o' which in Malay always pronounced as /ɔ/. In Iban, it can be /ɔ/ or /o/ depending on words as well as dialect of the sect. For this occurrence, there is no way to represent the right sound which do not exist in Malay and thus, if the sounds appear more than once in a sentence, respondents will notice the slightly 'distorted' sound.

6.2.6.5 Sentences with Expression

Malay sentence with expressions can be categorised into two: interrogative and exclamatory sentences. The perceptions of expressiveness for Malay however may be too subtle in a formal setting as what the Malay TTS is intended for when compared to the actual usage of expressiveness in normal conversation. It is also believed that due to the expressiveness in Malay, synthesised sentences are still lacking in conviction (when such sentences occur) and thus for the expressive part of the research needs an extensive study of itself before the issue can be reflected in an Iban polyglot TTS.

Thus, it is understandable that Iban polyglot synthesised speech also faces the same issue. The most fundamental issue with the Iban polyglot synthesiser is the nature of the language itself. This polyglot synthesiser is meant to cater for conversational Iban. Iban usage is directed more towards conversation.

Originally, it was thought that expressive Iban sentences will be rated poorly for intelligibility and effort. It is however not the case. But it does effect the quality and likeability rating.

6.3 Summary

This chapter described the method used to use a Malay speech synthesiser as an Iban speech synthesiser. The prerequisite Iban data was discussed, comparison and data matching between the two languages was also presented. Then a formal study of the Iban synthesiser for Iban speaker was conducted. The respondents' comments and behaviours' were also described. The comparison between general respondents and expert respondents were meant to show the different between expert opinion and general opinion. Issues of anomalies of the polyglot synthesised speech was addressed and suggestions on how these issues can be solved have also been brought forward. In the final chapter, the thesis will conclude as a whole with the hope that this study can be duplicated and revised to produce better speech synthesis or can be applied in other resource poor language synthesis development.

Chapter 7

Conclusion

Speech synthesis has advanced for more than half a century and reached a level of maturity for many well resourced languages with various synthesis techniques and is capable of producing high quality synthetic speech. Not only have speech scientists worked rigorously on developing systems, but the study of language itself has also reached a certain maturity and although one cannot say that the study of the language is complete, the studies that have already been conducted on language have been very thorough. However, speech synthesis for minority languages which lack of resources is still a challenge. The first or focal language in this thesis, Malay, is a widely used language with different dialects and in some instances categorised as a different language. It is used in Malaysia, Indonesia, Brunei, Singapore, Africa, the south of Thailand, the Cocos Islands as well as Sri Lanka.

The Malay language of Malaysia has been studied extensively by linguists since the 1950's. For Iban, however, progress started much later when compared to Malay.However, it has undergone rigorous development with the belief that it is possible to preserve the language and ensure that it survives to the younger generation by keeping a record of the text and historical related images. The study of Iban speech is less developed and is still undergoing work.

7.1 Synopsis of the Thesis

This thesis described research on building text-to-speech synthesis systems (TTS) for resource poor languages using available resources from other languages and using the

proposed general approach to building cross-linguistic polyglot TTS.

Sometimes, in machine translation, the translated text itself is not sufficient to convey the meaning. Therefore there can be a focus on the context domain or rather, the semantic roles within the text, to make the understanding of the text better than systematic syntactic translation. As in this research, the research has focussed less on what are the features of speech that can make a good synthesiser. In fact, unit selection-based and HMM-based speech synthesis have repeatedly shown very high quality performance already. The focus of this research is on the language features that can contribute to the reuse of cross language data in synthesised speech. The languages will need to be closely related in order to be able to be synthesised without data. However, what would be the criteria?

7.1.1 Phoneme Existence and Substitution

Chapter 4 showed that at the phoneme level, the duration of Malay and English were indeed close despite that they are not related at all in terms of typology, history or etymology except for loan words. This implied that the duration model proposed by Klatt, specifically the inherent and minimum duration threshold is possible to be reused for the Malay language. However, the duration model together with the rules may not conform properly with the Malay language as a previous study showed. If one creates a synthesiser using other resources, what would be the best resource to be reused? Thus, the study on phoneme substitution perception in was conducted.

There are two experiments conducted and presented in Chapter 4. The first is the study on phoneme confusion using standalone syllables. Second is the study on phoneme confusion using standalone words and also words in similar context.

For the first study, three sets of meaningless sounds were constructed. The first set is the sounds with CV structure where the phones of the consonant that is being studied are at the onset position. The second set is the sounds with VC structure where the phones of the consonant that are being studied is at the phones of the coda position. The third set is the sounds which have CVC structure where the phones being studied are at the onset and coda of the syllables respectively. Therefore, the third set have double the sounds from the CV or VC set.

The experiments have shown that substitutions are possible for selected phonemes in cases where the supposed phonemes do not exist in the target language. Based on the outcome of the first experiment groups (study on standalone syllables), it is found that when building the confusion matrices, respondents tend to be 'confused' more for coda phonemes than onset phonemes. Based on the second experiment, when no context is given, the substitute phoneme cannot be perceived as the intended word in consistence. However, when the context of the words is given, it is found that the substitute phonemes can be perceived as the intended word.

As a conclusion, substituted phonemes can indeed selectively replaced. Using the constructed confusion matrices, one can see the best substitute possible for each phone. Ideally, this substitute can be applied perfectly if a global phoneme is available in the resource collections.

Following that idea, the source language can be used as long as the target language source phonemes are a subset of the source language. Considering where there are missing target language phonemes, the source can provide replacements, then almost all resource rich, like English, German and French can be reused. However that was not the case based on prior studies of adaptation of Malay-English, Malay-Spanish or Malay-Afrikaans. Malay-Indonesian may be able to produce intelligible speech as has been done by other studies; however this was shown not to be the case for English and Spanish.

7.1.2 Prosody of Malay and Iban

Chapter 5 presented the comparison of duration used in the Malay recording with the Klatt duration model. Based on the comparison, the difference between the extracted value and the value listed by Klatt showed that it loosely fits between the inherent and minimum duration models. This indicates that the analysis showed that the duration of phonemes can fit into the Klatt duration model. It supports the plausibility of using other pre-recorded or phoneme information from another language.

Then a brief visual comparison of three sentences by a Malay native speaker set against the corresponding Iban were shown as spectrograms. From this the F0 contour, intensity contour and rough duration estimation could be depicted. Despite having different sets of sentence lengths, both languages showed close similarity in their duration, pitch and

energy contour. This does not indicate anything other than it is just another speech signal. Therefore the study on the closeness and pattern similarity was conducted to investigate how closely the Malay synthesiser can reproduce the synthesised speech of Iban text and the results compared to a recording of a native Iban speaker.

The purpose of running the experiment, was to see if there are any similarity values for well identified sentences when compared to a native speech recording. However, it is also known that the pitch values would not have a good RMSE values or even correlation values. Three levels of comparisons were conducted. The first were the comparison of RMSE and correlation at the phone level. Then, the study compared the RMSE and correlation of the syllables and words level.

Five experts respondents were given a set of Iban synthesised polyglot speech using Malay synthesiser. Based on their overall impression rating, the synthesised speech were grouped into good, fair and slightly poor categories and the RMSE and correlation means were observed.

The most interesting aspect of the outcome of the observation lay in the syllable boundary and word boundary analysis. From the outcome of the study, although Malay is a non-stressed language and Iban is a stressed language, the correlations of F0 mean and the energy mean were consistently low, but the energy minimum always has a better correlation value than the mean and the maximum corresponding energy comparison and so did the F0 minimum as compared to the F0 mean and F0 maximum. The duration's correlation is high for syllables and words levels are as compared to phoneme level. This consistent pattern indicates that these two languages indeed share undocumented suprasegmental features and create similar rhythms in speech.

It was supposed that the RMSE indicates the characteristic divergence between the two sound waves for the given features while correlation indicates the similarity across two sound waves. This RMSE and correlation study showed that there are similarity of prosody that can be matched from Iban and Malay. This can be used as a features before one language is adapted into another resource. Even when the phone coverage for vowels and diphthongs are considered poor, this prosody comparison provides a statistical agreement that the two languages share some prosody characteristics with each other despite the 'looseness' of intonation criteria.

7.1.3 Iban Polyglot Speech Synthesiser

Originally, the Malay synthesiser could not produce intelligible, correct pronunciation for Iban at all even though it sounded natural. When the pronunciation dictionary was added, the pronunciation started to become accurate and unexpectedly natural - in certain sentences. However, the quality was sometimes variable until Iban phrase information was added and therefore phrasal pause were included in the text. The quality then became very much better than was expected.

Iban and Malay are languages from the same root. Indeed, there is a controversial claim that Iban is a dialect of Malay. This however has been proven incorrect by experiments conducted on the cognate of the two languages. It is inevitable that both languages share similar properties. Iban consonants are a subset of Malay's consonants and the Malay vowels and diphthongs are a subset of Iban's. This similar phoneme set might also be true for English-Malay properties. But in the study of constructing a Malay TTS using English sounds, the synthesised speech sounded foreign. The compelling difference between Iban-Malay and Malay-English is the rhythm in speech. This may seem arbitrary. However, from a previous study by Brown (1988) on the pronunciation of English by Malaysian and Singaporean speakers, a staccato effect was found although Brown (1988) made an extensive study on the context of the 'staccato' effect showing that it is different from the use of the term in a rhythmical or musical sense. He also stated that their English was lacking the stressed-based rhythm of native accents and therefore lacked many of the features of the connected speech which are products of rhythm in native accents (Brown, 1988). This may indicate that the effect of the first language on the second language as presented by Charteris-Black (2002). This is also the possible reason why the Malay TTS using English resource and synthesiser engine had a foreign effect.

The Iban polyglot synthesiser were built using a Malay HMM synthesiser with very minor modification. At the initial stage of adaptation, the hope was only to get a temporary working synthesiser which later more information can be added to create one dedicated for the language. However, what was obtained from this adaptation is more than what was hoped to be accomplished.

The Iban polyglot synthesiser uses a pronunciation dictionary to handle the grapheme-to-phoneme conversion. This made the pronunciation dictionary contains three languages, which are Malay, selective English and Iban pronunciations. In the pronunciation dictionary, the phoneme substitutions (for none existed Iban phonemes) are assigned to matching Malay phoneme(s). After an informal evaluation, the phrasal dictionary were added and phrasal boundary is applied to the synthesiser. Then, using the trained acoustic model of Malay, the synthesiser produced Iban sentences. Only then the formal evaluation was carried out. Chapter 6 showed the outcome from this evaluation.

The feedback from the respondents showed that the sounds produced were indeed intelligible but some required repetitive listening and required moderate effort to be understood. It was also found that some of the respondents were not sympathetic towards synthesised speech for their language or for synthesised speech in general. For that reason, likeability questions were asked. The respondents were also asked to rate Malay synthesised speech for the same reason. What was expected to be derived was the respondents opinion of synthesiser in general as well as the Iban synthesiser. Since the synthesiser was built for Malay, it was expected that they rated the Malay quite highly.

It was found that for some of these respondents, they were not really uncomfortable with Malay TTS, but the low rating were given mostly to Iban synthesised speech only. When presenting the feedback in Chapter 6, the experts rating were given as a threshold to identify if some of the non-expert respondents were actually against the technology or they simply do not think the Iban synthesiser is not up to the par. It was found that latter was the case.

An interesting point about Malay is that it has no stress or tone or any other intermediate level of prosodic organisation. The lack of a stress system in Malay raises the question of how prominence in speech should be approached. Despite being told that Iban is a stressed language, the ability for Iban to adapt very easily into a Malay TTS framework calls for revisiting the prosodic characteristic of Iban prosody structure. It may be possible that as with Malay, standalone words seem to be stressed as presented in Don et al. (2008), and then later it was found that penultimate stressing was no longer there when the words are in a complete sentence. In fact, the focalisation shifts according to the context of the sentence. This is different from other stressed languages like English where the stresses and accentuation are created by the changes at the prosodic structure

of the sentence. Therefore, in line with Don et al. (2008), Malay (and therefore Iban) would require a different theoretical framework to represent the prosodic pattern of the language.

As for now, the similarity of prosody between the two languages was the reason for the compatibility of the Iban polyglot speech synthesis using the Malay synthesiser.

7.2 Reusing Resources of Speech Synthesis for Closely-Related Language

This thesis is based on the following questions:

How much acceptance of substituting phonemes can a listener bear while the speech remains comprehensible? This research relies on phoneme substitution for when the desired phoneme of the target language is not available in the source language. However, it was found from previous studies as well as this thesis that the missing of more crucial or higher frequencies phonemes would not be tolerated. It was also found that respondents are forgiving when the missing phoneme has a very close substituted sound to the sound that is supposed to be present. Also, it was found that when one language characteristic is missing, listeners find this very noticeable. In this research glottalisation of selected end words, one of Iban's characteristics was, when missing, highly detectable by the respondents. It was also found that when diphthong replacements were use repetitively in a sentence, the overall quality will be noticeably realised by the respondents.

Can the context of the text help in overcoming a missing phoneme of the target language by providing a closer sound? Definitely. This is proven by substitution based on context in Chapter 4 and then again in Iban polyglot speech synthesis. Without context, respondents were very sensitive to the criteria of the words and substitution did not work consistently.

If the language does not have a specific stress or accent pattern, how closely does it fit into Klatt's duration model which was originally tested on American and British English? Prior research has shown that the Klatt's duration model

is applicable to many languages. However, when tested on Malay, there were some differences on selected phonemes between the minimum duration of the Klatt synthesiser compared to the duration recorded in Malay. The difference was approximately 10 ms for the unfit duration. However, based on the experiment, the duration range of Klatt's duration model was close to the one obtained for Malay speech with the exception of a few phones. Therefore, it is safe to say that Klatt duration model can be used as the basis for other language, but necessary adjustment will be required to make it acceptable in the target language.

How is it possible to test one language's closeness to another based on speech produced? Would it be sufficient and conclusive? Prior studies on the closeness of one language to another were conducted by using a lexicostatistic approach, measuring the cognate using distance measurements. This is easier to accomplish if the writing systems are similar. However to find similar aspects in speech may require a different comparative evaluation. Therefore a comparison based on RMSE and correlation was conducted. If the language has definitive or objective rules of pronunciation, the tendency of matching spectral values would be very high. This has been shown for Mandarin which is a tonal language and therefore to ensure the correct word context is said, the pronunciation is more objective. This rule may be true for a stressed language like English although the pronunciation is not as stringent. Of these types of languages, the tonal particularly will show a very high correlation and low RMSE. However, for languages which cannot be classified as tonal or stressed, the pronunciation is more subjective and therefore, the variation will cause the correlation not to be as strong or the RMSE as small as compared to tonal and stressed languages. This was what was found in Section 5.3.4 where Iban native speech compared to Iban polyglot TTS was not as good as Dong et al., 2007 but similarly close to Dusterhoff and Black, 1997. This approach however would not be conclusive but a good guide before fitting a language into another synthesiser.

Is it possible to create a synthesiser using very limited target language resources or no target language resource at all? The answer is, yes to limited resources, although conditionally. The synthesiser used in this thesis used HMM-based

synthesis. In HMM TTS, the speech was reproduced using spectral and excitation features from the trained speech. The recording for the training determined the output speech quality. By providing sufficient speech data for training and associated files, the focal language synthesiser can be produced. The Malay synthesiser consisted of 130 hours of recorded speech used for training. It took years to prepare the data, obtaining the speakers and standardising the data. To date, more than 199 speakers have been involved in preparing the speech data for training.

This showed how time consuming and energy consuming the preparation process was for a developing language like Malay. However, for an under-resourced language, the initial process itself would take much longer. Therefore, reusing another synthesiser to create a new language synthesiser would speed up the process tremendously. Not all languages can fit to another. Previous studies showed that odd speech synthesis sound would occur when matching is simply done without prior consideration. However this study has shown that indeed the process can be shortened when a closely-related language is selected to be used in the focal language synthesiser. The selection of a closely-related language requires new language data. A pronunciation dictionary and phrasal phrases is sufficient to produce fair quality speech synthesis.

