Preparation of MaDiTS Corpus for Malay Dialect Translation and Speech Synthesis System

Yen-Min Jasmina Khaw, Tien-Ping Tan

School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia.
jasminakhaw87@hotmail.com,tienping@cs.usm.my

Abstract

This paper presents our work in acquiring a Malay dialect translation and speech synthesis corpus. In this study, an architecture of speech corpus acquisition, which including Malay dialect translation and Malay dialect grapheme to phoneme (G2P), was proposed. The pronunciation dictionary for dialectal Malay was generated through G2P tool. As dialectal Malay is considered as scarce resource, dialectal translation rules were developed for translating standard Malay text into dialectal Malay. With this, Kelantanese Malay is chosen in this research as it is considered as one of the Malay dialect from Kelantan, which positioned in the northeast of Peninsular Malaysia. This dialect is very distinctive. Evaluation results showed that the selected sentences through proposed approach has a correlation coefficient of about 0.99, which mean that it is phonetically well balanced.

Index Terms: Malay dialect translation, Malay dialect grapheme to phoneme, speech synthesis corpus

1. Introduction

Malay is a language from the Austronesian family [1], one of the most widely spoken languages in the world. It is the official language in Malaysia, Indonesia, Singapore, and Brunei. There are several Malay dialects spoken in Malaysia such as Perak dialect and Kedah dialect. Malay dialect is very distinctive. They might not only differ in pronunciation, but they can also vary in term of vocabulary and maybe grammar. For example, the word "*saya"* /s a̯ j ə/ in standard Malay (English: I) will be "*kawe"* /k a w ɛ/ in Kelantan dialect. In term of grammatical structure, standard Malay and dialectal Malay have similar sentence structure. However, there are few differences in some special cases. One of the rules appears in Kelantan dialect is that the word "*dah"* /d ɔ h/ (standard Malay: sudah; English: already) appears after verb in a sentence. For example, 'Dia sudah tidur(v).' (English: He slept already.) in standard Malay became 'Dia tidur(v) dah.' in Kelantan dialect.

Speech corpus is an important resource for research and development of speech processing applications such as speech synthesis, automatic speech recognition, speaker recognition system, speech translation system and others. There are efforts in acquiring standard Malay speech corpus especially in the area of automatic speech recognition [2]. However, the work in acquiring dialectal Malay speech is limited. Collecting dialectal Malay speech is a challenging task. The reason is dialectal Malay speech is a scarce resource. Secondly, in term of orthography, there is no standard writing system in dialectal Malay. This paper explains our work in collecting standard Malay and dialectal Malay speech. The purpose of the corpus is for research in domain of speech synthesis and linguistic study. We envisage to use the corpus for dialect translation and speech synthesis and to study the possibility of transforming one dialect to another.

In this study, we focus on standard Malay and Kelantanese Malay. Kelantan dialect is a dialect spoken by residents staying in the state of Kelantan, which is a state in peninsular Malaysia. This dialect is not only spoken in the state of Kelantan, but across political boundaries [3]. It is very distinctive and might be difficult to learn as every single Kelantan words can bring various meaning. Most of the Kelantanese learned Kelantan dialect at home.

This paper is organized as follows. In section 2, background for some research in this study is reviewed. Speech corpus acquisition is described in section 3 while section 4 illustrated speech synthesis corpus recording. Building speech systhesis system is drawn in section 5. In section 6, it contains conclusion and future work.

2. Background

There are some previous studies on collecting corpus of unwritten language. For example, Kammu is one of an unwritten language from Loas which has no practical orthography. This language was collected in term of recording with its transcriptions are in principle phonemic, using IPA, except for common use of capitalization and punctuation [4]. Most publications on linguistic field methods emphasise that a collection of recorded, transcribed and analyzed texts is the most important source for the grammatical description of a previously unresearched language [5]. Besides, there is also an oral translation done in previous study for unwritten language [6]. In this study, we are standardizing a writing system for dialectal Malay and their grapheme to phoneme rules are determined.

To translate a text from one language to another, statistical machine translation and example-based translation approaches can be used. 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 [7] [8]. For example-based translation system, it makes use of linguistic knowledge. However, the need of developing ruled-based machine translation is raised for language with morphologically rich and less-resourced [8].

