A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System

Tien-Ping Tan[1], Sang-Seong Goh[2], Yen-Min Khaw[1]

[1]School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
tienping@cs.usm.my, jasminakhaw87@hotmail.com

[2]School of Humanities
Universiti Sains Malaysia
Penang, Malaysia
gohss@usm.my

***Abstract*—Malay is a language from the Austronesian family. Malay is the official language in Malaysia, Indonesia, Singapore, and Brunei. However, Malay spoken in different countries, and even within a country itself might vary in terms of pronunciation and vocabulary from one place to another. The Malay dialects in Malaysia can be grouped according to the states of the country. In this paper, we propose the architecture of a Malay dialect translation and synthesis system, that given a sentence in standard Malay, it translates and synthesizes an utterance in the dialect requested. The system consists of 3 modules, dialect translation system, dialect G2P system, and speech synthesis system. The outcome from this study is two folds. From linguistic viewpoint, it will help us understanding and appreciating the interesting differences in the Malay dialect in Malaysia, which is important to help preserve the dialect and culture in it. Secondly, the proposed system will be useful for people who like to learn a particular dialect or it can be used in places that require this facility. At this stage, we have completed the standard Malay system, and this paper presents our work so far.**

Keywords—Malay; dialect; speech synthesis

I. INTRODUCTION

Malay is a language from the Austronesian family [1]. It is one of the most widely spoken languages in the world. Malay is the official language in Malaysia, Indonesia, Singapore, and Brunei. However, Malay spoken in different countries and even within a country itself might vary in terms of pronunciation and vocabulary from one place to another. In Malaysia, the Malay dialect can be grouped according to the geographical distribution in Malaysia [2], and maybe further classified according to different areas. For example, Malay dialects spoken in Perak (northern state of Malaysia) can be classified according to five areas. The north part of Perak speak Petani and Kedah dialect, the south part speak Selangor dialect, slightly to the east part speak Rawa dialect, while area around the middle of Perak around Parit and Kuala Kangsar speak Perak dialect [3].

Even though in Malaysia, there is standard Malay, Malay dialects still flourished and widely used in many areas especially for unofficial matters. The reason is, speakers with the same dialect share the same origin, culture and social group. Often to get acceptance into the group, speaker must speak the same dialect. However, Malay dialects are distinctive, and might not easy to learn. They might not only differ in pronunciation, but they can also vary in term of vocabulary and maybe grammar. Therefore, a system that can translate a standard Malay sentence to another Malay dialect and synthesize it can be a useful tool for users to learn different Malay dialects. Furthermore, the system also has vast potential to be applied in many areas that require speech generation or synthesis application.

II. BACKGROUND

There is not much study on dialect translation and synthesis system. However, pronunciation modeling, translation and speech processing approaches are widely studied. The following section describes related works in these areas.

A. Pronunciation Modeling

In many languages such as English, French, German, and Mandarin, there is a documented way of how words are pronounced. The pronunciation of words can normally be found in a dictionary. The pronunciations are typically described using IPA (International Phonetic Alphabets) symbols. Other language such as Mandarin uses a different way for describing the pronunciation of Chinese characters. In Malay, some studies have been done to describe the standard Malay pronunciation [4][5]. However, works on Malay dialects are limited.

For analyzing the phonology of a language or dialect, perception test, acoustic phonetic analysis, and speech processing techniques can be used. Perception test is easy to be carried out. Perception test requires native listeners to listen to some sample of sounds, which differs only in a speech sound [6]. If the listener can distinguish the speech sound, then the speech sound is a phoneme of the language. On the other hand, acoustic phonetic analyze the acoustic features of the spoken signal using spectrogram [7][8]. However, acoustic analysis might not be easy for certain sounds. These approaches require expert knowledge to be carried out. In situations when experts are not available, speech-processing tools such as phoneme recognizer and automatic speech recognizer [9] can be useful to derive the phoneme set.