As far as the experiments conducted, creating a TTS without any speech data will not work. Respondents already have to embrace the concept of listening to synthesised speech and therefore introducing more dissimilar concepts will make the listening and understanding process harder. However, using very minimal language resources seems to work for Iban-Malay polyglot speech.

7.3 Lessons Learnt

Any natural language processing without target language data would be impossible to implement. However, by finding the subset or the overlap of the phonemes and manipulating them would make it possible to create the intended system although it will be lacking in some vital criteria. As described in Besacier et al. (2014), a scenario where some prior information of the target language is available, such as pronunciation dictionary, the language model and the language identification of the un-transcribed data, unsupervised acoustic modelling approaches are very useful to save time and costs. Although that is the case for an ASR, the similar principle can be said for a TTS.

Malay has a higher number of phonemes than Iban for vowels and consonants but Malay has fewer diphthongs than Iban. Most Iban vowels also exist in Malay but one vowel, /ɤ/, is not available. Therefore, the vowel was replaced with the existing vowel /ɔ/ even though that the phoneme is also available in Iban. Malay is severely lacking in Iban diphthongs. To facilitate this, the phonemes were treated based on the closest possible sound (or combination of sounds) available and some by retaining the same rules as in Malay. This is described in Section 6.2.6.3 (page 122). This has created a distinct difference from the intended sound.

From the evaluation of Iban speech, five factors that are possible to influence the bad rating were given in Section 6.2.6 in page 121. It is shown that one or possibly two occurrences would not affecting the rating much. However, when the occurrences were too frequent in one sentence, the respondents will automatically give bad rating especially in the likeability and quality ratings.

Since the overall rating is good, and all words with substituted phonemes were correctly identified in the sentences, it is concluded that phoneme substitution sufficiently worked for the Malay and Iban language pair. The same cannot be said for all languages; however, it is believed that the same can be said for closely-related indigenous languages. This should be explored as a future study.

7.4 Future Work

This research provides a beginning for reuse of existing TTS with minimal target language resources. The work concluded by showing that, with minimal data, a new language synthesiser can be produced. This is proven for this set of languages only. Future work could be done on other closely-related languages to test if this applicability can be extended to other languages, especially the indigenous closely-related languages to the focal language of this thesis. Further work can also further explore the possibility of adapting the language from a cousin branch of the stock rather than the siblings of the language stock. In order to get a complete synthesiser of the source language, a study of the minimum recording for training of the new language would also be helpful to determine how much recording would be necessary before the language can achieve a very good synthesised speech result.

The following will be a good continuant to this research.

Improve the Focal Language's Synthesiser. As concluded in Chapter 6, the Malay synthesiser was rated consistently good by the experts and most respondents. This is tallied with the respondent's feedback of the same system in the survey conducted by Khaw and Tan (2014). However, it was also found that the Malay synthesiser do not fully support interrogative and exclamative sentences. In fact, any extra expression if required would not have been produced accurately by the synthesiser as what would be expected by native speakers. Thus one way of improving the polyglot synthesiser would be to improve the focal synthesiser: Malay. Most mature synthesisers with massive data resources make use of the expressiveness in speech. It would be true for the Iban case and a possibility in future cases that if one wants to use Malay as a focal language of their polyglot synthesiser, having a better and more robust Malay synthesiser could create a better polyglot synthesiser.

Identify 'Feeble' Phonemes and adapt into Focal Language Data. In Chapter 6, an extended description on the phonemes occurrences that are likely to result in a poor rating were summarised. Despite phoneme substitution experiments has showed that the respondents tend to be 'forgiving' when the coda of the syllable sounding slightly off (being replaced by another). However, the intelligibility, effort, likeability and quality ratings also showed that when such phonemes occurred too many times in a sentence, respondents will rate it poorly. The best way to overcome this issue is by providing an ad hoc training or short recording of words or phrases in which these 'feeble' or 'flimsy' phoneme exist. This may not be possible to be done in the standard HMM-based synthesis since it is only ideal for HMM training to have recordings that cover all phones based on the frequency of usage in a particular language. Section 6.2.6 has identified five main factors that resulting poor ratings. These factors when occurring repeatedly or one after another in a sentence required a lot of effort to be interpreted by the listeners. Therefore, to reduce the effect of these feeble occurrences, some sample data should be added to mask or reduce the significant substituted phoneme effect of the system. For example, based on Section 6.2.6, overcoming mismatched of diphthongs and vowels is believed to aid in improving the respondents' evaluation ratings. Future work on selective training is required to improve the overall quality of the synthesised speech.

Adapting a different Training Algorithm. When using HMM-based speech synthesis, scientists have to deal with some of HMM limitations as described by Zen (2015). For example, the limitation described is an inconsistency of dynamic feature constraints which are not used in training stage but used in the synthesis stage (Zen, 2015). This will make the above improvement of *"Identify 'Feeble' Phonemes and adapt into Focal Language Data"* would be difficult to be carried out especially when the intended recording is extremely short. Therefore, it might be plausible to improve the polyglot synthesiser further while still using another language resource by using a different algorithm for the acoustic model training. In Zen (2015), they predicted that Long Short-term Memory Recurrent Neural Network (LSTM-RNN) would be the next dominant acoustic model in the future. It is said to have better consistency while still maintaining its' efficient training plus has lower latency than HMM approach despite being computationally more expensive than HMM, but better than any other statistical parametric approach.

Extend this work for other Minority Languages. The approach used in this thesis when adapting the synthesiser from Malay may seems like an oversimplified process. In a nutshell, the pronunciation dictionary is used, grapheme-to-phoneme converter were turned off, the pentaphone that were implemented in Malay were fully utilised, phrasal dictionary list were used for pausal insertion – specifically to provide a more controlled speech and more succinct/accurate pronunciation. Any instances which are not available in the pronunciation dictionary would not be able to be produced by the synthesiser. Theoretically, based on the substitution phoneme and RMSE and correlation experiments, this approach won't work for the Thai language. Thai is a tonal and stressed language with five types of tones and consists of very elaborate but specific pronunciations stressed rules. Since their rules are very specific, the pronunciation is more stringent and objective. The language itself has a wide consonant coverage and very rich with linguistic features. If one to apply the restricted process of adaptation as what has been done from Malay to Iban for Thai, it will not work. Due to Thai language characteristics, adapting Thai from Mandarin would be a better option since their acoustic model will need to represent their rich linguistics criteria. However, researchers are well aware that despite being a minority language of the world, Thai language does not face any lack of speech resources. The same cannot however be said for many minority languages in Malaysia which mostly are also under resources languages. It would be

beneficial to extend the research further by matching and adapting other minority languages from Malaysia as what has been done for Iban. Languages like Javanese, Banjar, Minang and Bugis for example would need to undergo validation whether it is a dialect or a different language. However, before any research can be conducted, a small team of willing speakers need to be obtained first. No matter how interesting the research of closely-related language is, it can never do without at least a willing and knowledgeable speaker who can describe his/her language accurately. Even the simplest question of whether it is a tonal language or a stressed language or none at all will be difficult to be answered especially when there was no formal written form of the language Researchers need to be forewarned that these languages, have higher usage of glottal and pharyngeal sounds and have richer aspirated and nasalised sounds which will make the grapheme-to-phoneme adaptation would not be as easily adaptable as one would think.

Appendix A

Phoneme Confusion

For the ease of readability, the outcome from four main studies of phoneme confusions are listed here.

A.1 Phoneme Confusions Matrices by Fant et al., 1966

TABLE A.1: Phonemes confusion for English Listeners by Fant et al., 1966

		Heard																						
		k	p	t	g	b	d	tʃ	dʒ	ʃ	f	θ	s	h	ʒ	v	ð	z	j	r	w	l	m	n
Spoken	k	10																						
	p		6											4										
	t			9							1													
	g				10																			
	b					10																		
	d						10																	
	tʃ							10																
	dʒ								10															
	ʃ									10														
	f										10													
	θ										8	1	1											
	s												10											
	h		2											8										
	ʒ														7			3						
	v															10								
	ð														1	3	5							
	z																	10						
	j																		10					
	r																			10				
	w																				10			
	l																1					9		
	m																						10	
	n																							10

A.2 Phoneme Confusions Matrices by Cutler et al., 2004

In Cutler et al., 2004 the confusion matrices are build to compare the confusion between native and non-native listener.

TABLE A.2: Phonemes confusion for Swedish Listeners by Fant et al., 1966

		Heard																
		k	p	t	g	b	d	ç	ʃ	f	s	h	j	v	r	l	m	n
Spoken	k	10																
	p		4							1		5						
	t	1	1	7								1						
	g				10													
	b					9				1								
	d					1	8			1								
	ç							10										
	ʃ							6	4									
	f					1				9								
	s										10							
	h											10						
	j												10					
	v													10				
	r														10			
	l															10		
	m																10	
	n																	10

A.3 Phoneme Confusions Matrices by Meyer et al., 2007

The matrix element denotes how often the phoneme in row was classified as the phoneme in column. Rows are normalized to 100%. Matrix elements with a value of zero are not plotted and elements <5 are plotted in light gray for reasons of readability. Inverted elements denote large differences between this confusions matrix (CM) in Table Table A.11 and Table A.12.

For the following vowel confusion matrix (Table **??**), gray-shaded elements highlight degradations that emerge when resynthesized signals instead of the original ones are used.

A.4 Phoneme Confusions Matrices by Lovitt et al. (2007)

In the issues of phoneme confusion in speech perception, most research presented the confusion of phoneme either at the initial stage of speech recognition process, during the analysis of the speech signals or at the phoneme recogniser of the speech synthesiser itself. Lovitt et al. (2007) identified three stages where confusions phoneme can happened. The chart shows the list at all stages. It is colour coded. The italic blue phonemes are phones

TABLE A.3: Confusion matrix for initial consonants at 0 dB SNR category for English listeners (Cutler et al., 2004).

	Response																						
Stimulus	pie p	tie t	car k	far f	thin θ	see s	she ʃ	chin ʧ	hi h	be b	do d	go g	very v	there ð	zoo z	joke ʤ	yell j	my m	no n	lie l	row r	win w
p	**15.4**	2.9	4.6	10.4	3.3				39.2	6.7	1.3	2.1	2.5	1.3		0.8	1.3	1.3	0.4	0.4	0.8	0.8
t	10.4	**19.6**	9.6	5.4	6.3	1.3	0.8	0.4	27.9	2.9	1.3	3.3	0.4	3.3	1.3	0.4	0.8	0.4	0.8	0.4		0.8
k	11.7	14.6	**25.8**	1.7	2.1		0.4	0.8	27.9	2.1	0.8	2.1	0.4	0.8			0.4	0.4	2.1			1.7
f	22.9	2.1	3.8	**19.2**	7.5	0.4			14.2	8.8	1.3	0.8	3.8	5.8		0.4	1.7					3.8
θ	12.5	5.4	3.8	13.3	**18.3**			0.4	10.4	7.5	2.1	1.3	3.8	14.6	0.8		0.4	0.8	0.4	0.4	0.4	1.7
s	0.4	2.1	0.4	9.2	10.0	**51.7**	2.1	0.4	2.1	2.5				9.6	8.8							
ʃ	0.4		0.4			0.8	**76.7**	19.6				0.4	0.4	0.4		0.4						
ʧ		5.0	0.8	1.7	0.8		1.3	**83.8**	0.4			0.4		1.7		2.9	0.4		0.4			
h	14.6	5.0	4.6	9.6	4.6	0.4		0.4	**36.7**	7.1	0.4	1.7	2.9	2.1		0.4		1.7	0.4	0.4	0.4	1.7
b	2.1		1.3	5.8	5.8				15.0	**19.6**	1.3	1.3	5.0	8.3	0.4	0.8	3.8	10.8	0.4	4.2	1.7	3.8
d		2.9		1.3	7.9	0.4		0.8	4.6	7.9	**14.6**	2.9	0.8	14.6	0.4	0.8	7.1	3.3	19.6	6.7		0.4
g	1.7	0.8	1.3	2.9	2.5	0.4			10.4	3.8	2.5	**29.6**	4.2	2.1	0.4	0.8	19.2	1.3	7.9	2.9	1.3	1.3
v	2.9	1.7	0.4	5.8	4.2	0.8		0.4	8.8	18.3	0.4	3.3	**17.5**	14.6		0.8	2.1	4.2	1.7	0.8	1.7	5.0
ð		1.3		1.3	14.6	0.8		0.4	1.7	9.6	4.2	3.3	5.8	**30.4**	4.2	2.1	1.3	1.3	4.6	10.0		0.8
z		1.7			9.2	2.5			0.8	1.7	1.7	2.5	8.3	21.3	**31.3**	2.1	0.4	0.8	3.3	1.3	1.7	7.5
ʤ	0.4	0.4	0.4	0.4	2.5	0.4		4.6	1.3	1.7	4.2	2.9	0.4	8.3		**68.8**	0.8	0.4	0.4	1.3		
j		0.8			0.4				2.9	3.3	5.4	3.3	1.3	1.3	1.7	2.9	**65.8**	2.5	2.5	1.7		2.1
m	0.4		0.4	1.7	0.4				3.3	3.8	0.8	1.3	6.3	0.4	0.4		0.8	**63.8**	5.8	5.8	1.3	1.7
n					0.8				0.4	0.4	0.4		0.4	0.4	0.4		0.4	12.5	**77.9**	4.2	0.8	0.4
l	0.4				2.5			0.4	0.8	5.0	1.7	2.1	3.3	4.2	0.4		2.1	12.5	5.0	**54.2**	2.1	0.8
r	0.8	0.4	1.3	1.3					7.1	5.4	0.4	2.1	5.0	0.8				2.5	0.4		**68.8**	2.9
w	0.8	0.4							1.7	4.2		0.8	2.9	0.4			4.2	5.8		2.9	0.4	**73.3**

TABLE A.4: Confusion matrix for initial consonants at 0 dB SNR category for Dutch Listeners (Cutler et al., 2004).

	Response																						
Stimulus	pie p	tie t	car k	far f	thin θ	see s	she ʃ	chin ʧ	hi h	be b	do d	go g	very v	there ð	zoo z	joke ʤ	yell j	my m	no n	lie l	row r	win w
p	**30.8**	3.3	9.2	9.6	2.9			0.4	19.2	11.7	1.3	1.3	2.9	1.7	0.4		0.8	0.8	1.3	1.3		1.3
t	24.6	**14.2**	12.5	7.5	7.9	0.8		2.9	11.3	7.1	0.4	2.1	1.3	1.7				2.1	3.3	0.4		
k	25.0	7.9	**25.8**	3.8	4.2	0.4	0.8	0.4	13.8	4.2	1.3	3.8	1.3	0.8	0.4	0.4	1.3	0.4	1.7	1.3	0.4	0.8
f	24.6	2.1	9.2	**15.0**	7.1		0.4	0.4	9.2	15.0	1.7	2.9	5.4	4.2		0.4	0.4	0.4	0.4		0.4	0.8
θ	18.8	6.3	3.8	13.3	**12.1**	0.4	0.4	0.4	7.1	14.2	2.5	1.7	2.9	7.5			0.4	1.3	2.9	2.9		0.8
s	0.4	2.5	0.4	12.5	24.6	**30.4**	0.8	0.4		0.8	1.3		3.3	7.9	14.6							
ʃ		0.4			1.3	6.7	**72.5**	18.3									0.8					
ʧ	3.3	4.2	1.3	2.1	2.5	1.3	4.6	**70.8**	1.3	1.3	0.4		0.4	0.8		5.4	0.4					
h	26.3	4.6	12.1	11.3	5.0	0.4	0.4	0.8	**17.9**	8.3	1.3	0.4	4.6	1.7	0.4		0.8		0.8	1.7	0.8	0.4
b	7.5	0.4	5.8	9.2	1.7	0.4		0.4	12.5	**28.3**	2.5	0.4	4.6	2.1			2.9	7.1	2.9	5.0	1.3	5.0
d	2.5	2.1	1.3	1.3	5.4				8.8	12.1	**10.8**	2.5	2.1	12.9	0.4	0.4	6.3	4.2	12.5	12.5		1.7
g	3.3	1.3	9.2	2.9	2.5		0.4	1.3	9.2	10.0	5.0	**17.1**	1.7	3.3		0.8	24.2	0.8	2.5	2.5	0.4	1.7
v	7.5	2.9	2.5	8.8	5.0	0.4			6.7	30.0	1.3	1.7	**9.6**	7.9			2.1	3.8	1.7	0.4	2.5	5.4
ð	2.5	1.3	2.5	1.7	14.6	2.1	0.4	0.8	2.1	17.1	10.0		1.3	**18.8**	1.3	1.7	1.3	1.3	3.3	12.1	0.8	2.9
z		0.8	1.3	1.3	9.6	3.3	0.4	1.3		7.9	5.0		2.5	23.8	**27.1**	1.3	2.5	2.1	5.0	0.4	0.8	3.8
ʤ	4.2	0.4	2.5	0.4	2.1		0.8	18.3	2.1	2.1	6.7	2.1		7.5		**40.4**	5.8	0.4	1.3	2.9		
j	1.3		0.8	0.8	0.8		0.8	0.4	2.1	4.2	2.5	1.3	0.4	1.3	0.4	4.6	**69.6**	2.5	4.2	1.3		0.8
m	3.8	0.8	2.5	2.9	0.4				2.1	9.6	0.8		2.5					**50.0**	10.0	5.0	5.4	4.2
n					0.4				2.1	1.7	1.3					0.8	0.4	12.9	**73.8**	4.6	0.4	1.7
l	5.8	1.3	1.7	1.3	1.3			0.4	2.1	8.3	1.7	0.8	2.5	3.3			1.3	10.0	4.2	**46.7**	2.1	5.4
r	2.5	1.3	1.7	2.1				0.4	5.8	14.6	0.8	2.1	1.3	0.8				0.8	0.4	0.4	**58.3**	6.7
w	1.7		0.4	0.4	0.4	0.4			1.7	5.8	0.8		2.1	0.4				5.8	0.8	2.5	1.7	**75.0**

TABLE A.5: Confusion matrix for final consonants at 0 dB SNR category for English listeners (Cutler et al., 2004).