The phonology of a language or dialect can be analyzed through perception test, acoustic phonetic analysis or speech processing techniques. Malay dialects can differ in term of phoneme set. However, the different pronunciations of Malay among dialect are often systematic. There is a standard Malay G2P tool proposed in previous research by applying G2P conversion rules. The pronunciation of the words is then generated. For example, the word 'nasi' (English: rice) is given pronunciation of /n a s i/ where graphemes <n>, <a>, <s> and <i> are mapped to corresponding phonemes of /n/, /a/, /s/

and /i/ in a general replacement rule developed in previous study of standard Malay G2P tool [9].

A speech synthesis system is a system that converts text to speech. There are some different approaches for speech synthesis such as articulatory synthesis, formant synthesis, concatenative synthesis and HMM synthesis [10][11][12][13]. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. In previous research, standard Malay speech synthesis system has been developed [14]. It will be extended by applying dialectal Malay translated from standard Malay corpus into speech synthesis system which will be useful for people who like to learn a particular dialect, or it can be used in places that require this facility.

3. Speech Corpus Acquisition

There are some requirements that need to be fulfilled when acquiring a speech synthesis corpus. In term of environment, the recording must be done in a noise free studio. For recording, there are some criteria to be met such as expressiveness control, easy to segment, speaking rate control, prosody structure control, and voice beauty [15]. A speaker that can meet the five criteria mentioned above, then a good quality of synthesis system can be produced. In term of the content of the speech, the speech corpus should cover as many speech contexts as possible. A considerable amount of speech recordings with carefully selected sentences is very important for developing a good quality of speech synthesis system. One possible source of dialectal speech is from dialog speech. However, dialog speech is less suitable to be used as speech synthesis corpus because of the speed of the discourse and also the richness in emotion in an uncontrolled recording might not be desirable. Moreover, dialog speech normally does not cover a lot of phonetic context compared to read speech. Therefore, read speech is used instead of dialog speech. For preparing read speech, sentences must be selected from text corpus. Although dialectal text is available, the writing is not standardized and the quantity is very small.

In this paper, we present our work in acquiring a Malay dialect translation and speech synthesis corpus. Figure 1 shows the overview of the steps taken.

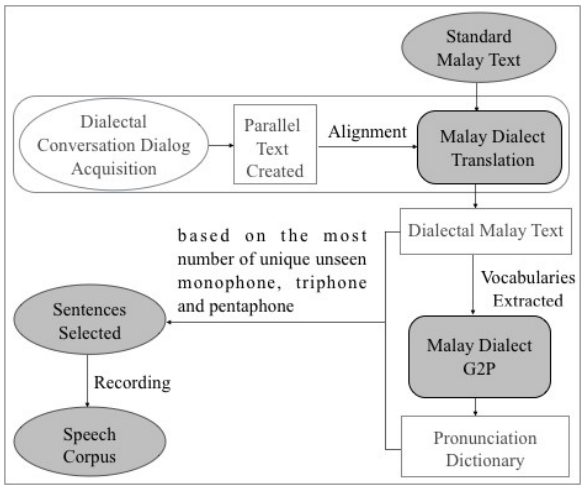

Figure 1: *Overview on speech corpus acquisition.*

The first step is to acquire some dialog speech. The dialog speech is manually transcribed and a parallel corpus is prepared to produce translation rules and unique dialectal vocabularies. With the learned translation rules and vocabularies, standard Malay text corpus is translated into dialectal Malay since we have a large standard Malay text corpus [16]. Then, G2P tool developed in this study is used to convert dialectal words to their corresponding pronunciation. To achieve phonetically balanced in preparing sentences for recording, which to be useful in creating dialectal speech synthesis system, the dialectal sentences were selected from the translated corpus based on the information of monophones, triphones and pentaphones. Finally, recording for the selected sentences was carried out.

