Towards Designing A High Intelligibility Rule Based Standard Malay Text-To-Speech Synthesis System

Zakiah Hanim Ahmad, Othman Khalifa
International Islamic University Malaysia, Malaysia
zakiahhanimahmad@yahoo.com.my

Abstract

Although text-to-speech (TTS) technology has gained some interest from amateur and professional researchers in developing a Standard Malay (SM) text-to- speech synthesizer, however, up to this day, there is rarely any high intelligible TTS system which is freely accessible to be implemented and introduced to the community of SM speakers. Therefore, identification of the core components required for the development of SM TTS system especially in establishing the NLP module should be carried out intensively. This paper presents a rule-based text- to- speech synthesis system for Standard Malay, named SMaTTS. A*n intelligible and adequately natural sounding formant-based speech synthesis system with a light and user-friendly Graphical User Interface (GUI) was developed. Result and suggestion for future improvements is discussed. The available Malay TTS synthesizers, the algorithms and speech engine in used, as well as their strong and weak points for each of the system are discussed in this paper. Assessment was made at all possible levels; phoneme, word and sentence level. The overall performance of the system is analyzed using Categorical Estimation (CE) for a comprehensive analysis. Result and suggestion for future improvements is discussed.*

I. Introduction

Text-to-Speech has become a crucial part of speech technology. The ability of a machine or computer to read any text aloud would enable a device to be used in various situations and purposes. The scope of work will be focusing on first of all, establishing a TTS system that would be able to pronounce common valid SM words, and then having the system correctly pronounces other loan words available in SM. We are looking forward to establish a system that would be able to utter SM sentence or text correctly and adequately intelligible to approximate a native SM speaker, whereas the intonation, rhythm and pitch are correctly ruled out. Main work is to construct a flexible phone database that is able to be used to utter flexible SM words, while in term of the result; main works are concentrated in producing intelligible single words.

II. Available SM TTS Synthesis Systems

Most of the Malay speech synthesizers proposed are generally small scale projects based on other open source or commercial speech synthesizer. There are many TTS systems available nowadays but only few can be considered as able to give great contributions to the world of Malay speech synthesizer. This section will discuss only the most significant systems or proposals in term of theoretical frameworks, NLP module, DSP module or even the system itself.

The first research ever recorded in designing a TTS synthesis system for SM was carried out by Aini Hussain, Salina Abdul Samad and Kuek Teik Soon from University Kebangsaan Malaysia, named SUM, acronym of Sintesis Ucapan Melayu. This synthesis-by-rule synthesizer is based on Klatt's formant synthesizer, KLSYN88 with the second version added some breathiness and flutter effect. Like any other formant-based TTS synthesis system, a flexible system structure of the software has made future improvements possible. However, there was an obvious minus point in the lack of naturalness of the output speech [1] on both version of SUM.

The most successful Malaysia text-to-speech software was thought to be FASIH which was launched by Mimos Bhd in 2005. MIMOS claimed to be the first and the most successful Malay TTS engine due to the ability of the system to produce an unrestricted vocabulary of Standard Malay with natural sounding speech, compared to other machine-sounded Malay TTS system. MIMOS applied Natural Language Processing (NLP) for prosodic analysis by assigning prosody information to the raw input text to generate

978-1-4244-1692-9/08/$25.00 ©2008 IEEE

output speech [2]. Given a text, NLP module for FASIH includes the part-of-speech (POS) tagging in determining the type of each word, phrase and sub-phrase classification for pause allocation, new word type and pitch contour assignment.

This diphone based concatenative TTS system uses time domain MBROLA as its speech synthesizer [3] where MBROLA itself was inspired by MBR-PSOLA algorithm. The diphone database which is specially adapted to the requirements of the synthesizer was obtained via hybrid Harmonic/Stochastic analysis-synthesis of the database, resulted in the flexibility of parametric speech models while keeping the computational simplified. The output of speech can either be directly spoken via the computer speaker or saved as wave (.wav) files [2]. The first version of Fasih was successfully commercialized and used in training software, QuickDo [4]. Other applications included e-mail reading, language-based training software and other typical voice service applications. Having inherited MBROLA diphone-based method, the drawback of such TTS system is that it must have all the knowledge and data of SM for grapheme-to-phoneme processing.