	Response																				
Stimulus	lip p	hot t	sick k	off f	path θ	pass s	fish ʃ	such tʃ	grab b	odd d	egg g	love v	smooth ð	buzz z	beige ʒ	edge dʒ	am m	on n	ring ŋ	ill l	far r
p	**50.0**	16.3	14.2	5.8	5.8				0.8	0.4	0.4		2.1				0.4				0.8
t	5.0	**77.1**	5.8	0.8	4.6			2.1		0.4		0.4	2.5	0.4							
k	11.3	12.5	**63.3**	0.8	5.0			2.1	0.4	0.4	0.4	0.8	2.1								
f	10.0	10.0	6.7	**45.0**	12.9		0.8		0.8			1.3	5.4		0.8		0.4			0.4	1.7
θ	9.2	17.9	4.2	30.8	**19.2**	0.8	0.4	0.8	0.4	0.4	0.4	2.5	7.5		0.4			0.4	0.8		0.4
s	0.8	2.9	0.8	12.9	8.8	**65.4**	2.9	0.4			0.4		1.7	0.8	0.4						
ʃ					0.4	1.3	**80.8**	14.2					0.4		2.5	0.4					
tʃ	0.4	3.8	0.4					**89.6**							1.7	4.2					
b	1.3	1.3	4.2	3.8	2.5	0.4		0.4	**35.0**	10.4	9.2	15.4	4.6	0.4	2.5	1.3	2.1	1.7		0.4	1.3
d		3.3	0.4	3.8	2.5	1.7			3.8	**42.9**	4.6	6.7	5.8	1.7	5.4	5.8	0.4	5.8	2.9		0.4
g	0.4	3.3	1.3	2.9	5.4	0.4		0.8	5.4	9.2	**35.4**	14.2	5.4	0.8	2.1	1.7	1.3	2.1	2.5	0.4	0.8
v	0.4	0.8	1.3	9.2	2.5			0.4	2.9	4.6	7.9	**47.5**	5.8	0.4	3.8	1.7	2.9	1.3	0.8	0.8	1.7
ð		2.1		1.7	4.2	2.5	0.4	0.4	2.9	22.5	5.0	17.5	**16.7**	5.8	5.4	7.5		1.3	0.8	0.4	0.4
z	0.4				2.5	7.5	0.4	0.4	1.3	10.0	1.3	12.5	9.6	**37.1**	5.4	4.6	0.4	2.5		0.4	0.8
ʒ				0.4		0.8	2.5	2.1		2.5	1.3	4.2	4.6	3.8	**51.7**	23.3	0.4	1.7		0.4	0.4
dʒ	0.4	0.8			0.4		0.4	3.3	0.4	5.8	2.5	0.8	0.4	0.4	17.9	**64.6**	0.4				0.8
m		0.4		1.3	0.4				1.7	1.3	2.1	7.1	0.4			0.4	**56.3**	12.1	14.2	0.4	0.8
n					0.4				0.4	5.4	1.3	3.3	1.7	1.3	1.3	0.8	12.5	**59.6**	10.4		0.4
ŋ		0.4	0.4	0.8				0.4	0.4	1.3	9.2	5.8	1.3		0.4		15.4	25.4	**35.0**	0.8	1.7
l	0.4	0.8	0.8	7.9	0.4					0.8	0.8	6.7	3.3		0.4	0.4	1.3	0.4		**70.8**	0.8
r		0.4	0.4	1.3	0.8				0.4	0.8	0.4	3.8	1.3			1.3	1.3	0.4		0.8	**84.2**

TABLE A.6: Confusion matrix for final consonants at 0 dB SNR category for Dutch listeners (Cutler et al., 2004).

	Response																				
Stimulus	lip p	hot t	sick k	off f	path θ	pass s	fish ʃ	such tʃ	grab b	odd d	egg g	love v	smooth ð	buzz z	beige ʒ	Edge dʒ	am m	on n	ring ŋ	ill l	far r
p	**24.2**	13.8	11.7	5.8	8.8			0.4	21.3	5.8	2.1		3.3		0.8		0.4	0.4		0.4	0.8
t	4.6	**45.0**	5.4	1.7	9.2	1.7		1.3	1.7	20.4	0.4	0.4	5.4	0.4	0.4	0.8				0.4	0.8
k	8.3	12.5	**44.6**	4.2	5.0	0.4	0.4	0.8	2.5	3.8	12.5	0.4	1.7			1.3		0.4			1.3
f	7.5	21.7	6.7	**22.1**	9.6	1.7	1.3	0.4	6.3	10.4	0.4	3.3	5.0	0.4	0.8	0.4	0.8			0.8	0.4
θ	7.9	24.6	3.8	17.9	**17.5**	0.8	0.4	1.7	2.9	7.9	0.4	2.5	9.2		0.8	1.3					0.4
s		3.8		17.1	14.6	**37.5**	5.0	0.8	0.4	0.4		2.5	10.0	4.6	1.7	0.8					0.8
ʃ					0.4	6.7	**66.7**	10.4			0.4		1.3	1.7	10.0	2.1					0.4
tʃ	0.4	0.8		0.4	1.3	0.4	6.7	**42.5**	0.8	3.3	0.4		3.3		5.4	34.2					
b	5.0	7.5	6.7	2.9	5.4			0.4	**30.4**	15.0	3.8	7.9	5.0	1.3	0.4	2.9	2.5	0.8		0.8	1.3
d	1.3	16.3	0.4	2.5	5.8	0.8	0.4		2.1	**39.6**	2.1	3.8	7.9	1.7	2.9	5.8	0.4	1.7	2.5		2.1
g	0.8	12.9	4.6	0.8	7.5			0.8	2.1	20.4	**25.8**	4.2	5.0	0.8	0.8	6.3	2.5	1.7	0.8	0.8	1.3
v	1.3	12.9	3.8	12.1	5.4		1.7	0.4	4.6	15.8	3.3	**15.8**	5.4	1.3	1.3	1.3	2.1	3.3	1.3	4.2	2.9
ð	0.4	11.3	0.8	4.2	8.3	3.3	1.7	1.3	2.5	29.2	2.1	10.0	**8.3**	2.9	1.3	7.9	0.4	1.3		0.8	2.1
z		3.8		1.7	10.0	12.1	2.1	2.5	0.4	8.8	2.1	5.0	10.0	**25.8**	5.0	2.1	2.1	1.7	1.3	1.3	2.5
ʒ		3.3	0.4		2.1	2.5	14.2	4.2	0.4	3.8	0.4	1.3	4.2	6.7	**45.0**	9.2	0.4	0.4		0.8	0.8
dʒ		2.9		1.7	2.9	0.8	2.9	13.3	0.4	8.8			4.2	0.4	9.6	**51.7**					0.4
m		9.2		0.4	4.6	0.8				5.8	1.3	2.1	2.1	0.8			**41.3**	20.0	8.3	2.1	1.3
n	0.4	9.6		1.3	1.3	0.4	0.4	0.4	0.8	8.3		0.4	2.1	2.1	0.4	2.1	8.3	**48.3**	9.2	2.1	2.1
ŋ		6.7	0.4	2.1	2.1	0.4			2.5	5.8	6.3		1.7	1.7			13.8	22.5	**30.4**	2.1	1.7
l	0.4	8.3	1.7	5.8	3.3		0.4			9.2	1.3	2.1	1.7	0.4			0.8	2.5	0.4	**57.5**	4.2
r	0.4	7.9	0.4	0.8	3.8		0.4		1.7	6.7		1.7	2.9	0.4	0.4	0.8		2.1		1.7	**67.9**

TABLE A.7: Confusion matrix for initial vowels at 0 dB SNR category for English listeners (Cutler et al., 2004).

	Response														
Stimulus	beat i	bit ɪ	wait eɪ	bet ɛ	bat æ	hot ɑ	cut ʌ	caught ɔ	boat oʊ	cook ʊ	boot u	buy aɪ	boy ɔɪ	shout aʊ	bird ɚ
i	**78.9**	8.3	0.3	2.7				0.3		0.3	1.8	1.2			3.9
ɪ	1.5	**81.8**	0.9	8.0	0.9		1.2	0.3		0.3	1.5	0.9			1.8
eɪ	5.7	5.4	**74.4**	4.5	5.7				0.3	0.3		0.3		0.3	1.2
ɛ	0.6	4.2	2.4	**84.2**	2.7		1.2					0.3			3.0
æ		1.2	6.5	3.9	**78.3**	0.6		1.2				0.3	0.3	4.8	2.1
ɑ		0.6	1.2	0.3	9.8	**42.3**	12.5	26.8	0.9			0.6		0.9	1.8
ʌ			0.3		1.2	12.5	**64.9**	8.3	1.8	1.2		1.2		4.2	3.0
ɔ		0.3	0.3		0.6	36.3	4.5	**47.3**	3.9	1.2		0.6	2.1	0.9	
oʊ		0.3				4.8	1.2	0.9	**69.6**	8.9	6.5		3.3	2.1	0.3
ʊ						2.1	14.0	2.1	0.9	**63.7**	6.8	0.3	3.0	2.1	0.9
u	3.6	1.5	0.3			0.6	3.0	1.5	1.8	19.3	**62.5**	0.3	1.2	1.8	1.2
aɪ		8.3	2.1				0.3					**87.2**	0.6		0.6
ɔɪ	0.3		0.6	0.3			0.3	0.6	1.2	0.6	0.3		**92.9**	3.0	
aʊ	0.3		0.3	2.1	0.6	3.3		7.1	2.4	0.3	0.3		0.9	**81.5**	0.3
ɚ	0.6	0.3		1.5			1.5	0.3							**95.5**

TABLE A.8: Confusion matrix for initial vowels at 0 dB SNR category for Dutch listeners (Cutler et al., 2004).

	Response														
Stimulus	beat i	bit ɪ	wait eɪ	bet ɛ	bat æ	hot ɑ	cut ʌ	caught ɔ	boat oʊ	cook ʊ	boot u	buy aɪ	boy ɔɪ	shout aʊ	bird ɚ
i	**75.6**	16.7	1.8	1.8		0.3	0.6			1.5			0.3	0.3	0.9
ɪ	1.5	**86.0**	0.6	5.4	0.3		0.3		0.3	0.3	1.5	1.2			2.4
eɪ	25.0	14.6	**46.7**	6.5	4.2		0.3	0.6	0.3				0.6		1.2
ɛ	0.3	12.5	0.9	**58.3**	25.0		0.3							0.3	2.1
æ		1.8	1.2	33.6	**56.3**	0.3	0.6					0.9		4.2	1.2
ɑ	0.3			0.6	13.4	**29.2**	31.5	15.5	1.2	0.3	0.3	3.9	0.3	1.2	1.8
ʌ			0.3	0.9	4.8	27.4	**44.3**	7.7	3.9	0.3	0.6	1.5	0.3	3.9	4.2
ɔ					0.6	63.4	5.4	**22.0**	2.4	0.6	0.9	0.6	3.0		1.2
oʊ			0.6			8.6	0.9	2.1	**53.0**	14.9	14.9	0.9	1.8	1.2	0.6
ʊ	0.6	0.3			0.3	11.6	4.2	2.4	3.9	**50.9**	17.3	0.6	3.6	2.4	1.5
u	8.6	1.5	0.3		0.3	2.7	0.9	1.8	4.8	46.7	**29.2**	0.3		1.8	1.2
aɪ		3.0	24.7	0.6	2.7		0.3	0.6	0.6			**64.3**	0.9		2.4
ɔɪ			1.5			3.6		0.3	0.9	0.6	0.3	0.9	**90.5**	1.2	0.3
aʊ			0.3	2.7	0.6	2.7		2.7	16.4	0.9	0.9	2.4	0.3	**70.2**	
ɚ	0.9	0.3	0.3	0.9		0.3	19.0	0.9	0.3			0.6		0.6	**75.9**

TABLE A.9: Confusion matrix for final vowels at 0 dB SNR category for English listeners (Cutler et al., 2004).

	Response														
Stimulus	beat i	bit ɪ	wait eɪ	bet ɛ	bat æ	hot ɑ	cut ʌ	caught ɔ	boat oʊ	cook ʊ	boot u	buy aɪ	boy ɔɪ	shout aʊ	bird ɚ
i	**93.5**	0.3		3.7				0.3			1.4			0.3	
ɪ	0.9	**84.4**	0.3	10.5	0.3		2.0		0.3					0.3	
eɪ	0.6	2.0	**91.5**	2.0	2.8							0.9			
ɛ	0.6	6.3	2.3	**73.6**	8.5		2.3	2.3		0.3		0.3		0.3	0.3
æ		0.6	1.1	12.2	**82.7**			1.1			0.3	0.3			0.3
ɑ			1.1	0.9	8.2	**33.5**	24.4	27.0	0.6	0.3		0.3		1.1	
ʌ			0.9	2.3	6.0	11.4	**65.3**	11.1	0.3	0.9	0.3		0.3	0.3	
ɔ			0.9		2.6	23.9	3.7	**65.3**	0.9	0.6				0.9	
oʊ	0.3	0.3		0.3		1.7		0.6	**90.6**	0.3	0.9		1.4	2.3	
ʊ	0.3			0.6		2.0	21.6	0.6	0.6	**68.2**	2.6	0.3	0.3		
u	3.7			0.9		0.3	0.9	0.6	0.3	6.8	**81.8**		0.3	2.6	
aɪ		6.0	0.9	0.3		0.3		0.3	0.3			**91.5**		0.3	
ɔɪ						0.6	0.3	0.9	1.4	0.9	0.3	0.3	**92.0**	2.3	
aʊ	0.3			0.9	0.6	2.0		1.7	8.2	0.6			1.7	**82.4**	0.6
ɚ	0.3	0.3		1.1			0.3								**97.7**

TABLE A.10: Confusion matrix for final vowels at 0 dB SNR category for Dutch listeners (Cutler et al., 2004).