3.1. Dialectal Conversation Dialog Acquisition

To obtain translation rules and vocabularies of dialectal Malay, a parallel corpus is required. Therefore, a dialectal dialog corpus was first recorded to obtain a dialectal text corpus and to be analyzed, before translating it manually into standard Malay. Recordings were conducted in a noise free room where two speakers who were discussing a topic are separated. They spoke through a microphone and telephone. They discussed on particular topics of interest for example Kelantan privileges, social problems and marriage issue. The more topics were discussed, the more dialectal Malay were covered. The recording was recorded in 22kHz using AKG C414XLII microphone and CoolEdit software. The conversation was then transcribed based on the graphemes used in standard Malay with correspondence phonemes. For example, /anɔʔ demɔ sakiʔ/ is transcribed as "*anak dema sakit*" (English: My child is sick.). The transcription process is standardized in term of spelling. Some of the examples of phonemes and graphemes correspondence for Kelantanese Malay are showed in Table 1. The transcribers will refer to the phoneme and grapheme correspondence table while transcribing the conversations.

Table 1. *Examples of phoneme and grapheme correspondence.*

Phoneme	Grapheme	Phoneme	Grapheme
/a/	a	/e/	e
/n/	n	/m/	m
/ɔ/	a	/s/	s
/ʔ/	k/t	/k/	k
/d/	d	/i/	i

About 5 hours of Kelantanese Malay conversation were recorded. There were 10 speakers involved in the conversation recording. Two participants carried out a conversation, which took 10 minutes for each different topic. Each conversation contains around 200 sentences.

3.2. Malay Dialect Translation Rules

Since there is no written form for dialectal Malay, grapheme used in dialectal Malay is based on standard Malay grapheme. So, there are no much changes in spelling for the dialectal words compare to standard Malay. In this study, Levenstein distance and statistical approach were used to align standard Malay and Kelantanese Malay words in parallel sentences without using any dictionary. While aligning the parallel text, translation rules and biligual dictionary were built. The dialectal Malay word with the shortest distance is aligned to the target standard Malay word. However, more than one word in dialectal Malay might be aligned to the same word in standard Malay because they have nearly same distance. Therefore, the dialectal Malay words that are aligned to the same standard Malay word will be compared again and get aligned. For words in dialectal Malay sentences which do not aligned to any standard Malay word will be aligned to the left-

word by default. Each vocabulary might appear several time in several different parallel sentences. The probability of each standard Malay word that it matches to Kelantanese Malay words is counted. In step 5 of Figure 2, the word "*saya"* (English: I) is matched to several possible words such as "*kawe"* /k a w ɛ/ (English: I), "*dema"* /d e m ɔ/ (English: I), "*sera"* /s ə ɣ ɔ/ (English: feel) and "*sayu"* /s a j u/ (English: plaintive). The word pair with the highest probability is used as reference of alignment, where bilingual dictionary was built. Example is shown in step 6 where the word "*saya"* (English: I) is matched to "*kawe" /k a w ɛ/*, "*isteri"* (English: wife) is matched to "*bini"* /b i n i/ and "*dulu"* (English: ago) is matched to "*mula" /m u l ɔ/*. By referring to the bilingual dictionary built, the aligned word is refined. Figure 2 shows the proposed approach for aligning parallel sentences for languages without a written form using standard Malay and Malay dialects. The sentence used in the example is 'Saya bawa nasi.' (English: I brought rice.).

