Building a Unit Selection Speech Synthesiser for Malay Language Using FESTVOX and Hidden Markov Model Toolkit (HTK)

Wai Keong Loo[1*], Siew Hock Ow[2] and Roziati Zainuddin[3]

[1,2]Department of Software Engineering, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[3]Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia

**Corresponding author. E-mail: david_loowaikeong@yahoo.com*

ABSTRACT

This paper describes the development of unit selection speech synthesiser for Malay language by using the Festival speech synthesiser system. The development processes involved the study of the phonological nature of language, letter-to-sound modeling and clustering on the phone unit database. An HMM-based speech recognition tool kit and Hidden Markov Model Toolkit (HTK) were also applied to label the unit boundaries automatically. To evaluate the naturalness of speech, an overall quality test on the synthesisers was also conducted to compare the naturalness of the earlier Multi Band Resynthesis Overlap and Add (MBROLA) diphone concatenation synthesiser and the new FESTVOX unit selection synthesiser.

Key words: Unit Selection Speech Synthesiser; Malay Language; Letter-to-Sound Modeling; Hidden Markov Model Toolkit, HTK; MBROLA Diphone Concatenation; FESTVOX Unit Selection Synthesiser

INTRODUCTION

Until the year 2001, there was only one text-to-speech system in the Malay language, namely, SUM (*acronym for Sintesis Ucapan Melayu*) which was developed by researchers from the National University of Malaysia. The synthesiser model is based on the Klatt's formant synthesiser. However, there are quite a number of shortcomings on this synthesiser. The synthesised speech lacks naturalness, but yet no effort has been made on the high level analysis and improvement in prosody (Hussain et al., 1999). Owing to the demand of local telecommunication industry, there was indeed a need for a quality speech synthesiser in the Malay language. This initiated research into the Multi Band Resynthesis Overlap and Add (MBROLA) diphone concatenation approach (Dutoit et al., 1996) and unit selection approach (Hunt and Black, 1996).

The Malay language is one of the western branches of Austronesian languages and is widely spoken among Malay-speaking countries such as Malaysia and Indonesia. The term "standard Malay" (SM) is a term that is basically accepted by the speech community to be the norm or the prestige dialect, which is also the official language in Malaysia. It is widely believed that the so-called "standard Malay" is based on the Johor-Riau Malay (JM) dialect, mainly spoken in the southern part of the peninsular Malaysia. There are other three dialects, namely, Kelantan Malay (KM), Ulu Muar Malay (UMM) and Langkawi Malay (LM) which are spoken in the different parts of peninsular Malaysia (Teoh, 1994).

Festival is a complete diphone concatenation and unit selection TTS system for British and American English, Spanish and Welsh (Black et al., 2000). It offers a general TTS research and development framework that includes tool sets for building a unit selection synthesiser in a new language (Black and Lenzo, 2003). The key idea behind the unit selection technique is to avoid the modification of speech segments so that the acoustic features of the unit are well preserved. In the unit selection approach, a large number of units are recorded and labeled. These units are then clustered according to their acoustic features and phonetics context. An appropriate unit from the pool of segments will be selected during the speech synthesis by searching through the pre-built indexed clusters. The search result would be the unit with a minimum cost of concatenation [Tokuda and Black, 2002]. The selected units are finally joined at the optimal position without any modification to the signal. Unlike the MBROLA diphone concatenation approach, the duration and the pitch level of the diphone units are adjusted towards the target prosodic effects.

The following sections focus on the phonetics aspect of the Malay language that covers the classification of the speech sounds and their syllabic structure; the lexicon and the letter-to-sound (LTS) modeling; the voice-building processes such as the automatic labeling of speech segments using HMM model; and the overall quality of the synthesisers.

PHONOLOGICAL VIEW OF THE MALAY LANGUAGE

There are some common features between the Malay language and English language. Firstly, Malay language is a phonetic language and it is written in Roman characters. Secondly, all syllables in the Malay language are pronounced almost equally and it is thus, considered as a non-tonal language. In general, there are 6 main vowels and 26 consonants in standard Malay. Nineteen of the consonants, /m/, /n/, /f/, /l/, /s/ and /y/ are pronounced almost the same way as in English.

In Malay language, the syllabic structure is well-defined and can be unambiguously derived from a phone string. The basic syllable structure of the Malay language is generated by an ordered series of three syllabication rules. The linguists claimed that Malay is a Type III language (Teoh, 1994), namely, of CV(C) type in which every syllable must have an onset and nucleus. Based on the CV(C) structure, coda is optional for the syllable in Malay language and open syllables are commonly found.