	Response														
Stimulus	beat i	bit ɪ	wait eɪ	bet ɛ	bat æ	hot ɑ	cut ʌ	caught ɔ	boat oʊ	cook ʊ	boot u	buy aɪ	boy ɔɪ	shout aʊ	bird ɚ
i	**97.4**	0.3	0.3	1.7								0.3			
ɪ	0.3	**95.5**		2.3		0.3								0.3	1.4
eɪ	6.0	4.3	**84.1**	3.1	1.1			0.3			0.3	0.9			
ɛ	0.3	15.6	0.3	**60.5**	22.2			0.3						0.3	0.6
æ		0.9	4.8	39.2	**51.7**	0.6	1.1		0.3			0.6			0.9
ɑ	0.3	0.6		0.6	16.8	**23.6**	22.7	28.4	2.6	0.3	0.3	1.7	0.3	2.0	
ʌ		0.3	0.6	1.7	10.2	25.0	**41.5**	16.8	1.4	0.3	0.3	1.4		0.6	
ɔ					4.8	34.7	5.4	**50.0**	3.4		0.3		0.3	0.6	0.3
oʊ		0.6			0.3	11.1	0.6	4.8	**69.6**	2.8	6.0		2.6	1.7	
ʊ		0.6				5.4	4.3	4.5	1.7	**73.3**	6.8	0.6	1.4	0.6	0.9
u	17.9	0.3		0.6		0.6	1.4	1.1	0.9	31.3	**45.2**	0.3		0.6	
aɪ		3.4	14.8	0.3	1.1	0.3	0.6	0.6	0.9			**74.1**	1.4	0.3	2.3
ɔɪ			0.3			2.3		0.6	0.9	0.3		1.7	**93.8**	0.3	
aʊ		0.3	0.3	2.0	0.6	4.0	0.6	5.1	17.0	0.3	0.9	3.1	0.9	**64.5**	0.6
ɚ	0.3	0.3		0.9	0.3	0.6	7.1	0.3				0.6		0.3	**89.5**

TABLE A.11: Confusion matrix for consonant phonemes, derived from human speech recognition tests with re-synthesized speech at an SNR of 0 dB (Meyer et al., 2007).

	p	t	k	b	d	g	s	f	v	n	m	ʃ	ts	l
p	58	4	9	10	1	5		4	8					
t	2	74	5		10	4							4	
k	4	2	70		1	18			3				1	
b	11		2	44	3	13			26		1			2
d		2		4	74	13			3	3				
g	3		11	2	2	73			5	2	1			1
s							87	5				4	1	2
f	1		1				1	89	7			1		
v	1			16	1	6		3	70		2			1
n				1	4	1			2	61	12			19
m				4		3			5	12	58			18
ʃ							1	2				98		
ts		3					2						95	
l				2	4	2			2	4				88

TABLE A.12: Confusion matrix for consonant phonemes, derived from ASR experiments for which training and test data at 0 dB SNR were used (Meyer et al., 2007).

	p	t	k	b	d	g	s	f	v	n	m	ʃ	ts	l
p	54	1	11	16	7	3		1	5	1			1	
t	9	27	9	3	17	1			1				33	
k	16	2	55	3	7	9		1	1				6	
b	15	1	3	39	8	7		2	16	5	3		1	
d	3	3	1	6	60	9			4	5	2		3	5
g	9	1	14	9	13	39			8	4	2			2
s	1	1	1		1		83	5		1	1	3	5	
f	7		1	7	1		1	76	6			1	1	
v	3	1	2	27	7	8		9	36	2	3			3
n		1	1	3			1		3	61	18		1	11
m	1		1	5	1	1			4	31	54			2
ʃ							1	3				96		
ts		9	1				1	1			1		89	
l	1		1	3	2	2			2	19	3	1		66

TABLE A.13: Confusion matrix for vowel phonemes, derived from HSR tests with original signals at -10 dB SNR (Meyer et al., 2007).

	a	a:	ɛ	e	ɪ	i	ɔ	o	ʊ	u
a	80	13	4				4			
a:	5	85					7	3		
ɛ			83	11	6					
e			1	85	7	8				
ɪ			4	4	73	19			1	
i				9	3	89				
ɔ			2		1		82	5	9	
o							1	88	7	4
ʊ					4	1		7	69	19
u						2		14	3	81

which are confused at all of the analysis stages of the phoneme recognizer. The bold red phonemes are major confusions which appeared only in the posterior probability and phoneme recognizer confusions.

Figure A.1 only showed the major confusions for each phoneme. The phonemes are in order of probability for their respective columns. Many low probability confusions were eliminated however the majority of the total number of confusions are represented for each phone.

Phone	Pronunciation Confusions	Frame Confusions	Phoneme Confusions
iy	**ix**, **ih**	**ix**, ey, **ih**	**ix**, **ih**, y
ih	**ix**, **iy**, ax, **eh**	**ix**, **eh**, **iy**	**ix**, **eh**, **iy**
eh	**ih**, **ix**	*ae*, **ih**, *ah*, **ix**	**ih**, *ae*, *ah*, **ix**
ae	**eh**, ix	**eh**, *aw*, *ay*	**eh**, ah, *ay*, *aw*
ix	**ih**, **ax**, en, **iy**	**ih**, **ax**, **iy**	**ih**, **ax**, **iy**
ux		*uw*, *ih*, *ix*, *iy*	*uw*, *ix*, *ih*, *iy*
ax	**ix**, **ah**, ih	**ix**, **ah**	**ix**, **ah**
ax-h		*ix*, t, *ax*	*ix*, *ax*, p
uw	**ux**, ix, **uh**	**ux**, **uh**, *l*	**ux**, *l*, **uh**
uh	ix, er, **ax**	**ax**, ih, ix	**ax**, ux, ah
ah	**ax**, ix	*eh*, *aa*, **ax**	**ax**, *eh*, *aa*
ao	**aa**	**aa**	**aa**
aa	**ah**, **ao**	**ao**, *ay*, **ah**	**ao**, **ah**, *ay*
er	**axr**, ax, **r**	**axr**, **r**, *eh*	**axr**, **r**, *eh*
axr	**er**, **r**, **ax**, **ix**	**er**, **r**, **ix**, **ax**	**er**, **r**, **ix**, **ax**
ey	**eh**	*iy*, *ih*, **eh**	*ih*, *iy*, **eh**
ay	**aa**	**aa**, ah, ae	**aa**
oy	**ao**, **ow**	**ao**, **ow**	**ao**, r, **ow**
aw	**aa**	*ae*, **aa**, *ow*, ay, *eh*	*ae*, **aa**, *ow*, *eh*, l
ow	ax, uh	*l*, *ao*, *ah*	*l*, *ah*, *ao*
p		*t*, *k*, *f*	b, *t*, *k*, *f*
t	dx, q, d	*k*, *p*, ch, s	d, *p*, *k*
k		*t*, *p*, *g*	*t*, *g*, *p*
q			
b	**v**	*p*, **v**, dh	*p*, d, **v**
d	dx, **t**	**t**, jh, *g*, *dh*	*g*, **t**, *dh*
g		*k*, *d*, t	*d*, *k*
m	**em**	*n*, **em**	*n*, **em**
n	nx, en	*m*, *ng*, en	*m*, *ng*, nx
ng	**n**	**n**, *m*	**n**, *m*
nx		*n*, *dx*, *m*	*n*, *dx*, *m*
dx		*dh*, *v*, *n*, *nx*	*dh*, *nx*, d, *v*, *n*
f		*s*, *th*, *v*, z	*s*, *th*, *v*
th	**dh**, t	*f*, v, **dh**	*f*, t, **dh**, b
s	**sh**, **z**	**z**, *f*, **sh**	**z**, *f*, **sh**
sh		*s*, *zh*, *ch*	*s*, *zh*, *ch*
v	f	*dh*, *z*, f	*dh*, *z*, b
dh	th, d	v, f, th	d, b
z	**s**, zh	**s**, *v*	**s**, *v*, zh
zh	jh, z, **sh**, ch	**sh**, z, s	ux, **sh**, en
ch	**sh**	**sh**, *jh*, *t*, *s*	*t*, *jh*, **sh**, *s*
jh	zh	z, *ch*, zh	y, d, *ch*, t
l	**el**	*ow*, *w*, **el**	*w*, **el**, *ow*
r	**axr**, **er**	**er**, **axr**	**axr**, **er**
y	ix, ux	*iy*, *ih*, ux	*iy*, *ih*
w		*l*, *ao*, *uw*	*l*, *ao*, *uw*
em	m	m, uw, ax, *n*, en	ux, *n*, w
en	**ix**, **n**	**n**, **ix**, *m*, ng	**ix**, **n**, *m*
eng			
el	**l**	*ow*, **l**, *ao*, *ax*	**l**, *ow*, *ax*, *ao*
hv		*hh*	*hh*
hh	**hv**	**hv**, *q*, f	*q*, **hv**

FIGURE A.1: Only the major confusions for each phoneme are shown (Lovitt et al., 2007).

Phone Groups	Phone Groups
iy, ix, ih	hv, hh
ax, ah	ng, n, nx, en
ae, eh	th, dh
uw, ux	s, z
ao, aa	em, m
er, axr, r	b, v
zh, sh	el, l
short and long silences	

FIGURE A.2: 15 sets of phonemes which have prolific confusion patterns so the distinction is assumed to be arbitrary. Thus errors between members in a group are not counted as errors in the recognition class evaluation (Lovitt et al., 2007).

A.5 Vowels Phoneme Confusions for Malay

The vowels confusions for Malay were only studied based on three phonemes: /a/, /e/ or /ə/ and /i/

TABLE A.14: Phonemes confusion for vowels at initial position

Observed Phoneme Identified by Listeners		**a**	**e**	**i**	**o**	**u**
Speech Synthesiser's Phones Production	**a**	388	6	1	1	
	e	48	267	40	19	22
	i		18	360		

TABLE A.15: Phonemes confusion for vowels at final position

Observed Phoneme Identified by Listeners		**a**	**e**	**i**	**o**	**u**
Speech Synthesiser's Phones Production	**a**	392	4			
	e	82	210	47	40	17
	i	1	16	374	5	

TABLE A.16: Phonemes confusion for vowels at middle position

Observed Phoneme Identified by Listeners		**a**	**e**	**i**	**o**	**u**
Speech Synthesiser's Phones Production	**a**	766	16	1	7	2
	e	56	595	37	3	65
	i	3	23	764		2
	o				29	7

Appendix B

List of English Word List used in the Study

B.1 Word List for Intelligibility Test

In this section, the words are given without further information. Respondents were requested to identify all the words. When the respondents find themselves unable to provide a valid word for what they heard, they were requested to spell out to the best possible spelling to what they think they heard.

1. cheap
2. chip
3. job
4. jab
5. pay
6. pair
7. bang /bæng/
8. bank
9. beef
10. beev /biːv/
11. beige
12. beishe /beɪʃ/
13. mosse /mɒs/
14. mosque
15. dip
16. dib /dɪb/
17. fife /faɪf/
18. five
19. git
20. jit /dʒɪt/
21. ring
22. wing
23. hase /heɪs/
24. haze
25. kit
26. kig /kɪg/
27. loch
28. lokh /lɒkh/
29. judge
30. charge
31. march
32. marge
33. mass
34. maz /mæs/
35. measure /meʃə/
36. measure
37. mug
38. muk /mʌk/
39. peeg /piːg/
40. peak
41. yawn
42. worn
43. big
44. dig
45. pig
46. pik /pɪk/
47. pit
48. theen /θiːn/
49. thin
50. thees /ðiːs/
51. this
52. vail /veɪl/
53. fail
54. sure
55. zure /ʒʊə/
56. bat
57. pat
58. tat /tæt/
59. pid /pɪd/
60. tid /tɪd/
61. demper /dempə/
62. temper
63. sack
64. zack /zæk/
65. sip
66. zip

B.2 Word List for Contextual Perception Test

This word list consist of words with CVC structure. Most of the words have consisten onset and coda with other words in the list except for one word.

1. laugh, rough, loaf, lift
2. lid, wed, lead, load
3. sane, zen, son, sign
4. wood, wade, laid, weed
5. pale, pull, bull, pill
6. most, must, nest, mist
7. load, lead, red, loud
8. wait, wit, wet, yacht
9. sail, zeal, sell, soul
10. weak, walk, leak, wake
11. lamp, limp, lump, ramp
12. lame, lime, yam, loom
13. blank, blink, plank, blunk
14. fuss, vase, fiss, fess
15. bless, bliss, please
16. check, chick, jock, chuck
17. drunk, drank, drink, trunk
18. trick, truck, drake, track
19. spell, still, spool, spill
20. wrung, wrong, lung, wrung
21. brine, print, brain, bran
22. sting, stung, stink
23. brain, brine, train, bran
24. braid, trade, bread
25. bred, dread, bread
26. braid, pride, bred
27. green, groan, crane, grain
28. fine, vine, fawn
29. weak, rake, wake, walk
30. lime, warm, lame
31. lime, ram, lame, loom
32. lamp, limp, wimp, lump
33. block, black, plague, bleak
34. pipe, dope, pup
35. pope, boop, poop, pup
36. net, need, knit, not
37. need, net, nod
38. loaf, love, lift, laugh
39. would, wait, weed, wade
40. pull, pal, pay, pill
41. most, must, miss, mast
42. lead, let, load, loud
43. wait, what, weed, wet
44. sail, soul, say, sell
45. sane, sung, sin, sign
46. wake, wig, walk, week
47. lame, loon, lime, lum
48. black, blog, bleak, block
49. blank, blink, bland
50. chick, check, chug, choke
51. bless, blaze, bliss, blass
52. drunk, drink, dring, drunk
53. flick, flock, flag, flack
54. trick, truck, track, (in)trigue
55. track, trip, truck, trick
56. spell, spay, spill, spool
57. stuck, stag, stock, steak
58. rang, wrong, ran, wring
59. clock, cloak, clog, cluck
60. clock, click, clot, cluck
61. stiff, stove, stuff, staff
62. sting, stand, stung
63. string, stream, strung
64. brain, brine, brim, bran
65. grain, grin, groom, groan
66. hail, hall, haw, howl
67. fine, found, fund, fend
68. lamp, limp, lamb, lump
69. pipe, pup, pod, pope
70. tongue, tank, ting
71. fuss, faze, fiss, face