Malay dialects can differ in term of phoneme set. For example, Perak dialect contains the phoneme /ɛ/ and /ɔ/ [3], but not the standard Malay. The pronunciation of Malay words may also differ among dialect, but often the differences are systematic. For example, a Malay dialect in

978-0-7695-4886-9/12 $26.00 © 2012 IEEE
DOI 10.1109/IALP.2012.14

Negeri Sembilan (a southern state of Malaysia), words with prefix "ber" is read as /b o r/, but in standard Malay it is read as /b ə r/ [10]. Another example is, the final grapheme "a" is pronounced as /ɛ/ in Perak dialect [3], but it is read as /ə/ in standard Malay. Besides differences in pronunciation, the vocabulary might also be different. For example in Malay dialect from Kedah, the word for "awak" (English: you) is "hang", "air" or water is "ayak" and "sen" is called "kupang". These words may not exist in written form. In term of grammar, they are similar.

B. Machine Translation

In term of machine translation, there are many approaches for translating a text from one language to another. Current state of the art approaches are statistical machine translation and example based translation. Statistical machine translation is a class of approaches that make use of a combination of probabilistic models to choose the most probable translation, for a sequence of words in the source language, given the target language [11][12]. On the other hand, example based translation system make use of linguistic knowledge and the main idea behind example based machine translation is translation by analogy [13][14]. However, there are also approaches that include statistical knowledge into example-based machine translation to make the translation more accurate.

C. Speech Synthesis System

With the linguistic knowledge, speech synthesis system will make use of the information for generating speech given a text. A speech synthesis system is a system that converts text to speech. There are many different approaches to speech synthesis: articulatory synthesis, formant synthesis, concatenative synthesis and HMM synthesis [9][15][16][17]. Typical speech synthesis architecture consists of a speech generation module and a training module. The speech generation module consists of a text analyzer, linguistic analyzer and a speech generation module, while the training module makes use of the pronunciation dictionary and an acoustic model. The text analyzer will analyze a sentence, and convert it to pronunciation using a pronunciation dictionary and/or a grapheme to speech sound system [18]. As for the acoustic modeling of the speech sounds, to obtain the models, speech corpus have to be acquired to train them using data driven approach. A waveform generator will then convert the parameters to speech. The type of speech sound for example, word, syllable, phone and others to create is depend on the speech synthesis architecture, requirements and the speech corpus acquired.

III. Proposed Malay Dialect Translation and Synthesis Speech System

In this project, we propose a Malay dialect translation and synthesis system. Given a sentence in standard Malay, the system will convert the sentence to an equivalent Malay dialect, and then generate or synthesize the corresponding speech. The proposed system consists of the following modules: Malay Dialect Translation Module, Malay Dialect Grapheme to Phoneme Module and Malay Dialect Synthesis Module. See Figure 1.

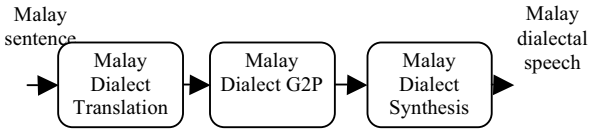

Figure 1. Malay Dialect Translation and Synthesis System

A. Dialect translation system

Since the amount of dialectal resources for Malay is near to none, statistical machine translation is not a good choice. Thus, Example Based Machine Translation (EBMT) approach will be a more reasonable solution. Furthermore, most of the Malay dialects are similar in term of grammatical structure; EBMT is a more reasonable solution.

B. Malay dialect G2P system

Figure 2 below shows our proposed Malay dialect G2P architecture. It is a rule-based system.