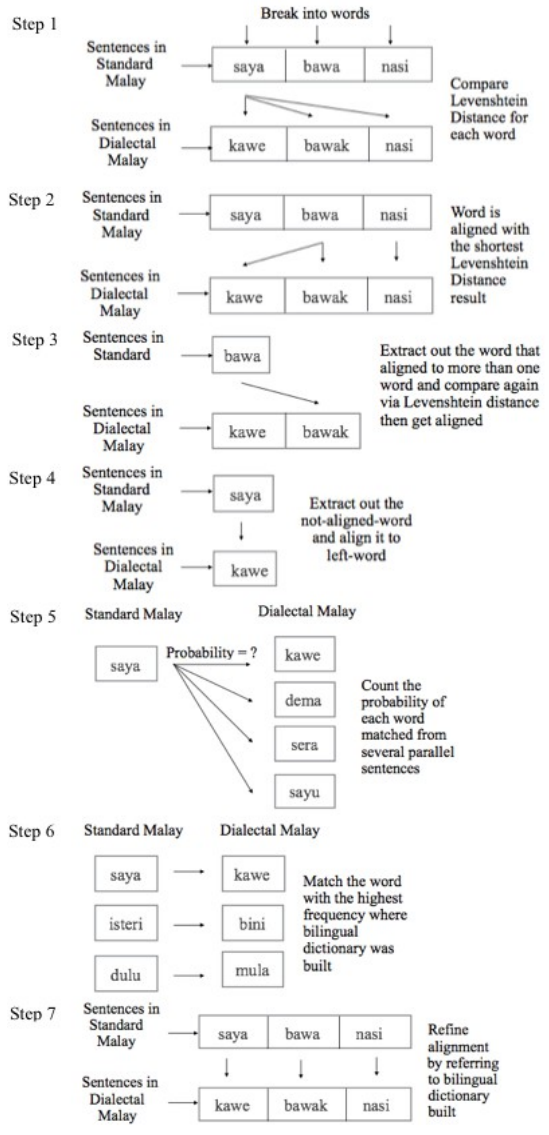

Figure 2: *Proposed approach for parallel text alignment*.

Besides, from the alignment of standard Malay-Kelantanese Malay parallel text conducted, we found some unique sentence structure in Kelantanese Malay as below.

1) **Swap Past Perfect Word of 'Sudah' Rule**: The word "*sudah"* (English: already) appears after a word of the verb in a Kelantanese sentence. In Kelantanese Malay, the word "*sudah"* is replaced with "*dah"*. For example, 'Dia sudah makan(v).' in standard Malay becomes 'Dia makan(v) dah.'.

2) **Swap Intensifier Rule:** The words "*sangat"* (English: very), "*sungguh"* (English: very) and "*benar"* (English: really) are the intensifier which appears after a word of adjective in a Kelantanese sentence. For example, 'Dia sangat letih(adj).' in standard Malay becomes 'Dia letih(adj) sangat.'.

3) **Suffix 'kan' Replacement Rule:** The suffix /-kan/ is replaced with the prefix /pə-/ for root word starting with consonant except 'h' or prefix /pəɣ-/ for a root word starting with vowels and consonant 'h', 'h' is dropped. For example, the word "*tidurkan"* in standard Malay (English: snooze) is pronounced as /p ə t i d o/ in Kelantanese Malay.

4) **Double Consonant Rule:**

a) ***Words made up of three syllables***: The first syllable is dropped and replaced by raising the length of the first consonant in the second syllable of the word. The dropped syllable could be a prefix or phonological features of a word. For example, the word "*bersusah"* (English: burden) is pronounced as /ss u s ɔ h/ where *ber-* is a prefix.

b) ***Words that having double quotation marks***: The first element of the word is aborted and at the same time the initial consonant in the second element of the first syllable is doubled. For example, the word "*kura-kura"* (English: tortoise) is pronounced as /kk u ɣ ɔ/.

c) ***Words with preposition in front****:* The preposition is deleted and the first consonant of the word is elongated. It involved only preposition of *di* (English: in), *pada* (English: at) and *ke* (English: to). For example, '*ke sini'* (English: come here) is pronounced as /ss i n i/.

The proposed alignment algorithm was evaluated by calculating precision and recall. Around 400 sentences are randomly chosen from the parallel text for evaluation. The average precision and recall achieves 0.9147 and 0.9172, which are above 90%. Through alignment conducted, around 300 standard Malay words were aligned to the new vocabularies in Kelantanese Malay. For example, the word *saya* in standard Malay (English: I) was aligned to the word *dema* with pronunciation of /d e m ɔ/ in Kelantanese Malay. With the vocabularies and translation rules found, dialectal Malay text was produced. Words in translated dialectal text were then extracted to convert into pronunciation. A Malay dialect grapheme to phoneme (G2P) was proposed in this study to convert dialectal words to their pronunciation, that is illustrated in section 3.3.

3.3. Malay Dialect Grapheme to Phoneme

Architecture for Malay dialect grapheme to phoneme (G2P) was proposed to estimate the pronunciation of a dialectal word. In this architecture, it is flexible in adding and removing rules. This Malay dialect G2P system is a rule-based tool. At the bottom layer of the proposed architecture of rule-based G2P system, the Malay graphemes are defined. At the middle layer, there are 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 are not pronounced according to the rules. At the top layer, there is Malay dialect that makes use of the G2P conversion rules to generate pronunciation of a word. Figure 3 shows the architecture of Malay dialect G2P system.