Appendix C

Malay Syllabification

This syllabification code is considered incomplete for Malay. There is no morphological analyser and no conversiton code for special case syllabification. For example, the word *terakru* will be syllabify as *te-rak-ru* while it is supposed to be *ter-ak-kru*. It will produces error when loan word like 'skirt' is run into the program.

```
#-------------------------------------------------------------------------
# Name          :   Phoneme2VCDG3.py                                      #
# Purpose       :   This program convert the phoneme transcription from #
#               input file: wordInPhoneme.txt to its corresponding      #
#               consonant-vowel-diphthongs-cluster consonants (each #
#               stand for:C-V-D-G respectively). Output will print  #
#               the phonemic transcription and its corresponding    #
#               VCDG into file: VCDG.txt
                                                                      #
# Author        :   Hana                                                  #
# Last Modified :   11/04/2013                                            #
# Copyright     :   (c) Hana 2013                                         #
# Licence       :   GPL (you cannot remove or modify the line prior to  #
#
          the #!/usr/bin/python)
                                                                          #
#-------------------------------------------------------------------------

#!/usr/bin/python

class wordFeatList(object):
    pass

vocab = wordFeatList()

############################## Copy input file content
diphList = ""
tempDiph = ""
diphFile = open("vowClustList.txt", 'r') # open file for reading
for diphLine in diphFile:
```

```
        clustDiph = diphLine.strip().split()
        tempDiph = ''.join(clustDiph)
        diphList = diphList + tempDiph + "\n"
diphFile.close()

vowList = ""
tempVow = ""
vowFile = open("vowList.txt", 'r') # open file for reading
for vowLine in vowFile:
        clustVow = vowLine.strip().split()
        tempVow = ''.join(clustVow)
        vowList = vowList + tempVow + "\n"
vowFile.close()

cCList = ""
tempCClust = ""
consClustFile = open("consClustList.txt", 'r') # open file for reading
for cClustLine in consClustFile:
        clustCons = cClustLine.strip().split()
        tempCClust = ''.join(clustCons)
        cCList = cCList + tempCClust + "\n"
consClustFile.close()

consList = ""
tempCons = ""
consFile = open("consList.txt", 'r') # open file for reading
for consLine in consFile:
        cons = consLine.strip().split()
        tempCons = ''.join(cons)
        consList = consList + tempCons + "\n"
consFile.close()
############################

w = open("wordInPhoneme.txt", 'r') # open file for reading
for line in w:
    try:
        if line:
            word = line.strip().split()
    except IndexError:
        break

    vocab.vWord = word[1]
    vocab.VCD = ""

    curDiph = ""
    curVow = ""
    curCC = ""
    curCons = ""

    index = 0
    clustPhone = None
    while index < len(vocab.vWord): #index ini mengira character dalam word
   semasa
        try:
            if vocab.vWord[index+1]:
                clustPhone = vocab.vWord[index] + vocab.vWord[index+1]

        except IndexError:
            clustPhone = ""
```

```
        if clustPhone: #still have two words bind
            #can have a few conditions.
            #The two characters can be diphthongs, consonant cluster,
            #or a consonant+another consonant or a consonant+vowel
            #or a vowel+consonant.
            #Need to handle the following first
            #1) diphthongs
            #2) consonant cluster (tS, dZ, ks)

            if clustPhone in diphList:
                curDiph = ''.join("D")
                vocab.VCD = vocab.VCD + curDiph
                index = index + 1
                curDiph = ""

            elif clustPhone in cCList:
                curCC = ''.join("G")
                vocab.VCD = vocab.VCD + curCC
                index = index + 1
                curCC = ""
            #If it is not the first two
            #Do the CVDG assignment individually
            #(to each phoneme)
            else:
                if vocab.vWord[index] in vowList:
                    curVow = ''.join('V')
                    vocab.VCD = vocab.VCD + curVow
                    curVow = ""

                elif vocab.vWord[index] in consList:
                    curCons = ''.join('C')
                    vocab.VCD = vocab.VCD + curCons
                    curCons = ""

        else: #definitely not a cluster phoneme, so process individually
    straight away
            if vocab.vWord[index] in vowList:
                curVow = ''.join('V')
                vocab.VCD = vocab.VCD + curVow
                curVow = ""
            if vocab.vWord[index] in consList:
                curCons = ''.join('C')
                vocab.VCD = vocab.VCD + curCons
                curCons = ""

        index = index + 1
     #print(vocab.VCD)
     #open file for appending words into the output file
     with open("VCDG.txt", "a") as wordVCDG:
        wordVCDG.write(vocab.vWord + "\t")
        wordVCDG.write(vocab.VCD + "\n")
     wordVCDG.close()

w.close()
```

```
#-------------------------------------------------------------------
# Name   : Syllabidication3.py
# Purpose: This program do syllabification. This program accept
#       input file: VCDG.txt. It will produce an output of
#   syllabify words
# Author        : Hana
# Last Modified : 14/04/2013
# Copyright     : Hana 2013
# Licence       : GPL (you cannot remove or modify the line prior
#                 to the #!/usr/bin/python)
#-------------------------------------------------------------------

#!/usr/bin/python


VCList = "" #don't change my location. I need to be a global variable

oriList = open("wordInPhoneme.txt", 'r') # open file for reading
for graphemeLine in oriList:
    grapheme = graphemeLine.strip().split()
    ortho = grapheme[0]
    phoTrans = grapheme[1]

    tempVC = ""
    tempVCList1 = ""
    tempVCList2 = ""
    tempVCList3 = ""

    VCDGFile = open("VCDG.txt", 'r') # open file for reading
    for vcdLine in VCDGFile:
            vcdg = vcdLine.strip().split()
            phonemic = vcdg[0]
            tempVC = vcdg[1]
            if phoTrans == phonemic:
                tempVCList1 = ''.join(ortho)
                tempVCList2= ''.join(phonemic)
                tempVCList3 = ''.join(tempVC)
                VCList = VCList + tempVCList1 + "\t" + tempVCList2 + "\t"+
   tempVCList3 + "\n"

                curSyl = "" #hold the current value of syllable structure
                curGrapheme = "" #hold the current value of grapheme
   structure
                curCVBreak = "" #hold the current value of CVCV structure
                curCV = "" #hold CVC value used to compare with the if-else
   rule
                index = 0 #generic counter
                CVCount = 0 #counter for CV string
                graCount = 0 #counter for grapheme string
                sylCount = 0 #counter for SAMPA string

                while index < len(tempVCList1): #using the orthographic/
   grapheme length for syllabification
                    try:
                        if tempVCList1:
                            curCV = tempVCList3[CVCount:CVCount+4] #curCV
   hold the CV structure to be compared in the if else list
                            #print("Nilai curCV sekarang ialah: ", curCV)
```

```
                        #print("Current index is: ", index)

                except IndexError:
                    break

                if (curCV == "CVCV" or curCV == "CVCD" or curCV == "CVVC"
 or curCV == "CVV"):
                    curCVBreak = curCVBreak + ''.join("CV-")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+2]+ "-")
                    if ("N" in tempVCList2[sylCount:sylCount+2]  or "J"
in tempVCList2[sylCount:sylCount+2]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3]+ "-")
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+2] + "-")
                        graCount = graCount + 2 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 2
                    index = graCount

                elif (curCV == "CVCC" or curCV == "CVCG"):
                    curCVBreak = curCVBreak + ''.join("CVC-")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+3]+ "-")
                    if ("N" in tempVCList2[sylCount:sylCount+3]  or "J"
in tempVCList2[sylCount:sylCount+3]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4]+ "-")
                        graCount = graCount + 4 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3] + "-")
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    CVCount = CVCount + 3
                    sylCount = sylCount + 3
                    index = graCount

                elif (curCV == "CDCV" or curCV == "CDCD"):
                    curCVBreak = curCVBreak + ''.join("CD-")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+3]+ "-")
                    if ("N" in tempVCList2[sylCount:sylCount+3]  or "J"
in tempVCList2[sylCount:sylCount+3]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4]+ "-")
                        graCount = graCount + 4 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3] + "-")
                        graCount = graCount + 3 #counter for grapheme-
break rep
```

```
                    CVCount = CVCount + 2
                    sylCount = sylCount + 3
                    index = graCount

                elif curCV == "CVC":#the last three remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("CVC")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+3])
                    if ("N" in tempVCList2[sylCount:sylCount+3]  or "J"
in tempVCList2[sylCount:sylCount+3]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4])
                        graCount = graCount + 4
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3])
                        graCount = graCount + 3
                    CVCount = CVCount + 3
                    sylCount = sylCount + 3
                    index = graCount

                elif curCV == "CDC": #the last syllable
                    curCVBreak = curCVBreak + ''.join("CDC")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+4])
                    if ("N" in tempVCList2[sylCount:sylCount+4]  or "J"
in tempVCList2[sylCount:sylCount+4]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+5])
                        graCount = graCount + 5 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4])
                        graCount = graCount + 4 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 3
                    index = graCount

                elif curCV == "CV": #only the last two remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("CV")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+2])
                    if ("N" in tempVCList2[sylCount:sylCount+2]  or "J"
in tempVCList2[sylCount:sylCount+2]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3])
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+2])
                        graCount = graCount + 2 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 2
```

```
                    index = graCount

                elif (curCV == "VCCD" or curCV == "VCCV"):
                    curCVBreak = curCVBreak + ''.join("VC-")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+2]+ "-")
                    if ("N" in tempVCList2[sylCount:sylCount+2]  or "J"
in tempVCList2[sylCount:sylCount+2]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3]+ "-")
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+2]+ "-")
                        graCount = graCount + 2 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 2
                    index = graCount

                elif curCV == "VC": #only the last two remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("VC")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+2])
                    if ("N" in tempVCList2[sylCount:sylCount+2]  or "J"
in tempVCList2[sylCount:sylCount+2]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3])
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+2])
                        graCount = graCount + 2 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 2
                    index = graCount

                elif curCV == "CD": #only the last two remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("CD")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+3])
                    if ("N" in tempVCList2[sylCount:sylCount+3]  or "J"
in tempVCList2[sylCount:sylCount+3]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4])
                        graCount = graCount + 4 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3])
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 3
```

```
                    index = graCount

                elif curCV == "GVCV": #only the last two remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("GV-")
                    if ("tS" in tempVCList2[sylCount:sylCount+3] or "dZ"
in tempVCList2[sylCount:sylCount+3]):
                        curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+3]+ "-")
                    else:
                        print("Not yet defined")
                    if ("N" in tempVCList2[sylCount:sylCount+3]  or "J"
in tempVCList2[sylCount:sylCount+3]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3]+ "-")
                        graCount = graCount + 3 #counter for grapheme-
break rep
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+2]+ "-")
                        graCount = graCount + 2 #counter for grapheme-
break rep
                    CVCount = CVCount + 2
                    sylCount = sylCount + 3
                    index = graCount

                elif curCV == "GVC":#the last three remain in the
grapheme
                    curCVBreak = curCVBreak + ''.join("GVC")
                    if ("tS" in tempVCList2[sylCount:sylCount+4] or "dZ"
in tempVCList2[sylCount:sylCount+4]):
                        curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+4])
                    else:
                        print("Not yet defined")
                    if ("N" in tempVCList2[sylCount:sylCount+4]  or "J"
in tempVCList2[sylCount:sylCount+4]):
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+4])
                        graCount = graCount + 4
                    else:
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+3])
                        graCount = graCount + 3
                    CVCount = CVCount + 3
                    sylCount = sylCount + 3
                    index = graCount

                elif (curCV == "V"):
                    curCVBreak = curCVBreak + ''.join("V")
                    curSyl = curSyl + ''.join(tempVCList2[sylCount:
sylCount+1])
                    curGrapheme = curGrapheme + ''.join(tempVCList1[
graCount:graCount+1])
                    graCount = graCount + 1 #counter for grapheme-break
rep
                    CVCount = CVCount + 1
                    sylCount = sylCount + 1
                    index = graCount
```

```
                    elif (curCV == "VCVC" or curCV == "VCVG" or curCV == "
    VGVC"):
                        curCVBreak = curCVBreak + ''.join("V-")
                        curSyl = curSyl + ''.join(tempVCList2[sylCount:
    sylCount+1]  + "-")
                        curGrapheme = curGrapheme + ''.join(tempVCList1[
    graCount:graCount+1] + "-")
                        graCount = graCount + 1 #counter for grapheme-break
    rep
                        CVCount = CVCount + 1
                        sylCount = sylCount + 1
                        index = graCount

                    elif curCV == "VCCV":
                        curCVBreak = curCVBreak + ''.join("VC-")
                        if ("N" in tempVCList2[sylCount:sylCount+2]  or "J"
    in tempVCList2[sylCount:sylCount+2]):
                            curGrapheme = curGrapheme + ''.join(tempVCList1[
    graCount:graCount+3]+ "-")
                            graCount = graCount + 3 #counter for grapheme-
    break rep
                        else:
                            curGrapheme = curGrapheme + ''.join(tempVCList1[
    graCount:graCount+2]+ "-")
                            graCount = graCount + 2 #counter for grapheme-
    break rep
                        CVCount = CVCount + 1
                        sylCount = sylCount + 1
                        index = graCount

                    else:
                        index = index + 1
                        CVCount = CVCount + 1
                        sylCount = sylCount + 1
                        graCount = graCount + 1
                print("\nCurent word is: ", tempVCList1)
                print("CV structure is: ", curCVBreak)
                print("Syllable structure of the grapheme is: ", curGrapheme)
                print("Syllable structue of the SAMPA is: ", curSyl)

                with open("Grapheme-Phoneme-Syllable.txt", "a") as syllabify:
                    syllabify.write('{0:<15} {1:^15} {2:^15} {3:^15} {4:^15}
     \n'.format(tempVCList1, curGrapheme, tempVCList2, curSyl, curCVBreak)) #
    this line print a correct format)
                syllabify.close()

                #print('{0:<10} {1:^10} {2:^10}  \n'.format(tempVC, phonemic,
     tempVC, )) #this line print a correct format
     VCDGFile.close()
     #print(VCList)
oriList.close()
#print(VCList)
```

Appendix D

The Iban Sentences List

1. Badu bemain ditu!
   Stop playing here!

2. Cis, berani nuan ngemeli ka aku!
   How date you insult me this way!

3. Oh, aku enda ingat baru ga!
   Owh, I have forgotten again!

4. Sida deka ngirup apa?
   What would they like to drink?

5. Kati nuan bisi hari?
   Are you free now?

6. Nama pemanjai pembaris tok?
   How long is this ruler?

7. Sapa penghulu ba endur tu?
   Who is the magisterial in this area?

8. Aku enda puas ati enggau nuan.
   I am really annoyed with you.

9. Minta anang bekenyak agi.
   Really appreciate if you don't repeat this again.

10. Kami matau Pulau Semaya.
    We hike around Semaya Island.

11. Anang nginti ba pegong tu.
    Do not fish in at this lake.

12. Aku beguna ka orang ke nemu ngemudi ka perau.
    I need someone who can steer a small boat.

13. Kadang-kadang iya ngingat ka ari biak iya.
    Sometimes he/she reminicense of all his story when he was young.

14. Pendiau enggau pemanah iya lalu lengkas diterima bala maioh.
    His/her charming manner had him immediately accepted into the community.

15. Orang ti empu pengawa nya enda betanggup jawap ngagai utai ti lenyau.
    The management is not answerable for any loss of personal belongings.

16. Pengereja pengawa kasar nya diukum sipat laban ditemu nyadong candu.
    The labour received corporal punishment for being caught red handed distributing opium.

17. Tuboh iya lumpoh laban iya patah tulang belakang leboh ke bebadi di jalai raya.
    His/her body is paralysed because of the broken back bone resulting from the car accident.

18. Nyawa iya tama lalat leboh tindok tengah hari tadi.
    A fly flew into his/her mouth while he/she was sleeping at noon.

19. Berasai bediri bulu pegu aku udah ninga cerita ti ngenakutka nya.
    I have a goosebump while I was listening the horror story.

20. Orang ke beperangai ngemaroh ka diri enggai ngulai orang ke sumbong.
    Humble people will also avoid condescending people.

21. Bansa Punan ngembuan pengelandik nyumpit ti ngelui ari orang mayuh.
    Punan people have the extraordinary expertise in bamboo shooting.

22. Penyanyi ke tebilang nya dirandau bala pemberita ba bilik alai belelak.
    The well known singer is being interviewed by the reporter in the lounge room.

23. Apai aku udah ngelekatka dua iti mentol lampu ba bilik mandi.
    My father has install two bulb of light in the wash room.

24. Serai, lia enggau lengkuas ditutok lumat-iumat dikena ngulai dagin nya.
    Lemongrass, ginger, galanger are pounded into fine fiber as the paste for the meat.

25. Minyak binjin ari Malaysia mayuh dijual ngagai menoa tasik.
    Malaysian petrol fuel are frequently exported to overseas.

26. Kami deka mela temuai nya makai.
    We are going to serve the food to those guests.

27. Tanah Besai Asia diseraraka Tasik Luar Atlantik ari Tanah Besai Amerika Utara.
    The Asian continent is separated by Atlantic Ocean from the North America's Continent.

28. Kati ulih aku betemu enggau Pengawai Pendidikan laban aku deka ngerejista anak aku masok besekula ba primari satu?
    May I see the Education Officer because I would like to register my child for the Standard One entry?