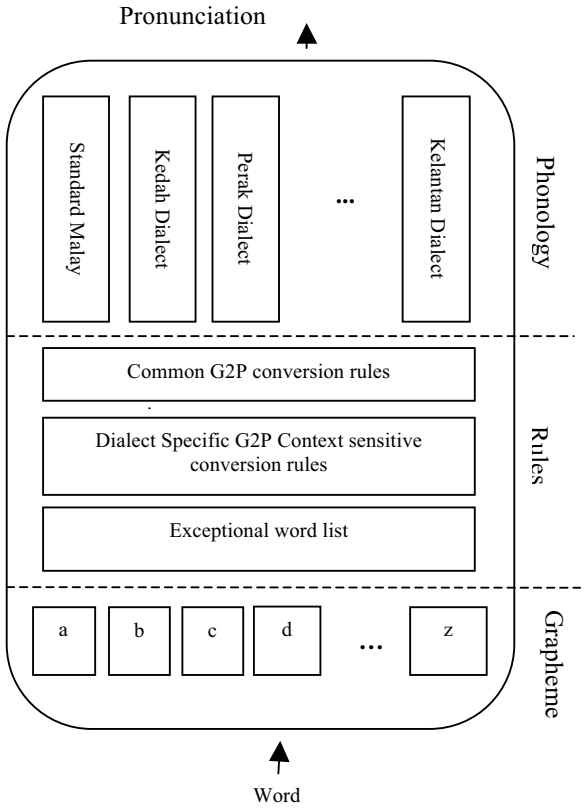

Figure 2. Rule-based Malay Dialect G2P System

At the bottom layer, we define the Malay graphemes. At the middle layer, we have the graphemes to phonemes conversion rules. The rules can be divided into 6 groups: context free rules that convert graphemes to phonemes without considering the context, context sensitive rules, context sensitive rules that require syllable information, context sensitive rules that require affixation information, context sensitive rules that require both syllable and affixation information, and also the dialect specific rules. There is also an exceptional word list, which includes

words that is not pronounced according to the rules. At the top layer, we have different Malay dialects that make use of different rules to generate pronunciation of a word.

C. *Speech Synthesis System*

For speech synthesis system, we propose to apply HMM speech synthesis system for converting speech to text. The reasons for using HMM speech synthesis system is because it requires less resource, and the HMM acoustic model created can be applied different transformations. In our case, we want to adapt a speech dialect to a different dialect.

After recording the corpus, using the acoustic model created from our Malay speech corpus (MASS) and also the dialect pronunciation dictionary, we will force align the speaker dependent dialectal speech an automatic speech recognizer. The purpose is to align the phones in the utterances and then train a speaker dependent Hidden Markov Model (HMM) acoustic model, which is phone based. Given a sentence, the sentence will need to be converted to pronunciations before the vocoder can synthesize the corresponding speech.

IV. PRELIMINARY SYSTEM

At this moment, we have completed the standard Malay module for G2P and speech synthesis, and in the process of studying and adding other Malay dialects.

A. *Standard Malay G2P Module*

The pronunciations of standard Malay words in phoneme can be determined from the grapheme, word morphology and the syllable structures [5]. A grapheme is the fundamental unit of written language, and a phoneme is "the smallest linguistically distinctive unit of sound".

There are 34 graphemes in Malay. It consists of all the 26 alphabets in English with the additional 8 graphemes with 2 letters ('ai', 'au', 'oi', 'kh', 'gh', 'sy', 'ng', 'ny'). Given a word, the G2P system will first segment the word to graphemes. Before the pronunciation of the word can be predicted, the affixation information of the words and syllable structure of the words need to be obtained. After that, the G2P will convert the graphemes to their respective sounds. We use a rule based G2P to convert a Malay word to their respective pronunciation. The reason for using a rule-based system instead of another approach e.g. a Hidden Markov Model is because rules-based system is generally more accurate and simple to be implemented, especially when the number of rules involved is not many. In our case, there are seven rules for converting words to pronunciation, namely general replacement rule, schwa rule, glottal stop insertion rule, general glottal stop rule, final "r" deletion rule, glide insertion rule, last syllable rule, and duplicate grapheme rule. There are 36 phonemes in standard Malay [4]. Six of them are vowels, three are diphthongs, and 27 are consonants. However, two of the phonemes /**θ**/ and /ð/ that are rarely used by the speakers are replaced with /t/ and /d/ respectively.

1) General replacement rule

This is the context dependent rule for mapping grapheme to phoneme. The word *diberi* (English: given) is converted to /d i b ə r i/.