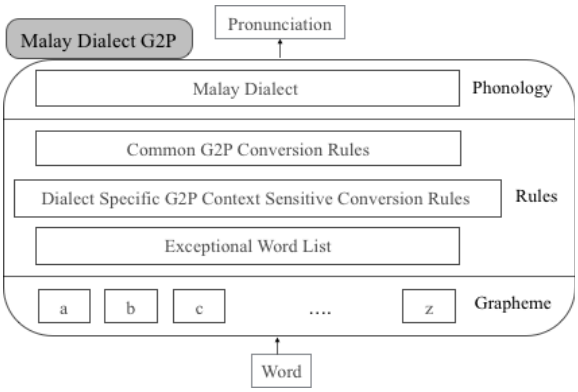

Figure 3: *Rule-based G2P system for Malay dialect.*

From the acoustic analysis of some recording done, the number and the type of consonants and vowels used in Kelantanese Malay were determined. Therefore, a Kelantanese Malay G2P tool was developed. Fifteen of grapheme to phoneme conversion rules are applied to automatically generate the Kelantan dialect pronunciations. For example, one of the G2P rules found is *final 'a' rule*. The grapheme 'a' at the end of a word is pronounced as /ɔ/, such as, the word "*gila"* (English: crazy) is pronounced as /g i l ɔ/.

Thousand frequently used original standard Malay words from Malaysia's local news web pages (700MB) [16] are extracted out. The pronunciation for each of the word extracted is then generated through the G2P tool built for Kelantan dialect. The produced pronunciation dictionary was compared with the manually verified pronunciation dictionary of Kelantan dialect. The phoneme accuracy achieves 99.14%. It showed that the G2P tool developed is effective and reliable.

With the translated Malay dialect text and pronunciation dictionary created, sentences were selected such that those with the most number of unique unseen monophones, triphones and pentaphones were selected, so that the context can be evaluated as many as possible. A proper weight was assigned to each phone while selecting the sentences. Sentences which are having maximized number of unseen monophone, triphone and pentaphone, with highest score will be selected first. To ensure that the selected sentences are phonetically well balanced, the phone distribution in the translated Malay dialect corpus was calculated. The following shows the formula for selecting sentences with the most number of unique unseen monophones, triphones and pentaphones from the translated dialectal text.

$$A_v = \sum_{x=1}^{i=A} \left(\frac{P_{vi} + T_{vi} + M_{vi}}{3} \right)$$

where A_v is the maximum score of monophone, triphone and pentaphone, P_v is the score of pentaphone, T_v is the score of triphone and M_v is the score of monophone, i is the iteration and A is the size of sentences.

$$P_v = \sum_{x=1}^{i=P} \frac{1}{v} \qquad T_v = \sum_{x=1}^{i=T} \frac{1}{v} \qquad M_v = \sum_{x=1}^{i=M} \frac{1}{v}$$

where P is the number of unique pentaphone, T is the number of unique triphone and M is the number of unique monophone.

4. Speech Synthesis Corpus Recording

We have recorded speech for 3 speakers. Two of the speakers, one male and one female, read only standard Malay sentences, while the third speaker spoke standard Malay and Kelantanese Malay speech. The recording in standard Malay can be used in the process of transformation from standard Malay to dialectal Malay for speech synthesis purpose. Around two thousands of sentences (about 4 hours) in text were selected for each standard Malay and Kelantanese Malay, with phonetically well balanced based on the phoneme distribution. To have an idea of the phone distribution in the selected sentences compare to the general phone distribution of standard Malay and Kelantan dialect, a correlation coefficient between these two are calculated. Table 2 shows the correlation coefficient for the selected sentences in standard Malay and Kelantan dialect calculated from the speaker who speak standard Malay and Kelantanese Malay.

Table 2. *Correlation coefficient for selected sentences in standard Malay and Kelantan dialect.*

Type	Correlation Coefficient
Standard Malay	0.9923
Kelantan Dialect	0.9893

The result shows that the selected sentences have a correlation coefficient of about 0.99 for standard Malay and Kelantan dialect, which means that it is phonetically well balanced. As our proposed approaches achieved high accuracy, they are reliable and effective to be applied. Figure 4 and Figure 5 show the phone distribution among selected sentences for standard Malay and Kelantan dialect.