29. Anak mayau kaban aku udah parai.
    My friend's kitten is dead.

30. Aku deka ngagai pasar enggau sulu aku.
    I am going to the market with my boyfriend.

31. Terima kasih laban nulong aku.
    Thank you for helping me.

Appendix E

The Complete Respondents Feedback on Iban Polyglot Synthesiser

Table E.1: Overall Respondents Rating

		Mean	Mean Expert
Iban	Intelligibility	4.590804598	4.745747126
	Effort	4.070038314	4.570114943
	Likability	3.487279693	4.22183908
	Quality	3.374482759	4.090114943
Std Dev	Intelligibility	0.657449631	0.516147633
	Effort	0.853133467	0.756016608
	Likability	0.927649279	1.024354875
	Quality	0.836441783	1.06113606
Std Err	Intelligibility	0.120106352	0.094454296
	Effort	0.155949721	0.138597835
	Likability	0.169364812	0.187020757
	Quality	0.152712678	0.193736052
CI	Intelligibility	0.235408449	0.18513042
	Effort	0.305661453	0.271651756
	Likability	0.332428844	0.367982123
	Quality	0.29968841	0.380837349
Malay	Intelligibility	5	5
	Effort	4.866666667	5
	Likability	4.288888889	5
	Quality	4.266666667	5
Std Dev	Intelligibility	0.230940108	0
	Effort	0.230940108	0
	Likability	0.441253477	0
	Quality	0.19245009	0

Table E.2: A short summary of overall respondents rating

	Iban	**Malay**	**Std Dev Iban**	**Std Dev BM**
Intelligibility	4.590804598	5	0.657449631	0.230940108
Efford	4.070038314	4.866666667	0.853133467	0.230940108
Likability	3.487279693	4.288888889	0.927649279	0.441253477
Quality	3.374482759	4.266666667	0.836441783	0.19245009

Table E.3: Overall expert rating

	Iban	Malay	Malay Std Dev	Iban Std Dev
Intelligibility	4.745747126	5	0	0.516147633
Efford	4.570114943	5	0	0.756016608
Likability	4.22183908	5	0	1.024354875
Quality	4.090114943	5	0	1.06113606

Table E.4: Individual sounds intelligibility rating by all respondents

Question	**Mean**	**Std Dev**	**Std Err**
2	3.6	1.4040757	0.362530787
3	4.866666667	0.351865775	0.090851353
4	4.6	0.736788398	0.190237946
5	4.933333333	0.25819889	0.066666667
6	4.666666667	0.6172134	0.159363815
7	4.933333333	0.25819889	0.066666667
9	4.933333333	0.25819889	0.066666667
11	4.666666667	1.046536237	0.270214494
12	4.066666667	1.032795559	0.266666667
13	3.733333333	1.709915063	0.441498171
16	4	1.133893419	0.292770022
17	4.133333333	1.060098827	0.27371634
18	3.933333333	1.099783528	0.283962886
20	4.466666667	1.060098827	0.27371634
21	4.466666667	1.125462868	0.290593263
24	4.266666667	1.387014608	0.358125632
26	4.466666667	1.125462868	0.290593263
27	4.933333333	0.25819889	0.066666667
28	4.866666667	0.351865775	0.090851353
29	4.8	0.414039336	0.106904497
30	4.933333333	0.25819889	0.066666667
32	4.8	0.414039336	0.106904497
33	4.733333333	0.59361684	0.153271209
35	4.933333333	0.25819889	0.066666667
36	4.933333333	0.25819889	0.066666667
37	4.933333333	0.25819889	0.066666667
308	4.8	0.560611911	0.144749373
310	5	0	0
315	4.866666667	0.351865775	0.090851353
328	4.466666667	1.060098827	0.27371634

TABLE E.5: Individual sounds effort rating by all respondents

Question	**Mean**	**Std Dev**	**Std Err**
2	3.133333333	1.505545305	0.388730126
3	4.266666667	0.961150105	0.24816789
4	3.733333333	1.222799287	0.315725418
5	4	1	0.25819889
6	4.333333333	0.723746864	0.186870637
7	4.533333333	0.915475416	0.236374736
9	4.6	0.632455532	0.163299316
11	3.6	0.91025899	0.235027861
12	3.6	1.502379066	0.387912607
13	3.333333333	1.67616342	0.432783534
16	3.4	1.183215957	0.305505046
17	4.066666667	1.032795559	0.266666667
18	2.733333333	1.222799287	0.315725418
20	4.466666667	0.833809388	0.215288658
21	4.2	1.146423008	0.296005148
24	3.6	1.298350602	0.335232684
26	3.6	1.055597326	0.272554058
27	4.666666667	0.487950036	0.125988158
28	4.266666667	0.883715102	0.228174258
29	4	0.9258201	0.239045722
30	4.466666667	0.833809388	0.215288658
32	4.066666667	0.883715102	0.228174258
33	3.733333333	1.032795559	0.266666667
35	4.266666667	0.798808637	0.206251503
36	4.533333333	0.639940473	0.16523192
37	4.866666667	0.351865775	0.090851353
308	4.733333333	0.59361684	0.153271209
310	4.733333333	0.59361684	0.153271209
315	4.466666667	0.516397779	0.133333333
328	4.133333333	0.990430402	0.25572803

Table E.6: Individual sounds likability rating by all respondents

Question	**Mean**	**Std Dev**	**Std Err**
2	2.266666667	0.457737708	0.118187368
3	3.6	1.055597326	0.272554058
4	3	0.755928946	0.195180015
5	3.733333333	0.961150105	0.24816789
6	3.6	0.985610761	0.254483604
7	3.8	1.320173149	0.340867241
9	3.533333333	0.743223353	0.191899445
11	3.066666667	0.798808637	0.206251503
12	3.2	0.941123948	0.242997159
13	2.733333333	1.667618776	0.430577316
16	2.666666667	0.816496581	0.210818511
17	3.533333333	1.245945806	0.321701824
18	2.8	1.014185106	0.261861468
20	3.6	1.298350602	0.335232684
21	3.733333333	1.222799287	0.315725418
24	2.733333333	1.222799287	0.315725418
26	3.133333333	1.302013093	0.336178335
27	3.933333333	0.883715102	0.228174258
28	3.333333333	1.112697281	0.287297202
29	3.6	0.985610761	0.254483604
30	3.733333333	1.162919151	0.300264434
32	4.133333333	0.915475416	0.236374736
33	3.933333333	0.961150105	0.24816789
35	3.6	0.91025899	0.235027861
36	4.2	0.560611911	0.144749373
37	4.4	0.632455532	0.163299316
308	3.8	1.014185106	0.261861468
310	3.933333333	1.099783528	0.283962886
315	3.466666667	0.990430402	0.25572803
328	3.333333333	1.2344268	0.318727629

Table E.7: Individual sounds quality rating by all respondents

Question	**Mean**	**Std Dev**	**Std Err**
2	2.466666667	1.060098827	0.27371634
3	3.466666667	0.743223353	0.191899445
4	3	1.253566341	0.323669437
5	3.6	1.055597326	0.272554058
6	3.533333333	0.833809388	0.215288658
7	3.8	0.774596669	0.2
9	3.6	0.736788398	0.190237946
11	3	0.9258201	0.239045722
12	3.133333333	1.125462868	0.290593263
13	3.2	1.264911064	0.326598632
16	2.533333333	0.990430402	0.25572803
17	3.733333333	1.032795559	0.266666667
18	2.533333333	0.833809388	0.215288658
20	3.466666667	1.060098827	0.27371634
21	3.466666667	1.125462868	0.290593263
24	2.4	0.632455532	0.163299316
26	3.066666667	1.279880947	0.330463839
27	3.733333333	0.961150105	0.24816789
28	3.133333333	0.743223353	0.191899445
29	3.2	0.774596669	0.2
30	3.866666667	0.915475416	0.236374736
32	3.8	1.082325539	0.279455252
33	3.866666667	0.915475416	0.236374736
35	3.733333333	0.703731551	0.181702705
36	4.066666667	0.703731551	0.181702705
37	4.266666667	0.59361684	0.153271209
308	3.666666667	0.899735411	0.232310684
310	3.866666667	0.990430402	0.25572803
315	3.866666667	0.833809388	0.215288658
328	3.6	1.121223821	0.289498746

TABLE E.8: Individual sounds intelligibility rating by experts

Question	**Mean**	**Std Dev**	**Std Err**
2	4.2	0.836660027	0.374165739
3	4.6	0.547722558	0.244948974
4	4.4	0.894427191	0.4
5	5	0	0
6	4.6	0.547722558	0.244948974
7	5	0	0
9	5	0	0
11	5	0	0
12	4	0.707106781	0.316227766
13	5	0	0
16	4	0.707106781	0.316227766
17	4.4	0.894427191	0.4
18	4	0.707106781	0.316227766
20	5	0	0
21	4.4	0.894427191	0.4
24	5	0	0
26	4.8	0.447213595	0.2
27	5	0	0
28	4.8	0.447213595	0.2
29	4.6	0.547722558	0.244948974
30	5	0	0
32	5	0	0
33	4.8	0.447213595	0.2
35	5	0	0
36	5	0	0
37	5	0	0
308	5	0	0
310	5	0	0
315	5	0	0
328	4.8	0.447213595	0.2

TABLE E.9: Individual sounds effort rating by experts

Q	Mean	Std Dev	Std Err
2	3.8	0.447213595	0.2
3	4.4	0.894427191	0.4
4	3.8	1.643167673	0.734846923
5	5	0	0
6	4.8	0.447213595	0.2
7	5	0	0
9	5	0	0
11	3.4	0.894427191	0.4
12	4.2	0.447213595	0.2
13	4.6	0.547722558	0.244948974
16	3.4	0.894427191	0.4
17	4.6	0.547722558	0.244948974
18	2.6	0.894427191	0.4
20	5	0	0
21	4.8	0.447213595	0.2
24	5	0	0
26	4.2	0.447213595	0.2
27	5	0	0
28	5	0	0
29	5	0	0
30	5	0	0
32	5	0	0
33	5	0	0
35	4	0	0
36	5	0	0
37	5	0	0
308	4.8	0.447213595	0.2
310	5	0	0
315	4.8	0.447213595	0.2
328	5	0	0

TABLE E.10: Individual sounds likeability rating by experts

Question	**Mean**	**Std Dev**	**Std Err**
2	2.2	0.447213595	0.2
3	3.2	0.447213595	0.2
4	3	1	0.447213595
5	5	0	0
6	3.8	0.447213595	0.2
7	4.4	0.894427191	0.4
9	4.2	0.447213595	0.2
11	3.6	0.894427191	0.4
12	3.8	0.447213595	0.2
13	4.8	0.447213595	0.2
16	2.8	0.447213595	0.2
17	4.6	0.547722558	0.244948974
18	2	1	0.447213595
20	4.8	0.447213595	0.2
21	4.8	0.447213595	0.2
24	3.6	1.516575089	0.678232998
26	4.6	0.894427191	0.4
27	4.8	0.447213595	0.2
28	4	1.224744871	0.547722558
29	4.4	0.894427191	0.4
30	5	0	0
32	4.8	0.447213595	0.2
33	5	0	0
35	3.6	0.894427191	0.4
36	4.8	0.447213595	0.2
37	5	0	0
308	4.8	0.447213595	0.2
310	5	0	0
315	4.2	1.095445115	0.489897949
328	4.6	0.894427191	0.4

TABLE E.11: Individual sounds quality rating by experts

Question	**Mean**	**Std Dev**	**Std Err**
2	3.2	0.447213595	0.2
3	3.8	0.836660027	0.374165739
4	3.8	0.447213595	0.2
5	5	0	0
6	4.4	0.547722558	0.244948974
7	4.4	0.894427191	0.4
9	4.4	0.547722558	0.244948974
11	3.8	1.095445115	0.489897949
12	4.2	0.447213595	0.2
13	4.8	0.447213595	0.2
16	2	1	0.447213595
17	4.6	0.547722558	0.244948974
18	2	1.224744871	0.547722558
20	4.8	0.447213595	0.2
21	4.8	0.447213595	0.2
24	2.8	0.836660027	0.374165739
26	4.6	0.894427191	0.4
27	4.6	0.547722558	0.244948974
28	3	1	0.447213595
29	3.8	0.836660027	0.374165739
30	5	0	0
32	4.8	0.447213595	0.2
33	5	0	0
35	4.2	0.836660027	0.374165739
36	4.8	0.447213595	0.2
37	5	0	0
308	4.8	0.447213595	0.2
310	5	0	0
315	4.8	0.447213595	0.2
328	4.8	0.447213595	0.2

TABLE E.12: Complete individual sounds rating by all Iban respondents (number is rounded to two decimal points for readability)

	Intelligibility			**Effort**			**Likability**			**Quality**		
Question	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**
2 (3)	3.60	1.40	0.36	3.13	1.51	0.39	2.27	0.46	0.12	2.47	1.06	0.27
3 (6)	4.87	0.35	0.09	4.27	0.96	0.25	3.60	1.06	0.27	3.47	0.74	0.19
4 (6)	4.60	0.74	0.19	3.73	1.22	0.32	3.00	0.76	0.20	3.00	1.25	0.32
5 (4)	4.93	0.26	0.07	4.00	1.00	0.26	3.73	0.96	0.25	3.60	1.06	0.27
6 (4)	4.67	0.62	0.16	4.33	0.72	0.19	3.60	0.99	0.25	3.53	0.83	0.22
7 (5)	4.93	0.26	0.07	4.53	0.92	0.24	3.80	1.32	0.34	3.80	0.77	0.20
9 (6)	4.93	0.26	0.07	4.60	0.63	0.16	3.53	0.74	0.19	3.60	0.74	0.19
11 (4)	4.67	1.05	0.27	3.60	0.91	0.24	3.07	0.80	0.21	3.00	0.93	0.24
12 (5)	4.07	1.03	0.27	3.60	1.50	0.39	3.20	0.94	0.24	3.13	1.13	0.29
13 (8)	3.73	1.71	0.44	3.33	1.68	0.43	2.73	1.67	0.43	3.20	1.26	0.33
16 (9)	4.00	1.13	0.29	3.40	1.18	0.31	2.67	0.82	0.21	2.53	0.99	0.26
17 (12)	4.13	1.06	0.27	4.07	1.03	0.27	3.53	1.25	0.32	3.73	1.03	0.27
18 (10)	3.93	1.10	0.28	2.73	1.22	0.32	2.80	1.01	0.26	2.53	0.83	0.22
20 (14)	4.47	1.06	0.27	4.47	0.83	0.22	3.60	1.30	0.34	3.47	1.06	0.27
21 (9)	4.47	1.13	0.29	4.20	1.15	0.30	3.73	1.22	0.32	3.47	1.13	0.29
24 (10)	4.27	1.39	0.36	3.60	1.30	0.34	2.73	1.22	0.32	2.40	0.63	0.16
26 (11)	4.47	1.13	0.29	3.60	1.06	0.27	3.13	1.30	0.34	3.07	1.28	0.33
27 (11)	4.93	0.26	0.07	4.67	0.49	0.13	3.93	0.88	0.23	3.73	0.96	0.25
28 (11)	4.87	0.35	0.09	4.27	0.88	0.23	3.33	1.11	0.29	3.13	0.74	0.19
29 (9)	4.80	0.41	0.11	4.00	0.93	0.24	3.60	0.99	0.25	3.20	0.77	0.20
30 (6)	4.93	0.26	0.07	4.47	0.83	0.22	3.73	1.16	0.30	3.87	0.92	0.24
32 (12)	4.80	0.41	0.11	4.07	0.88	0.23	4.13	0.92	0.24	3.80	1.08	0.28
33 (16)	4.73	0.59	0.15	3.73	1.03	0.27	3.93	0.96	0.25	3.87	0.92	0.24
35 (6)	4.93	0.26	0.07	4.27	0.80	0.21	3.60	0.91	0.24	3.73	0.70	0.18
36 (6)	4.93	0.26	0.07	4.53	0.64	0.17	4.20	0.56	0.14	4.07	0.70	0.18
37 (5)	4.93	0.26	0.07	4.87	0.35	0.09	4.40	0.63	0.16	4.27	0.59	0.15
308 (5)	4.80	0.56	0.14	4.73	0.59	0.15	3.80	1.01	0.26	3.67	0.90	0.23
310 (4)	5.00	0.00	0.00	4.73	0.59	0.15	3.93	1.10	0.28	3.87	0.99	0.26
315 (8)	4.87	0.35	0.09	4.47	0.52	0.13	3.47	0.99	0.26	3.87	0.83	0.22
328 (11)	4.47	1.06	0.27	4.13	0.99	0.26	3.33	1.23	0.32	3.60	1.12	0.29

TABLE E.13: Complete individual sounds rating by all expert respondents (numbers are reduced to four decimal points for readability)