Phone Set Used in Speech Synthesiser

All the phones are defined in ASCII characters. Generally, the phone set covers 6 vowels (Table 1), 23 consonants (Table 2), 2 diphthongs (/ai/ and /au/) and 2 allophones. The two allophones, /k/ and /r/, are defined explicitly in the phone set. This is because they are acoustically different in spoken Malay. /r/ and /k/ usually appear at the syllable final or coda. /K/ represents a glottalised unreleased voiceless velar stop which occurs as a variant of /k/ in syllable final position. /R/ represents a weakly-articulated /r/ which occurs as a variant of /r/ in syllable final position.

Table 1. Classification of vowel sounds in Malay language.

	Front	**Centre**	**End**
High	i		u
Medium	e	at	o
Low		a	

Table 2. Consonants in standard Malay except those in brackets are loaned consonants.

Manner of Articulation	**Place of Articulation**					
	Labial	**Alveolar**	**Palate-alveoral**	**Palatal**	**Velar**	**Glottal**
Plosives–Voiceless	p	t			k	
Plosives–Voiced	b	d			g	
Fricative –Voiceless	(f)	s			(x)	h
Fricative–Voiced	(v)	(z)				
Affricate–Voiceless			c			
Affricate–Voiced			j			
Nasal	m	n		ny	ng, nx	
Roll		r				
Lateral		l				
Semivowel	w			y		

In some of the vowel sequences, when the first vowel is /i/ or /u/, as in *dia* or *dua*, it is followed by a glide of the type /y/ or /w/, respectively. The prominence of this glide varies, and it can be as prominent as the glide of *yang* or *wang*, or it can be much weaker. In view of this variable prominence, two semi-vowels /Ya/ and /Wa/, are included in the phonet set.

Similarly, when a diphthong is followed by a word-final consonant, as in *baik* or *baur*, the two vowels may be separated by a weak /y/ or /w/ brought about by a slight overshoot in the articulatory gesture. Hence, /Yi/ and /Wu/ are also considered as a phone type in text-to-speech (TTS). In considering the above variations in Malay sounds, there are a total of 36 phones defined in FESTVOX's phone list (Black et al., 2000).

LEXICON AND LETTER-TO-SOUND (LTS)

For Malay language, the alphabets in a word itself is good enough to identify its pronunciation. Not all words, however, can be pronounced exactly as it is written. Both lexicon and some form of LTS conversion are required.

The lexicon is not only to provide an accurate pronunciation of the word during the speech synthesis, but, it also plays a part in the process of building a TTS where the automatic labeling and segmentation method will refer to the pronunciations in the lexicon. There are 13,550 Malay word entries available in the lexicon. Each entry is associated with its pronunciation, syllable grouping of the phonemes and the stress level for each syllable. The stress level is expressed as "1" to indicate a primary stress and "0" for the secondary stress.

In reality, the input text to the TTS is an open context. It is impossible to include all possibilities into the lexicon. For this reason, the LTS prediction is needed as a fallback on the lexicon. In this research, both the rule-based and CART-based methods in LTS prediction were attempted.

Hand-Written Rules for LTS

Rules have traditionally been viewed as the primary source of knowledge in LTS conversion. Each rule is conditioned by *target, left context* and *right context*. The target is a string of input alphabets or a single alphabet which needs to be converted into the phoneme symbol defined in the phone set. The *left context* and *right context* is a string of a word's alphabets on the left and right position of the *target*.

The LTS rules are organised with the most specific cases first and end with a generic case. At the initial stage, 26 generic rules were defined for the 26 Roman characters. Each alphabet is mapped to its most common sound pronounced in Malay. With these generic rules, we can only get 26.25% of correct match on word pronunciation in the lexicon. The percentage increased when more specific rules were added into the rule set. When the rule set increased to 94 rules, there is a 71.0% match on the word level. However, the growth is constrained by a number of ambiguities in LTS conversion. Table 3 shows some of the most significant ambiguities.

Table 3. Ambiguity in LTS conversion.