TABLE 1: Context independent mapping of grapheme to phoneme

Grapheme	Phoneme	Grapheme	Phoneme
p	**p**	**j**	**ʤ**
b	**b**	**l**	**l**
t	**t**	**r**	**r**
d	**d**	**m**	**m**
k	**k**	**n**	**n**
q	**k**	**ng**	**ŋ**
g	**g**	**ny**	**ɲ**
s	**s**	**w**	**w**
x	**s**	**y**	**j**
h	**h**	**a**	**a**
f	**f**	**e**	**ə**
v	**v**	**i**	**i**
z	**z**	**o**	**o**
sy	**ʃ**	**u**	**u**
sh	**ʃ**	**ai**	**aj**
kh	**x**	**au**	**aw**
gh	**ɤ**	**oi**	**oj**
c	**ʧ**		

2) Schwa rule

The grapheme 'a' at the end of a word is pronounced as /ə/. This rule is applicable only for 'old' Malay words, and it's not applicable for proper nouns and borrowed words. For example the word *suka* (English: like) is pronounced as /s u k ə/. If the root word is appended with a suffix, the 'a' at the end of the root word is still pronounced as /ə/. For example, the word *sukakan*, which is formed by adding the suffix *-kan* to the root word *suka* is pronounced as /s u k ə k a n/.

3) Glottal stop insertion rule

A glottal stop /**?**/ is inserted between two specific sequences of vowel graphemes in a word. See Table 2.

TABLE 2: Glottal stop insertion rules

Grapheme sequence	Word	Pronunciation
aa	taat	/**t a ? a t**/
oa	doa	/**d o ? a**/

4) General glottal stop rule

The grapheme 'k' at the end of a syllable is converted to a glottal stop /**?**/. For example, the word *laksa* (English: noodle) is pronounced as /l a ? s a/.

5) Final 'r' deletion rule

The grapheme 'r' at the end of a word is not pronounced. However, there are some speakers that pronounce this final 'r'. For example the word *sukar* (English: difficult) is pronounced as /s u k a/.

6) Last syllable rule

For words with more than one syllable, the grapheme 'u' and 'i' at the last syllable of some contexts are converted to phoneme /o/ and /e/ respectively. Like schwa rule, if the root word is appended with a suffix, the rule is still applicable on the last syllable of the root word. This rule is also not applicable for proper nouns and borrowed words. See Table 3.

TABLE 3. Last syllable rules

Target grapheme	Following grapheme	Word	Pronunciation
u	k, h, p, m, ng, r	hidup	/**h i d o p**/
i	k, l, t, h, r, t, k	bilik	/**b i l e ?**/

7) Duplicate grapheme rule.

A sequence of two similar graphemes is converted to a single phoneme. This rule is used mostly for proper nouns. For example, *Azzam* is pronounced as /a z a m/.

There are some words where the rules described above do not applied. First, many of the words with grapheme 'e' should be transcribed to either /e/. Thus, the general mapping rule which map 'e' to /ə/ is not correct. In this case, human knowledge is required to determine the right mapping. Secondly, schwa rule does not apply to "new" Malay words and proper noun. For these words, speaker pronounce the final 'a' as /a/. Thirdly, English words are also often found in Malay text and speech. For these words, we simply convert English phoneme to the nearest Malay phoneme in perception. For these reasons, manual validation of the pronunciation needs to be done. Currently, we have validated more than 60 thousand words.

B. Standard Malay Speech Synthesis Module

The speech synthesis system used here is a HMM speech synthesis system (HTS Speech Synthesis System). The recording was done in a sound proof room, using AKG C414XLII microphone, and EMACOP software [19]. The sampling rate is set at 22kHz. The speech of two speakers, one native male and another native female, was recorded. Nearly 5 hours of speech was recorded from each speaker. After the speech corpus was acquired, we aligned the phone in the utterances to create speaker dependent acoustic models. Since manual alignment of the utterances is expensive and time consuming, we applied automatic alignment by force aligning the utterances using an automatic speech recognizer (Sphinx3 from CMU). The acoustic model of the automatic speech recognizer was trained using MASS speech corpus [20] that contains about 140 hours of speech, and our pronunciation dictionary. The aligned speech is then used to train acoustic model for the HTS speech synthesis system. From our study, we notice that, if we create the HTS speech synthesis acoustic model from the speaker's speech directly, the synthesized speech is not clear. This observation is similar to the finding in the field of automatic speech recognition, where an acoustic model created from a large speech corpus is better in decoding the speech of a particular speaker than using a small speaker dependent speech corpus alone.