	Intelligibility			**Effort**			**Likability**			**Quality**		
Question	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**	**Mean**	**Std Dev**	**Std Err**
2	4.2000	0.8367	0.3742	3.8000	0.4472	0.2000	2.2000	0.4472	0.2000	3.2000	0.4472	0.2000
3	4.6000	0.5477	0.2449	4.4000	0.8944	0.4000	3.2000	0.4472	0.2000	3.8000	0.8367	0.3742
4	4.4000	0.8944	0.4000	3.8000	1.6432	0.7348	3.0000	1.0000	0.4472	3.8000	0.4472	0.2000
5	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
6	4.6000	0.5477	0.2449	4.8000	0.4472	0.2000	3.8000	0.4472	0.2000	4.4000	0.5477	0.2449
7	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.4000	0.8944	0.4000	4.4000	0.8944	0.4000
9	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.2000	0.4472	0.2000	4.4000	0.5477	0.2449
11	5.0000	0.0000	0.0000	3.4000	0.8944	0.4000	3.6000	0.8944	0.4000	3.8000	1.0954	0.4899
12	4.0000	0.7071	0.3162	4.2000	0.4472	0.2000	3.8000	0.4472	0.2000	4.2000	0.4472	0.2000
13	5.0000	0.0000	0.0000	4.6000	0.5477	0.2449	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
16	4.0000	0.7071	0.3162	3.4000	0.8944	0.4000	2.8000	0.4472	0.2000	2.0000	1.0000	0.4472
17	4.4000	0.8944	0.4000	4.6000	0.5477	0.2449	4.6000	0.5477	0.2449	4.6000	0.5477	0.2449
18	4.0000	0.7071	0.3162	2.6000	0.8944	0.4000	2.0000	1.0000	0.4472	2.0000	1.2247	0.5477
20	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
21	4.4000	0.8944	0.4000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
24	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	3.6000	1.5166	0.6782	2.8000	0.8367	0.3742
26	4.8000	0.4472	0.2000	4.2000	0.4472	0.2000	4.6000	0.8944	0.4000	4.6000	0.8944	0.4000
27	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.6000	0.5477	0.2449
28	4.8000	0.4472	0.2000	5.0000	0.0000	0.0000	4.0000	1.2247	0.5477	3.0000	1.0000	0.4472
29	4.6000	0.5477	0.2449	5.0000	0.0000	0.0000	4.4000	0.8944	0.4000	3.8000	0.8367	0.3742
30	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
32	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
33	4.8000	0.4472	0.2000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
35	5.0000	0.0000	0.0000	4.0000	0.0000	0.0000	3.6000	0.8944	0.4000	4.2000	0.8367	0.3742
36	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
37	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
308	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000	4.8000	0.4472	0.2000
310	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000	5.0000	0.0000	0.0000
315	5.0000	0.0000	0.0000	4.8000	0.4472	0.2000	4.2000	1.0954	0.4899	4.8000	0.4472	0.2000
328	4.8000	0.4472	0.2000	5.0000	0.0000	0.0000	4.6000	0.8944	0.4000	4.8000	0.4472	0.2000

Bibliography

[1] Shahbudin Abdullah. *Sejarah Perkembangan Tulisan Jawi*. Internet. Accessed on December 2013. 2010. URL: `https://www.academia.edu/4903272/Sejarah\_Perkembangan\_Tulisan\_Jawi`.

[2] Zaharani Ahmad, Nor Hashimah Jalaluddin, Fazal Mohamed Mohamed Sultan, Harishon Radzi, and Mohd Shabri Yusof. "Pemerkasaan Jati Diri Bahasa Melayu: Isu Penyerapan Kata Asing translated as Empowerment and Configuration Malay Language Identity: Infiltration of Foreign Vocabularies Issue". In: *Jurnal Melayu* 6 (2011), pp. 13–27.

[3] S. Takdir Alisjahbana. *Dari Perdjuangan dan Pertumbuhan Bahasa Indonesia*. Vol. 3. 2. Pustaka Rakyat, 1957, pp. 66 –81.

[4] B. Baughman. *Speaking Iban*. Ed. by James T. Reuteler. CreateSpace Independent Publishing Platform, 2012.

[5] Alessandro Bausani. "The First Italian-Malay Vocabulary by Antonio Pigafetta". In: *Journal of East and West, NS* 11(4) (1960), pp. 229–248.

[6] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. "Automatic Speech Recognition for Under-resourced Languages: A Survey". In: *Speech Communication* 56 (2014), pp. 85–100.

[7] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. "The AT&T Next-Gen TTS System". In: *Joint Meeting of ASA, EAA, AND DAGA*. 1999, pp. 18–24.

[8] Alan W Black. "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling". In: *INTERSPEECH*. 2006, pp. 1762–1765.

[9] Alan W. Black and Andrew Hunt. "Generating F0 Contours from ToBI Labels Using Linear Regression". In: *ICSLP 1996*. ISCA, 1996.

[10] Alan W Black and Tanja Schultz. "Speaker Clustering for Multilingual Synthesis". In: *Proceedings of the ISCA Tutorial and Researc Workshop on Multilingual Speech and Language Processing*. Stellenbosch, South Africa, 2006.

[11] Alan W Black and Paul Taylor. "CHATR: A Generic Speech Synthesis System". In: *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*. Association for Computational Linguistics. 1994, pp. 983–986.

[12] Dwight Bolinger. "Accent is predictable (if you're a mind-reader)". In: *Language* (1972), pp. 633–644.

[13] Adam Brown. "The Staccato Effect in the Pronunciation of English in Malaysia and Singapore". In: *New Englishes: The Case of Singapore* (1988), pp. 115–128.

[14] Nick Campbell. "CHATR: A High-Definition Speech Re-Sequencing System". In: *Proceeding of 3rd ASA/ASJ Joint meeting*. 1996, pp. 1223–1228.

[15] Nick Campbell. *Segmental Elasticity and Timing in Japanese Speech*. IOS Press, 1992.

[16] Nick Campbell. "Timing in Speech: A Multi-Level Process". In: *Prosody: Theory and Experiment: Studies Presented to Gösta Bruce*. Ed. by Merle Horne. Dordrecht: Springer Netherlands, 2000, pp. 281–334.

[17] R. Carlson and B. Granström. "A Phonetically Oriented Programming Language for Rule Description of Speech". In: *Speech Communication* 2 (1975), pp. 245–253.

[18] Rolf Carlson and B. Granström. "Word Accent, Emphatic Stress, and Syntax in a Synthesis-by-Rule Scheme for Swedish". In: *Speech Transmission Lab QPSR* 1973 (1973), pp. 2–3.

[19] Jonathan Charteris-Black. "Second Language Fgurative Proficiency: A Comparative Study of Malay and English". In: *Applied linguistics* 23(1) (2002), pp. 104–133.

[20] Beverly A. Clark. "First- and Second-Language Acquisition in Early Childhood". In: *Issues in Early Childhood Education: Curriculum, Teacher Education, & Dissemination of Information. Proceedings of the Lilian Katz Symposium*. Champaign, IL, 2000, pp. 835–838. URL: `http://ecap.crc.illinois.edu/pubs/katzsym/clark-b.pdf`.

[21] Adrian Clyness and David Deterding. “Standard Malay (Brunei)”. In: *Journal of the International Phonetic Association* 41(2) (2011), pp. 259–268.

[22] Cynthia M Colwell and Kathleen D Murlless. “Music activities (singing vs. chanting) as a Vehicle for Reading Accuracy of Children with Learning Disabilities: A pilot study”. In: *Music Therapy Perspectives* 20(1) (2002), pp. 13–19.

[23] Franklin S Cooper, Pierre C Delattre, Alvin M Liberman, John M Borst, and Louis J Gerstman. “Some Experiments on the Perception of Synthetic Speech Sounds”. In: *The Journal of the Acoustical Society of America* 24(6) (1952), pp. 597–606.

[24] David Crystal. “Prosodic Development in The Transition into Language”. In: *Language Acquisition: Studies in First Language Development*. Ed. by Paul Fletcher and Michael Garman. Cambridge University Press, 1986, pp. 33–48.

[25] Thomas H. Crystal and Arthur S. House. “A Note on the Variability of Timing Control”. In: *Journal of Speech, Language, and Hearing Research* 31(3) (1988), pp. 497–502.

[26] Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. “Patterns of English Phoneme Confusions by Native and Non-native Listeners”. In: *The Journal of the Acoustical Society of America* 116(6) (2004), pp. 3668–3678.

[27] DBP. *Pusat Rujukan Persuratan Melayu*. Internet. Accessed on December 2013. 2008. URL: `http://prpm.dbp.gov.my/`.

[28] Michel Divay and Anthony J. Vitale. “Algorithms for Grapheme-Phoneme Translation for English and French: Applications for Database Searches and Speech Synthesis”. In: *Journal of Computational Linguistics* 23 (1997), pp. 495–523.

[29] Zuraidah Mohd Don, Gerry Knowles, and Janet Yong. “How Words can be Misleading: A Study of Syllable Timing and “Stress” in Malay”. In: *The Linguistics Journal* 3(2) (2008), pp. 66 –81.

[30] Minghui Dong, Haizhou Li, and Tin Lay Nwe. “Evaluating Prosody of Mandarin Speech for Language Learning”. In: *Journal of Chinese Language and Computing* 17(4) (2007), pp. 219–226.

[31] Kurt Dusterhoff and Alan W. Black. “Generating F0 Contours for Speech Synthesis Using the Tilt Intonation Theory”. In: *Proceedings of the ESCA Workshop 1997 on Intonation*. Athens, Greece, 1997.

[32] Yousif A. El-Imam and Zuraidah Mohd Don. "Rules and Algorithms for Phonetic Transcription of Standard Malay". In: *IEICE - Transaction of Information System* E88-D(10) (2005), pp. 2354–2372. ISSN: 0916-8532.

[33] G. Fant, B. Lindblom, and A. de Serpa-Leitao. "Consonant Confusions in English and Swedish. A pilot study". In: *STL-QPSR* 7(4) (1966), pp. 31–34.

[34] Douglas Fisher. "Early Language Learning with and without Music". In: *Reading Horizons* 42(1) (2001), p. 8.

[35] Hiroya Fujisaki and H Sudo. "Synthesis by Rule of Prosodic Features of Connected Japanese". In: *International Congress on Acoustics*. Budapest, 1971, pp. 133–136.

[36] J.B. Gilbert. *Clear Speech Teacher's Resource and Assessment Book: Pronunciation and Listening Comprehension in North American English*. Clear speech. Cambridge University Press, 2012.

[37] Mon-Shen Goh. *Malay TTS using Microsoft SAPI*. Tech. rep. Computer Aided Translation Unit, Universiti Sains Malaysia, 2004.

[38] Xavi Gonzalvo, Ignasi Iriondo, Joan Claudi Socoró, Francesc Alías, and Carlos Monzo. "HMM-based Spanish Speech Synthesis using CBR as F0 Estimator". In: *Proceesing of ISCA Tutorial and Research Workshop (ITRW) on Non Linear Speech Processing (NOLISP07)*. 2007, pp. 7 –10.

[39] Xavi Gonzalvo, Ignasi Iriondo, Joan Claudi Socoró, Francesc Alías, and Carlos Monzo. "Mixing HMM-Based Spanish Speech Synthesis with a CBR for Prosody Estimation". In: *Advances in Nonlinear Speech Processing*. Vol. 4885/2007. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007, pp. 78–85.

[40] Matthew Gordon. "A Factorial Typology of Quantity-Insensitive Stress". English. In: *Natural Language & Linguistic Theory* 20(3) (2002), pp. 491–552.

[41] Amran Halim. *Intonasi Dalam Hubungannya Dengan Sintaksis Bahasa Indonesia*. Seri ILDEP. Djambatan, 1984.

[42] Johan 't Hart, Rene Collier, and Antonie Cohen. *A Perceptual Study of Intonation - An Experimental-phonetic Approach to Speech Melody*. Institute for Perception Research, Eindhoven: Cambridge University Press, 1990.

[43] John M. Heinz and Kenneth N. Stevens. "On the Properties of Voiceless Fricative Consonants". In: *Journal of Acoustical Society of America* 33(5) (1961), pp. 589–596. DOI: 10.1121/1.1908734.

[44] Daniel J. Hirst. "Automatic Analysis of Prosody for Multilingual Speech Corpora". In: *Improvements in Speech Synthesis.* Ed. by E. Keller, G. Bailly, J.Terken, and M. Huckvale. Chichester, United Kingdom: Wiley Publisher, 2001, pp. 320–327.

[45] Arthur S House and Grant Fairbanks. "The influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels". In: *The Journal of the Acoustical Society of America* 25(1) (1953), pp. 105–113.

[46] William Howell and Demetrius James Sandford Bailey. *A Sea Dayak Dictionary in Alphabetical Parts, with examples and quotations shewing the use and meaning of words.* English. American Mission Press, 1900, pp. 596–616.

[47] Xeudong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* New Jersey: Prentice Hall PTR, 2001.

[48] Barbara Höhle, Ranka Bijeljac-Babic, Birgit Herold, Jürgen Weissenborn, and Thierry Nazzi. "Language Specific Prosodic Preferences During the First Half Year of Life: Evidence from German and French Infants". In: *Journal of Infant Behavior and Development* 32(3) (2009), pp. 262 –274. ISSN: 0163-6383.

[49] IPG. "Sejarah Perkembangan Bahasa Melayu, Perkamusan dan Terjemahan". In: ed. by Kementerian Pendidikan Malaysia. Dewan Bahasa Pustaka, 2011. Chap. Perkembangan Sistem Ejaan Rumi Bahasa Melayu, pp. 133–158.

[50] Sarah FS Juan, Vyonne Edwin, Chai Yeen Cheong, Lee Jun Choi, and Alvin W Yeo. "Adopting Malay Syllable Structure for Syllable Based Speech Synthesizer for Iban and Bidayuh Languages". In: *2011 International Conference on Asian Language Processing (IALP).* IEEE. 2011, pp. 216–219.

[51] Sarah Samson Juan and Laurent Besacier. "Fast Bootstrapping of Grapheme to Phoneme System for Under-resourced Languages - Application to the Iban Language". In: *WSSANLP 2013* (2013), p. 1.

[52] Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Tien-Ping Tan. "Using Closely-related Language to build an ASR for a very Under-resourced language: Iban". In: *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA 2014).* IEEE. 2014, pp. 1–5.

[53] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (Prentice Hall Series in Artificial Intelligence)*. 2nd ed. Prentice Hall, 2008.

[54] H. Kähler. *Grammatik der Bahasa Indonesia*. Vol. 3. 2. O. Harrassowitz, 1956, pp. 66 –81.

[55] Nik Safiah Karim, Farid M. Onn, Hashim Musa, and Abdul Hamid Mahmood. *Tatabahasa Dewan: Edisi Baharu*. 3rd ed. Kuala Lumpur: Dewan Bahasa dan Pustaka (DBP), 1996.

[56] Roy Kennedy and Amanda Scott. "A Pilot Study: The Effects of Music Therapy Interventions on Middle School Students' ESL Skills". In: *Journal of Music Therapy* 42(4) (2005), pp. 244–261.

[57] Yen-Min Jasmina Khaw and Tien-Ping Tan. "Preparation of MaDiTS Corpus for Malay Dialect Translation and Speech Synthesis System". In: *Proceedings of the 2nd International Workshop on Speech, Language and Audio in Multimedia 2014*. School of Computer Sciences, USM, 2014, pp. 53 –57.

[58] Katrin Kirchhoff. "Language Characteristics". In: *Multilingual Speech Processing*. Ed. by Tanja Schultz and Katrin Kirchhoff. San Diego, California: Elsevier, 2006, pp. 5–31.

[59] Dennis H. Klatt. "Linguistic uses of Segmental Duration in English: Acoustic and Perceptual Evidence". In: *Journal of the Acoustical Society of America* 59 (1976), pp. 1208–1221.

[60] Dennis H. Klatt. "Review of Text-to-Speech Conversion for English". In: *Journal of the Acoustical Society of America* 82(3) (1987), pp. 737–793.

[61] Dennis H. Klatt. "Synthesis by rule of segmental durations in English sentences". In: *Frontiers of Speech Communication Research* (1979), pp. 287–299.

[62] Dennis H Klatt. "Vowel Lengthening is Syntactically Determined in a Connected Discourse". In: *Journal of Phonetics* 3(3) (1975), pp. 129–140.