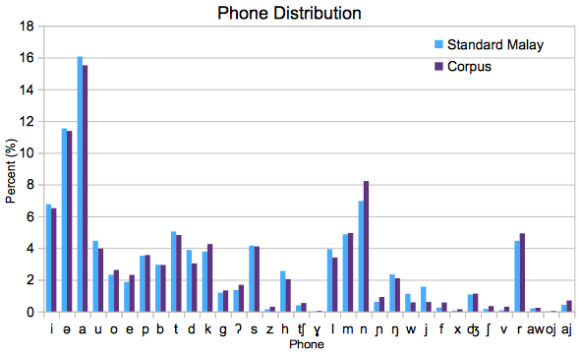

Figure 4: *Phone distribution for standard Malay.*

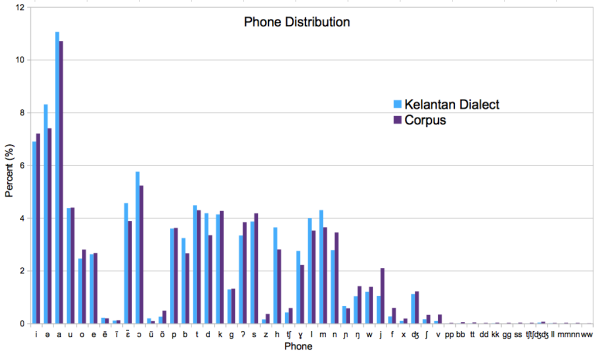

Figure 5: *Phone distribution for Kelantanese Malay.*

5. Building Speech Synthesis System

After acquired speech corpus, the phone in the utterances was aligned to create speaker dependent acoustic models. Since manual alignment of the utterances is expensive and time consuming, automatic alignment was applied by force aligning the utterances using an automatic speech recognizer, Sphinx3 from CMU. The aligned speech was then used to train acoustic model for the HMM speech synthesis system (HTS Speech Synthesis System). Currently, we have completed a first Kelantanese Malay speech synthesis system. The perception test carried out on 20 speakers to evaluate the quality of the Kelantanese Malay speech synthesis system compared to the natural speech of the speakers shows that, from the scale of 1 (bad) to 5 (excellent), the average rating given for the quality of the synthesize speech is "good", which is scale 4.

6. Conclusion

In this paper, we present our work in acquiring a dialectal Malay speech corpus, which including Malay dialect translation and Malay dialect grapheme to phoneme (G2P). With this, Kelantan dialect is chosen to be applied in the proposed architecture in this study. For future works, other dialectal Malay such as Kedah dialect and Perak dialect will be conducted using the proposed methodology.

7. Acknowledgement

This project is supported by the research university grant 1001/PKOMP/817068 from Universiti Sains Malaysia.

8. References

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

[2] K.-F Lee, H.-W Hon and R. Reddy, An overview of the SPHINX speech recognition system, Acoustics, Speech and Signal Processing, IEEE Transactions on, p.p. 35-45, 1990.

[3] H. Abdul, "Sintaksis Dialek Kelantan", Dewan Bahasa dan Pustaka, p.p. 3,1994.

[4] Uneson, Marcus, "Tone restoration in transcribed Kammu: decision-list word sense disambiguation for an unwritten language", Linkoping Electronic Conference Proceedings, vol 85, pp. 399-410, 2013.

[5] U. Mosel, Universität Kiel, Collecting data for grammars of previously unresearched languages, Draft for the International LingDy Symposium on Grammar Writing,, Tokio 8th.-10th Dec., 2009.

[6] F. R. Hanke and S. Bird, "Large-scale text collection for unwritten languages", Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1134-1138, 2013.

[7] 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.

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

[9] T. P. Tan and R. M. Bali, Malay Grapheme to Phoneme Tool for Automatic Speech Recognition, 2010.

[10] 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.

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

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

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

[14] T. P. Tan, S. S Goh and Y. M. Khaw, "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System", Proceedings of the 2012 International Conference on Asian Language Processing (IALP 2012), Hanoi, Vietnam, pp. 109-112, 13-15 November 2012.

[15] J. H. Tao, F. Z. Li, M. Zhang and H. B. Jia, "Design of speech corpus for Mandarin Text to Speech", 2008.

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