Another attempt to build a Malay speech synthesizer by adapting MBROLA algorithm was made by Nur-Hana Samsudin and Tanya Enya Kong [5] from University Science of Malaysia. The system used four syllable structure of consonant-vowel clusters (CV), vowel-consonant clusters (VC), consonant- vowel –vowel clusters (CVC) and vowel (V) cluster with a few sub-models proposed for exception such as loan words pronunciation. The database used prerecorded syllable segment from a native Malay speaker to avoid phonological problem derived from the use of Speech Application Programming Interface (SAPI) due to the fact that only American English phonological representation is used in the interface, hence yield to the sound of Malay Language being very foreign and awkward. However, segment discontinuation and distortion at the boundaries are obvious since the database was built without prosody modification.

Another Malay TTS synthesizer was established by Yousif A. El-Imam and Zuraidah Md. Don [6]. They proposed a system based on unit-selection methods with four synthesis units, namely, CV, VC, vowel-consonant- vowel clusters (VCV) and consonant-consonant clusters (CC). Each of the synthesis units contains 162, 162, 972, and 729 clusters respectively. All the input text would first tagged with this CV rules before the syllable segmentation that is used to process text utterance can be obtained.

The system which adapted from a previously developed synthesizer for Standard Arabic language also proposed a general linguistic analysis and phonological aspect of Standard Malay and loan words from Arabic language that can as well be implemented to the NLP module of our Malay TTS synthesizer system. A lexicon containing all the special properties such as abbreviations, acronyms, and special symbols will divide the user input into two fields, the orthography of the item and its pronunciation of the words or the representative word sequence. The database would be scanned for the first matching entry.

In Say It! System [7], the segmenting technique is to select the longest phoneme sequence and compare the selected sequence in the available syllable database. If matches occur, the sequence will be taken out and consider as a syllable unit. Else, the last phoneme in the done again with the reduced phonemes sequence. The process will be repeated until the match is found in the database. This technique does provide a simple implementation and produced quick result but the parsing could also be segmented wrongly.

There was also an attempt to build Malay TTS synthesis system made in 2004 by Tan Tian Swee for his Master Thesis [8]. With ‘festival’ as the speech engine, the diphone-based TTS synthesizer generates its database in residual-exited LPC (RELP) format. Festival is free software distributed under an X11-type license. The overall programming language of this speech engine is written in C++ and uses the Edinburgh Speech Tools Library for low level architecture. The LTS rule for this Malay TTS synthesizer is coded using Scheme programming language. There is some complexity in handling the LTS rules where all the syllable types and vowel pairs must be recognized by the system, thus restricted the flexibility of the system in uttering any possible SM words including words of foreign origins and scientific terms. Special digraphs handling such as <kh>, <ng>, and <sh> needed to be define in special rule due to the concatenative process of diphone, where by default, the system would read words such as “tangan” (hand) as <tangan> instead of <taŋan>.The total size of 3.4 Mbytes is also big for embeddable system. Moreover, until the time of this writing, the development of ‘festival’ seems to stop at beta version of 2.0 since December 2004 [9].

Other development of speech synthesiser for Malay language by using the Festival speech synthesiser system was made by Loo et al [10] from University of Malaya in 2007. The authors of this unit selection speech synthesiser claims that the system produces a more natural output speech that approximate the prosody of a human speech, compared to diphone

based concatenation synthesis which sounded more artificial. This claim was made upon comparison to diphone based MBROLA speech synthesis. On the other hand, the huge size of database demanded high memory requirements while there exists a significant spectral discontinuity at the joined of consonant-vowel (CV) and vowel-consonant (VC) cluster or wrong labelling of phone segmentation at the labelling stage. This major drawbacks decrease the ability of the users to perceive the word spoken by the system [10].

III. SM SYLLABLE STRUCTURES

SM is classified as type III in the class of languages [11] where syllables begin with an onset that is the initial consonant of a syllable. Nucleus is normally vowel and it can be either monophthong or diphthong while onset and coda are optional and can exist together or either one.