[63] Dennis H. Klatt and William E. Cooper. *Perception of Segment Duration in Sentence Contexts*. Ed. by A. Cohen and S.G. Nooteboom. Berlin: Springer Verlag, 1960, pp. 69 –89.

[64] Dennis H. Klatt and Laura C. Klatt. "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers". In: *Journal of the Acoustical Society of America* 87(2) (1990), pp. 820–857.

[65] John Kominek. "TTS from Zero: Building Synthetic Voices for New Languages". PhD thesis. Language Technologie Institute, School of Computer Science, Cernegie Melon University, 2009.

[66] John Kominek, Tanja Schultz, and Alan W Black. "Voice Building from Insufficient Data - Classroom Experience with web-based Development Tools". In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*. Bonn, Germany, 2007.

[67] D Robert Ladd and Nick Campbell. "Theories of Prosodic Structure: Evidence from Syllable Duration". In: *Proceedings of the 12th International Congress of Phonetic Sciences*. Vol. 2. 1991, pp. 290–293.

[68] D.R. Ladd and K. Silverman. "Vowel Intrinsic Pitch in Connected Speech". In: *Phonetica* 41 (1984), pp. 31–40.

[69] Peter Ladefoged and Keith Johnson. *A Course in Phonetics*. 6th. Boston: Thomson Wadsworth, 2010.

[70] L. S. Larkey. "Reiterant Speech: An Acoustic and Perceptual Validation". In: *The Journal of the Acoustical Society of America* 73 4 (1983), pp. 1337–45.

[71] Javier Latorre, Koji Iwano, and Sadaoki Furui. "New Approach to the Polyglot Speech Generation by means of an HMM-based Speaker Adaptable Synthesizer". In: *Speech Communication* 48(10) (2006), pp. 1227–1242.

[72] Andrea G Levitt. "Reiterant Speech as a Test of Non-Native Speakers' Mastery of the Timing of French". In: *The Journal of the Acoustical Society of America* 90(6) (1991), pp. 3008–3018.

[73] Haizhou Li, Mahani Aljunied, and Boon Seong Teoh. "A Grapheme to Phoneme Converter for Standard Malay". In: *COCOSDA 2005*. Jakarta, 2005.

[74] Alvin M Liberman, Franklin S Cooper, Donald P Shankweiler, and Michael Studdert-Kennedy. "Perception of the Speech Code". In: *Psychological Review* 74(6) (1967), pp. 431 –461.

[75] Lian Tze Lim. "Using Conceptual Vectors to improve Translation Selection". MSc thesis. School of Computer Sciences, Universiti Sains Malaysia, 2006.

[76] Andrew Lovitt and Jont B. Allen. "50 Years Late: Repeating Miller-Nicely 1955". In: *INTERSPEECH 2006*. 2006, pp. 2154–2157.

[77] Andrew Lovitt, Joel Praveen Pinto, and Hynek Hermansky. *On Confusions in a Phoneme Recognizer*. Tech. rep. IDIAP, 2007.

[78] Johari Madzhi. *Fonologi Dialek Melayu Kuching (Sarawak)*. Kuala Lumpur: Dewan Bahasa dan Pustaka, 1989.

[79] Shinji Maeda. "A Characterization of American English Intonation". PhD thesis. Massachusetts Institute of Technology, 1976.

[80] Yunus Maris. *The Malay Sound System*. Siri Teks Fajar Bakti, 1979.

[81] MBROLA-Group. *The MBROLA PROJECTS: Towards a Freely Available Multilingual Speech Synthesizer*. Internet. Accesed on 17th July 2009. Mons, Belgium: Théorie des Circuits et Traitement du Signal (TCTS), 2005. URL: `http://tcts.fpms.ac.be/synthesis/mbrola.html`.

[82] Bernd T. Meyer, Matthias Wächter, Thomas Brand, and Birger Kollmeier. "Phoneme Confusions in Human and Automatic Speech Recognition". In: *INTERSPEECH*. ISCA, 2007, pp. 1485–1488.

[83] George A. Miller and Patricia E. Nicely. "An Analysis of Perceptual Confusions Among Some English Consonants". In: *The Journal of the Acoustical Society of America* 27(2) (1955), pp. 338–352.

[84] Rafizah Mohd Rasidi. "Penilaian Penulisan Buku Tatabahasa Bahasa Melayu : Satu Analisis 'Teori' Atqakum". MA thesis. Jabatan Bahasa Melayu, Akademi Pengajian Melayu, University Malaya, 2000.

[85] Carmen Foncesa Mora. "Foreign Language Acquisition and Melody Singing". In: *ELT Journal* 54(2) (2000), pp. 146–152.

[86] Ee-Lee Ng, Alvin W Yeo, and Bali Ranaivo-Malançon. "Identification of Closely Related Indigenous Languages: An Orthographic Approach". In: *International Conference on Asian Language Processing (IALP'09)*. IEEE. 2009, pp. 230–235.

[87] S.G. Nooteboom. "The Posody of Speech: Melody and Rhythm". In: *The Handbook of Phonetic Sciences*. Ed. by W.J. Hardcastle and L. Laver. Oxford: Blackwell, 1997, pp. 640 –673.

[88] George Odam. *The Sounding Symbol*. Cheltenham: Stanley Thornes, 1995.

[89] A.H. Omar. *The Iban Language of Sarawak: A Grammatical Description*. Dewan Bahasa & Pustaka, 1981.

[90] Asmah H. Omar. *Nahu Melayu Mutakhir*. Dewan Bahasa dan Pustaka, Kementerian Pendidikan Malaysia, 1993.

[91] Douglas David O'Shaughnessy. "Modelling Fundamental Frequency, and its Relationship to Syntax, Semantics, and Phonetics". PhD thesis. Massachusetts Institute of Technology, 1976.

[92] Keiichiro Oura. "An Example of Context-dependent Label Format for HMM-based Speech Synthesis in English". In: *HTS-demo CMU-ARCTIC-SLT* (2011).

[93] Gordon E. Peterson and Harold L. Barney. "Control Methods Used in a Study of the Vowels". In: *The Journal of the Acoustical Society of America* 24(2) (1952), pp. 175–184.

[94] Gordon E. Peterson and Ilse Lehiste. "Duration of syllable nuclei in English". In: *Journal of the Acoustical Society of America* 32(6) (1960), pp. 693–703.

[95] Bromeley Philip. *A Seminal Study of Iban Alphabet*. Institute of Research, Development and Commercialization, Universiti Teknologi MARA, 2007.

[96] Janet Breckenridge Pierrehumbert. "The Phonology and Phonetics of English Intonation". PhD thesis. Massachusetts Institute of Technology, 1980.

[97] Silvia Quazza, Laura Donetti, Loreta Moisa, and Pier Luigi Salza. "ACTOR: a Multilingual Unit Selection Speech Synthesis System". In: *4th ISCA Tutorial & Research Workshop on Speech Synthesis*. Perthshire, Scotland, 2001, pp. 209–214.

[98] Bali Ranaivo and Nur-Hana Samsudin. *Bahasa Malaysia Phonemes*. Tech. rep. Computer Aided Translation Unit, Universiti Sains Malaysia, 2003.

[99] Bali Ranaivo-Malançon. "Automatic Identification of Close Languages–Case Study: Malay and Indonesian". In: *ECTI Transaction on Computer and Information Technology* 2(2) (2006), pp. 126–133.

[100] Bali Ranaivo-Malançon. "Computational Analysis of Affixed Words in Malay Language". In: *Proceedings of the 8th International Symposium on Malay/Indonesian Linguistics, Penang, Malaysia*. 2004.

[101] Harald Romsdorfer and Beat Pfister. "Multi-Context Rules for Phonological Processing in Polyglot TTS Synthesis". In: *Proceeding of Interspeech 2004 and The International Conference on Speech and Language Processing 2004 (ICSLP 2004)*. Jeju Island, Korea, 2004, pp. 845–848.

[102] Harald Romsdorfer and Beat Pfister. "Phonetic Labelling and Segmentation of Mixed-Lingual Prosody Databases". In: *Proceeding of Interspeech 2005*. Lisbon, Portugal, 2005, pp. 3281–3284.

[103] Harald Romsdorfer and Beat Pfister. "Text Analysis and Language Identification for Polyglot Text-to-Speech Synthesis". In: *Speech Communication* 49(9) (2007), pp. 697–724.

[104] K.N. Ross and M. Ostendorf. "A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis". In: *IEEE Transactions on Speech and Audio Processing* 7(3) (1999), pp. 295–309.

[105] Suhaila Saee, Lay-Ki Soon, Tek-Yong Lim, Bali Ranaivo-Malançon, and Enya Kong Tang. "From Raw Text to Morphological Rules for Iban Morphological Analyser". In: *International Conference on Asian Language Processing (IALP)*. IEEE. 2012, pp. 21–24.

[106] Suhaila Saee, Alvin W Yeo, and Jennifer Wilfred. "Sarawak Language Technology (SaLT) Initiative: Preservation of Sarawak Ethnic Languages". In: *Proceedings of the Borneo Research Council 9th Biennial International Conference (BRC '08)*. Borneo Research Council. Kota Kinabalu, Sabah, Malaysia, 2008.

[107] Jenny R. Saffran, Ann Senghas, and John C. Trueswell. "The Acquisition of Language by Children". In: *Proceedings of the National Academy of Sciences of the United States of America* 98(23) (2001), pp. 12874–12875.

[108] Jan P.H. van Santen and Julia Hirschberg. "Segmental Effects On Timing And Height Of Pitch Contours". In: *The 3rd International Conference on Speech and Language Processing 1994 (ICSLP'94)*. Yokohama, Japan, 1994, pp. 719–722.

[109] Kevin P Scannell. "The Crúbadán Project: Corpus Building for Under-Resourced Languages". In: *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. Vol. 4. 2007, pp. 5–15.

[110] Tanja Schultz and Alex Waibel. "Language adaptive LVCSR through Polyphone Decision Tree Specialization". In: *Workshop on Multi-lingual Interoperability in Speech Technology (MIST-1999)*. Leusden, The Netherlands, 1999, pp. 85–90.

[111] Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek. "SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems". In: *Proceedings of Interspeech*. Antwerp, Belgium, 2007.

[112] Heather A Schunk. "The Effect of Singing paired with Signing on Receptive Vocabulary Skills of Elementary ESL Students". In: *Journal of Music Therapy* 36(2) (1999), pp. 110–124.

[113] Melody Schwantes. "The Use of Music Therapy with Children who Speak English as a Second Language: An Exploratory Study". In: *Music Therapy Perspectives* 27(2) (2009), pp. 80–87.

[114] Norman Carson Scott. *A Dictionary of Sea Dayak*. School of Oriental and African Studies, University of California, 1956, pp. 596–616.

[115] Peter G Sercombe. "Adjacent Cross-Border Iban Communities: A comparison with Reference to Language". In: *Bijdragen tot de Taal-, Land-en Volkenkunde* (1999), pp. 596–616.

[116] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. "TOBI: A Standard for Labeling English Prosody". In: *Proceeding of the 2nd International Conference on Spoken Language Processing 1992(ICSLP'92)*. Banff, Alberta, Canada, 1992, pp. 867–870.

[117] Richard Sproat. "Multilingual Text Analysis for Text-to-Speech Synthesis". In: *Journal of Natural Language Engineering*. 1996.

[118] Jochen Steigner and Marc Schröder. "Cross-Language Phonemisation In German Text-To-Speech Synthesis". In: *Proceedings of Interspeech 2007*. Antwerp, Belgium, 2007, pp. 1913–1916.

[119] Karlheinz Stöber et al. "Speech Synthesis using Multilevel Selection and Concatenation of Units from Large Speech Corpora". In: *Verbmobil: Foundations of speech-to-speech translation*. Ed. by Wolfgang Wahlster. Springer, 2000, pp. 519–534.

[120] Panceras Talita, Alvin W Yeo, and Narayanan Kulathuramaiyer. "Challenges in Building Domain Ontology for Minority Languages". In: *International Conference on Computer Applications and Industrial Electronics (ICCAIE)*. IEEE. 2010, pp. 574–578.

[121] Tien-Ping Tan. "Automatic Speech Recognition for Non-Native Speakers". PhD thesis. Université Joseph-Fourier - Grenoble I, 2008.

[122] Tien-Ping Tan, Xiong Xiao, Enya Kong Tang, Eng Siong Chng, and Haizhou Li. "MASS: A Malay language LVCSR corpus resource". In: *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*. IEEE. 2009, pp. 25–30.

[123] Lim Beng Tat, Zaharin Yusoff, Tang Enya Kong, and Guo Cheng Ming. "Primitive-based Word Sense Disambiguation for SENSEVAL-2". In: *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics. 2001, pp. 103–106.

[124] Paul Taylor, Alan W Black, and Richard Caley. "The Architecture of the Festival Speech Synthesis System". In: *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*. International Speech Communication Association, 1998, pp. 147 –151.

[125] Keiichi Tokuda, Heigen Zen, and Alan W. Black. "An HMM-Based Speech Synthesis Applied to English". In: *Proceeding of IEEE Workshop on Speech Synthesis 2002*. 2002, pp. 227–230.

[126] Keiichi Tokuda, Yoshihiko Nankaku, Takechi Toda, Heishun Zen, Junichi Yamagishi, and Keiichiro Oura. "Speech Synthesis based on Hidden Markov Models". In: *Proceedings of the IEEE* 101(5) (2013), pp. 1234–1252.

[127] Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner. "From Multilingual to Polyglot Speech Synthesis". In: *Proceedings of Eurospeech 1999*. Budapest, 1999, pp. 835–838.

[128] N Umeda. "Another Consistency in Phoneme Duration". In: *The Journal of the Acoustical Society of America* 58(S1) (1975), S62–S62.

[129] Joseph Verguin. "L'accentuation en Malgache-merina et en Malais". In: *Orbis* 4(2) (1955), pp. 522–528.

[130] Wolfgang Wahlster. "Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System". In: *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000, pp. 3–21.

[131] R.O. Winstedt. *Malay Grammar*. Clarendon Press, 1927.

[132] Soo-Fong Yong, Bali Ranaivo-Malançon, and Alvin Yeo Wee. "NERSIL:The Named-Entity Recognition System for Iban Language". In: *PACLIC*. 2011, pp. 549–558.

[133] Rohani Mohd. Yusof. "Perkaitan bahasa Melayu dan bahasa Iban: Satu Tinjauan Ringkas". In: *Jurnal Bahasa* 3(3) (2003), pp. 457–475.

[134] Heiga Zen. "Acoustic Modeling in Statistical Parametric Speech Aynthesis – from HMM to LSTM-RNN". In: *Proceedings of Machine Learning in Spoken Language Processing (MLSLP)*. International Speech Communication Association, 2015.

Glossary

isolect a language or dialect; “coined” as a neutral term between ‘language’ and ‘dialect’.. 185

lexicostatistic the statistic /quantitative assessment of the genealogical relatedness of language. 26, 132, 185

logatome non-sense utterance; a meaningless artificial word that obeys all the phonotactic rules of a language. In the literature of this thesis, logatome words being used in the studies were also reiterant utterance, ‘mamama’ or ‘bababa’ produced using different prosody to identify the words or sentence focalisation. 185

morphology the mental system involved in word formation or to the branch of linguistics that deals with words, their internal structure, and how they are formed. The form includes the language morphemes and other linguistic units such as words, affixes, parts of speech and intonation/stress. . 185

orthographic phonography a writing system in which the words of a language are spelled representing elements of sound the word form - in phonetic sound form. 185

pentaphone two left and two right context of the phoneme in questions. 185

phonetic spelling system see orthographic phonography. 185

phonographic a writing system which the spelling is based on pronunciation. It may also refer to a system of shorthand writing based on sound. 185

polyphone a letter (or combination of letters) that has two or more pronunciations. Example: <c>is a polyphone. It can be pronounced like /k/ in car and /c/ or /tʃ/ in charcoal and /s/ in cell. 185

suprasegmental A suprasegmental is a vocal effect that extends over more than one sound segment in an utterance, such as pitch, stress, or juncture pattern. Suprasegmental is often used for: tone, vowel length, and features like nasalization and aspiration. 3, 185