Alphabet	Ambiguity in Sound	Examples
Al	/a/, /Yl/ or /ai/	baik -> /b/ /**a**/ /**Yl**/ /k/
		baiki -> /b/ /**ai**/ /k/ /i/
A	/a/ or /at/	tadika -> /t/ /a/ /d/ /i/ /k/ /**a**/
		jika -> /j/ /i/ /k/ /**at**/
R	/r/ or /R/	beranak -> /b/ /at/ /**r**/ /H/ /a/ /n/ /a/ /k/
		beramal -> /b/ /at/ /**R**/ /H/ /a/ /m/ /a/ /l/
U	/u/ or /O/	tunjuk -> /t/ /**u**/ /n/ /j/ /**u**/ /K/
		huruf -> /h/ /**u**/ /r/ /**O**/ /f/
E	/e/ or /at/	kemas -> /k/ /**at**/ /m/ /a/ /s/
		kemah -> /k/ /**e**/ /m/ /a/ /h/

Letter-to-Sound Rules in Tree Models

Recently, an attempt was made to automate the acquisition of LTS conversion rules. FESTVOX has provided the framework to applied CART tree model for the LTS prediction (Figure 1). This trainable method assumed that given those sets of words with correct phonetic transcription (lexicon), an automated training algorithm could capture its significant generalisations.

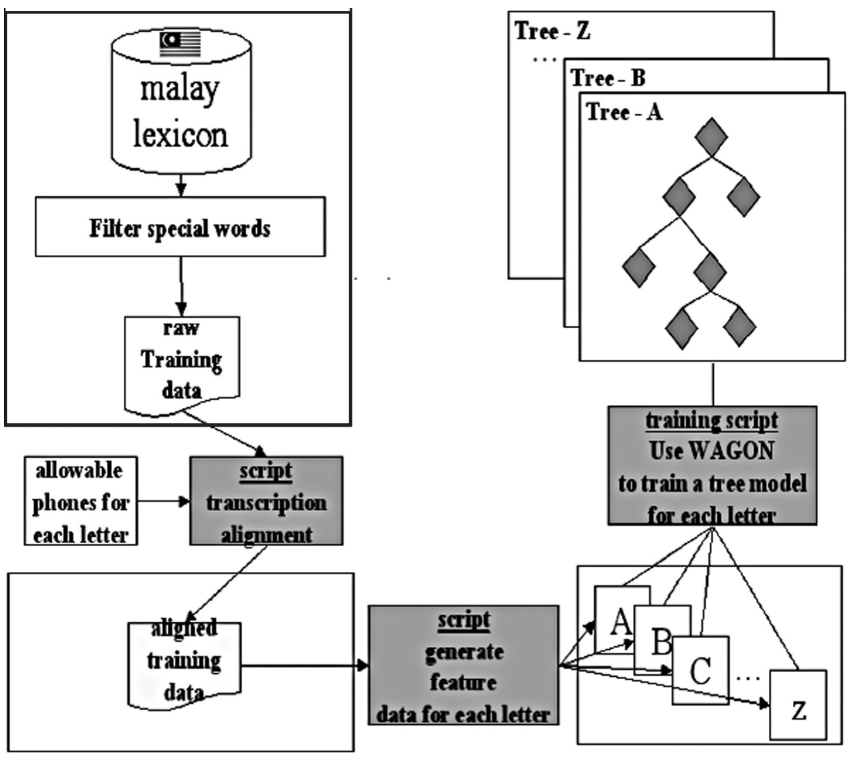

Figure 1. Schematic presentation of CART training in FESTVOX.

Ninety percent (90%) of word entries in the lexicon are used for training and the remaining 10% (1,304 words) are reserved to evaluate the output tree models. To ensure that the output tree models are accurate in representing the LTS conversion in SM, the special words such as abbreviations and loaned words from other languages are pre-filtered before they are used for the training. The overall test results for the CART prediction are given in Table 4.

Table 4. Summary of CART-based LTS evaluation.

	Total Test Data	**Correct (%)**
Letters	9,950	9,521 (95.68)
Words	1,304	964 (73.93)

The accuracy of the CART-based LTS prediction is about 73.0% which is slightly higher than the hand-written rules. Indeed, most of the CART-based predictions at the alphabet level achieved above 99.0% match with the lexicon except for a few alphabets, as shown in Table 5.

Table 5. Alphabets that are below 99.0% in LTS prediction.

Alphabets	**Total Test Data**	**Correct (%)**
U	472	416 (88.14)
A	1,793	1,593 (88.85)
I	750	693 (92.40)
E	873	821 (94.04)
R	545	519 (95.23)
K	536	512 (95.52)
O	190	182 (95.79)

Based on the results shown in Table 4 and Table 5, it is obvious that the ambiguity in predicting the vowels is relatively higher than the consonants in SM.