The perception test carried out on 20 speakers to evaluate the quality of the standard Malay speech synthesis system compared to the natural speech of the speakers shows that, from the value of 0 (bad) to 10 (very good), the average rating given for the quality of the synthesize speech is "good".

V. CONCLUSIONS

In this paper, we present our proposal for a Malay dialect translation and synthesis system. The preliminary system that we have constructed so far for standard Malay is very promising. On average, the speakers rate the standard Malay speech synthesis system to be "good". Nevertheless, there is still a lot of work needed to complete the whole system. Currently, we are in the process of studying and analyzing Malay dialects for Kedah and Kelantan to be added to the current system

VI. ACKNOWLEDGEMENT

This project is supported by the ERGS grant 203/PHUMANITI/6730035 from Ministry of Higher Education, Malaysia.

REFERENCES

[1] W. O'Grady, and J. Archibald, *Contemporary Linguistic Analysis: An Introduction*, Addison Wesley Longman, Toronto, 2000.

[2] J. T. Colins, "Malay Dialect Research in Malaysia: the Issue of Perspective", Bijdragen tot de Taal-, Land- en Volkenkunde, 1989, pp 235-264.

[3] Z. B. Ahmad, *The Phonology & Morphology of The Perak Dialect*, Dewan Bahasa dan Pustaka, 1991.

[4] Y. M. Maris, *The Malay Sound System*. Siri Teks Fajar Bakti, Malaysia, 1979.

[5] T-P. Tan, B. Ranaivo-Malançon, "Malay Grapheme-to-Phoneme Tool for Automatic Speech Recognition". *Malindo'09*, 2009

[6] P. Ladefoged, *Vowels and Consonants: An Introduction to the Sound of Languages*, Black Well Publishing, United Kingdom, 2000.

[7] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, Singular Thomson Learning, Canada, 2002.

[8] K. N. Sevens, *Acoustic Phonetics*, The MIT Press, Cambridge, Massachusetts, 2000.

[9] X.D. Huang, A. Acero, H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 2001.

[10] R. S. Hendon, The phonology and Morphology of Ulu Muar Malay, Yale University Publications, 1966.

[11] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, R. L. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics, v.19 n.2, 1993.

[12] D. Jurafsky, J. H. Martin, Speech and Language Processing, 2nd Ed, Pearson Education, 2009, pp. 910-942.

[13] H. A.-A. Mosleh, E. K. Tang, "Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema", Machine Translation Summit VII, 1999, pp 244-249.

[14] Mosleh Al-Adhaileh, Tang E.K Zaharin Yusoff, "A Synchronization Structure of SSTC and Its Applications In Machine Translation", COLING-02 on Machine translation in Asia, Volume 16, 2002.

[15] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addision-Wesley, 1999.

[16] E. Rank and H. Pirker, "Generating Emotional Speech with a Concatenative Synthesizer", ICSLP'98, 1998, pp. 671-674.

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", Eurospeech, 1999, pp.2347-2350.

[18] D. H. Klatt, "Software for a Cascade/ Parallel Formant Synthesizer", Journal of Acoustical Society of America, vol 67, 1980, pp. 971-995.

[19] D. Vaufreydaz, J. Bergamini, J. F. Serignat, L. Besacier, and M. Akbar, "A New Methodology for Speech Corpora Definition from Internet Documents," presented at LREC'00, Athens, Greece, 2000.

[20] T-P. Tan, HZ. Li, E. K. Tang, X. Xiao, E. S. Chng, "Mass: A Malay Language LVCSR Corpus Resource", Cocosda'09, Beijing, 2009, pp. 10-13.