BUILDING THE MALAY VOICE DATABASE

In building the Malay speech database, the recording script with 100 sentences are first prepared from various types of news appearing in the newspaper. In addition, based on available triphone patterns in the news script, we can identify the triphone patterns in the lexicon that did not appear in the news script at all. By comparing the amount of triphones in the news scripts and triphones in the lexicon, it is found that there are about 4.8% of triphone patterns not covered in the news corpus.

Based on the analysis, an additional 50 phone balanced sentences were generated into the script to increase the triphone coverage in the corpus. With the addition of phone-balanced corpus, the triphone coverage can be increased to 6.7% out of the theoretically-possible triphones in Malay language. The recording was done in a quiet recording room with a normal personal computer microphone. All

the 150 sentences were recorded at 16 KHz. Table 6 shows the basic statistics of the corpus.

Table 6. Basic statistics of the Malay female voice corpus.

	Malay News (100 Sentences)	Phone Balanced (50 Sentences)	Lexicon
Average Duration	19.4 sec	16.2 sec	-
Total Duration	2,148 sec	901 sec	-
Number of Words	2,133	500	13,550
Number of Phone	11,242	4,398	65,523
% Triphone	5.2	1.7	9.9

Automatic Phone Labeling

To support the building of unit selection voice, an HMM-based triphone model in Malay language was used for automatic labeling and segmentation. The automatic labeling system can quickly generate the phoneme labels for the sentence by referring to the lexicon. The generated labels were then matched with the wave file, using the HViTe program in HTK (Young, 1997). The speech corpus of 50 minutes was first labeled automatically using HTK. These HMM-labeled phone boundaries were then fine-tuned by human.

To evaluate the accuracy of HMM labeling, the HMM-aligned boundary set is compared with the final version of the hand-tuned boundary set. Figure 2 shows a histogram of timing error on HMM-forced alignment. The errors are measured against the result of the hand-tuned labels. Altogether, 1,700 units in 20 utterances were compared. As shown in Figure 2, the timing error of 1,349 units out of the total of 1,700 units is within the range of 150 ms.

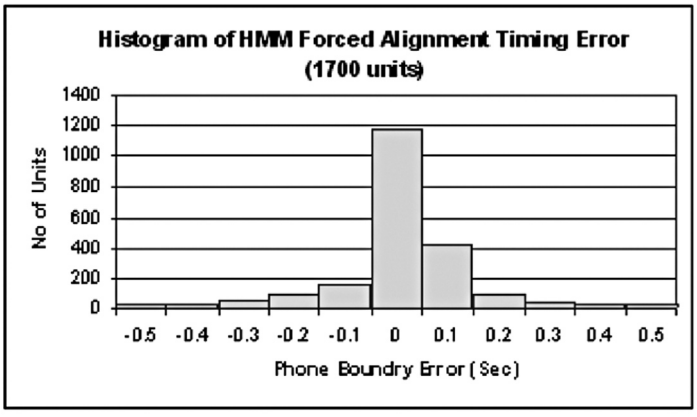

Figure 2. Histogram of HMM labeling timing error.

The striking plot of both consonant-vowel (CV) and vowel-consonant (VC) boundaries were found 10 ms away from the target position. In general, the majority

of the unit boundaries (including the vowel-vowel and consonant-consonant) are within the error range of 10-30 ms (Figure 3).

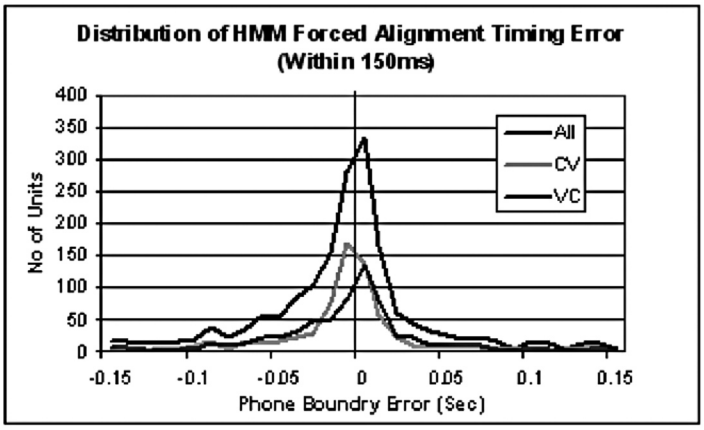

Figure 3. Distribution of HMM labeling timing error that is within the range of 150 ms.

Human labeling without the automatic pre-alignment will take approximately 30 minutes per sentence. Working from pre-aligned boundaries, the amount of time required is 30-40% shorter by just adjusting the segment's boundary position and correcting some of the transcription errors. As a result, the amount of work and attention on human labeling is reduced with the help of HMM pre-alignment.

OVERALL QUALITY TEST ON MALAY SPEECH SYNTHESISERS

The overall quality test was measured in mean opinion scores (MOS). The MOS test was conducted to assess the naturalness, pleasantness and the clarity of the synthesisers (Figure 4). Six sentences were synthesised from two different TTS and pre-recorded into wave files. Listeners were divided into two groups of 10 for each group. The first group listened to the speech generated from diphone concatenation synthesiser and the second group listened to the speech from the unit selection synthesiser. After listening to the six sentences, listeners were requested to give ratings on the perceived naturalness, pleasantness, clarity and the listening effort required over synthesised speech. The MOS for naturalness, pleasantness and clarity are scaled from 1 to 5 where: 1 - Bad, 2 - Poor, 3 - Fair, 4 - Good and 5 - Excellent.

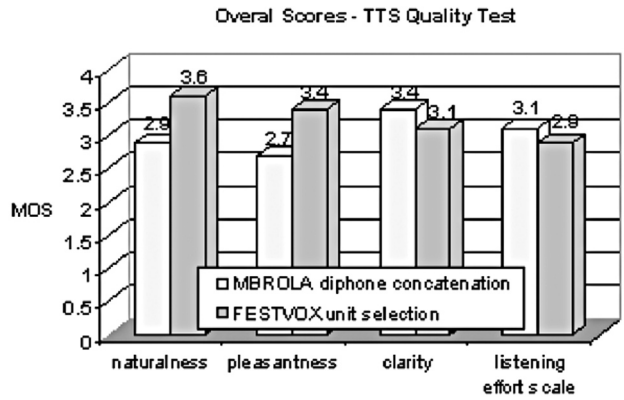

Figure 4. Overall quality test on both Malay synthesisers.

DISCUSSION

Based on the feedback, listeners encountered some difficulty in identifying some of the words that were synthesised by unit selection approach. These ëbad' words usually contain a significant spectral discontinuity at the CV (vowel-consonant) and VC (vowel-consonant) joined or the phone segment was wrongly labeled (considered as bad phones) during the labeling stage. These worst-joined segments in synthesised speech can decrease the understandability of the word level and even to the phrase level. In this respect, the inter-segmental joins in diphone synthesis are much smoother than unit selection synthesis. Adding more units into the unit database can minimize this issue of unit selection synthesis. We noticed that some vowel units such as /o/ and /u/ are still lacking in our unit selection inventory. Having more instances per unit type can reduce the concatenation distortion.

On the aspect of naturalness and pleasantness, unit selection synthesis is preferable to diphone concatenation synthesis, as shown in the test result. Listeners perceived that the speech in unit selection synthesis is more desirable as it sounds almost like human's prosody whereas the diphone concatenation synthesis sounds more artificial and robotic.

REFERENCES

Black, A. W., and K. A. Lenzo. 2003. Building synthetic voices. http://www.festvox.org/festvox/festvox_toc.html, accessed on 6 January 2005.

Black, A. W., P. Taylor, and M. Macon. 2000. Speech Synthesis in Festival. http://festvox.org/festtut/notes/festtut_toc.html, accessed on 30 August 2006.

Dutoit,T., V. Pagel, N. Pierret, O. van der Vreken, and F. Bataille. 1996. The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In proceedings of ICSLP '96, Philadephia, P.A. Vol. 3: 1393-1397.

Hunt, A., and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In proceedings of the ICASSP, Atlanta, Georgia, Vol 1: 373-376.

Hussain, A., S. Abdul Samad, and T. S. Kuek. 1999. Theory, methodology and Implementation of the malay text-to-speech system. Malaysian Journal of Computer Science. Vol.12(1): 28-37.

Teoh, B. S. 1994. The sound system of malay revisited. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Tokuda, K., H. Zen, and A. W. Black. 2002. An HMM-based speech synthesis system applied to english. http://www.cs.cmu.edu/~awb/papers/IEEE2002/hmmenglish.pdf, accessed on 30 August 2006.

Young, S. 1997. HTK Book, University of Cambridge. http://htk.eng.cam.ac.uk, accessed on 17 January 2005.