TABLE 1: SM GENERAL SYLLABLE STRUCTURE

Final (rhyme)		
Onset	Nucleus	Coda

SM vocabularies consist of words where most of its syllables follow the form of V, VC, CV, and CVC where C indicates consonant and V indicates vowel. However, taking that many loan words have been trough a lot of modification and has now been absorbed as Standard Malay, the syllable structure of SM has extended to various shape. Dewan Bahasa dan Pustaka (DBP) or the Institute of Language and Literature for Malay Language, (the regulatory body to promulgate changes to Malay Language) in 1985 has concluded that the possible cluster consonant (C) – vowel (V) might as well extend to the additional combinations of VCC, CVCC, CCV, CCVC, CCVCC, CCCV, and CCCVC.

From the examples given, we can hypothesize that all syllables in Standard Malay have the following general form:

$$S = C^x V C^y$$

Where S = syllable form, and

$$x = 0,1,2,3 \; and \; y = 0,1,2$$

$$x + y \leq 4$$

In this equation, C^x indicates 0 to x number of consonants, V indicates vowels which is the nucleus of a SM syllable and C^y indicates y number of consonants in syllable coda, while the number of consonants in one syllable [12] do not exceed four, having acknowledged that even though the spelling of loan words might have changed to imitate Malay semantical spelling, it is the nature of native speakers and other SM speakers to associate their utterance into the original C-V cluster of pairing consonant with vowel as possible. Even after eliminating the combination of CC clusters of five digraphs in SM, namely, <gh>, <kh>, <ng>, <ny> and <sy>, and the fact that not all combinations of consonants are applicable [13], the possible number of syllables is still tremendous.

The above CV combination is made possible since Malay Language has modified many loaned words before it can be recognized as Standard Malay. Consequently, The CVCCC or VCCC cluster are not possible due to the same reason. From our observation, SM phonotactic rules determine that a closed syllable cannot have coda in form of three or more of adjacent consonants and all the loan words in that particular form would be given Malay orthography system. For instance, the word “arch” and “March” that has syllables of VCCC and CVCCC each are borrowed into Standard Malay with the spelling changed to “arca” /ʌtʃ a/ and “Mac” /matʃ / respectively. These examples of lexical borrowing reflect the role of globalization in the shifting of Malay toward English in language contact situations with a lot of English lexical loaned to SM. Similarly, while some of the structures of the original words are kept, many have gone through adaptation. For example, the spelling of loaned word “television” has been changed to “televisyen” and has also adapted Malay pronunciation as /tɛ lɛ vi:ʃ en/. In contradiction, English word “gear” kept its structure and phonetical sound while the spelling of the word “metre” is changed to “meter” but still have the English phonetical representation. Thus, for the latter cases where the phonological system of Standard Malay are not followed and the morphemes are unidentified, there are four options left:

1. The terms or pronunciation of the words are stored in the `exception list` in the database. In our database, this exception of phonetical words or syllables is listed in file name ms_list while the rules are clarified in file ms_rules. This exception lexicon also stored abbreviations, numbers, etc.
2. New rules are set up. For example, any word ending with “-ter” or word ending with “–sion” are normally English loaned word where the underlying phonetic should remain its phonetical sound /tə/ and /ʃ ɛ n/ respectively.
3. The terms are tagged and rules are set up so that the pronunciation would be referred to English lexicon.
4. Malay pronunciation is forced if possible.

IV. LETTER-TO-SOUND (LTS) RULES

As to systematically investigate the related phonetical rules, we used the following generative phonology as described by Chomsky and Halle [11], which is based on rewrite rules and written in the following form:

$$p)\ a\ (f \rightarrow b \qquad \text{Equation 3.1}$$

which implies that segment a is written as segment b if immediately preceded by p string and followed by f string, where p and f can either be a single orthographic character, strings of characters, syllable, null or punctuation marks. El-Imam and Zaharah [6] has ruled out seven letter-to-sound (LTS) rules with each category has its own sub-rules totaled in 29 rules. All the seven rules are schwa deletion of the grapheme <a>, glottal stop insertion rules, final <r> deletion, diphthong generation rules, consonant deletion rules and vowel replacement rules. However, we need to make some applications before being applied to our system. Hence, having considered all possibilities, we have restructured the morphophonemic module to be in the following format [3] to best suit our TTS system. This way, we can also systematically modify, alter or revise the rules if necessary or if there is any mistake or inaccuracy in output can be detected easily:

1) Changing phonological features
   a) Vowels nasalization if vowels is preceded or followed by nasal consonants - A vowel is predictably nasalized if preceded or followed by nasal consonants, /m/, /n/, /ng/, and /ny/.
   b) Vowel as nucleus in closed final syllable - A vowel in a close syllable that forms a word [12], does not sound the way it used to if it were in any other position. For example, grapheme <i> in "pasir" (sand) does not share the same phone as in "hisap" (suck). The previous case was categorized as the third group of final 'r' deletion [6] [14].
   c) Prefixes in SM words of English origin that keep the English orthographical and spelling system, such as "pre-", "uni-", "de-" and "re-". The example of words with these prefixes are "universiti", "deformasi", and "reformasi".
   d) Diphthong generation rules - Diphtongs in SM are /ai/, /au/ and /oi/.
   e) Voice obstruents devoicing in syllable final position - Voice obstruents /p, b, g, d/ are devoiced in syllable final position [13].
   f) Glottal formation rules - After some research and experiments, we need to conclude that voiceless velar stop <k> at syllable coda position manifested to glottal stop although there have been some debate among the linguist over this problem [13]. For example, in our TTS system, "masak" is represented with /ma:sa?/.

2) Consonant deletion rules
   a) Digraphs "ch", "sy", "ny", and "ng", are deleted and replaced by the phoneme sequences of /tS/, /S /, /nj/ and /N/ respectively.

3) Consonant insertion rules
   a) Seven rules of vowel consequences. For example, in the group "ia", "iu", "io", and "i", the consonant "j' is inserted to result in "ija", "iju", "ijo", and "ij", respectively. There is also consonant "w" insertion in vowels consequence of "ua" and "ui" resulted in "uwa" and "uwi".
   b) Glottal stop insertion rules. This rule deals mainly with glottal stop insertion when a word begins with a vowel. [6]. For example, the word "ambil" (to receive) can have the glottal stop inserted to become "?ambeil". And some exceptions for few words such as "masalah" (problem), "kaedah" (method) [13] which is pronounced as /masʌ ?alax/ and /ka?edax/ respectively. Experiments also suggested that same vowel sequences /a-a/, /e-e/, /i-i/, /o-o/ and /u-u/ needed glottal stop insertion in between the vowels. For example, "suun" (rice noodle, Chinese loanword) and "rai" are pronounced as /su?on/ and /ra?i:/ respectively.

4) Vowel replacement rules
   a) The remaining vowels are replaced by its phonetic transcriptions [6]. For example grapheme <a> is replaced by the phoneme "a".

5) Stress shift - Stress shift refers to the change in the placement of stress or tone to reflect a contrast in lexical category. In Standard Malay, the positions of stress are mainly affected by the presence of schwa (surfaces, or etymological) in the root [15].

In his research, Mark Donohoue has made some comparisons on the stress shift in varieties of Malay and Indonesia. As for SM, we agreed with his analysis that schwa which is treated as epenthetic allows for a regular treatment of stress in most varieties of the language, but the behaviour of stress under affixation remains an independent variable. For example,

Case 1: “pasar” (market), “duduk” (sit), and “botol” (bottle)
Case 2: “besar” (big), “kecil” (small) and “betul” (correct)

In Case 1, stress is placed on the first syllable for each word while in the latter case, stress occurs in the final syllable. Stress shift occurs in suffixed words. In SMaTTS, dissyllabic root words are stressed in the first syllable.

As for complex sentences, Ann Delilkan [16] proposed that other than SM bare root words, adjacent syllables must contrast for stress where primary stress is shifted to the right of the entire complex word as far as possible, so that adjacent syllables would contrast for stress. She also claims that light syllables are not to be stressed in SM, except in prefixed or suffixed roots or when a prefix gets stressed in conjunction to her first claim that adjacent syllables contrast for stress. In addition, upon stressing these syllables, the stress placement in question produces unmarked (trochaically stressed) feet.

V. Database construction

Sinusoidal method is applied in producing sounds. For examples, a vowel sound is generated via adding the sine waves of the various harmonics where different vowels are produced using different mixtures of harmonic signal. Some consonants such as /p/ and /t/ are pre-recorded wave (.wav) files while /z/ is the combination of both [17]. The use of phoneme database significantly decreases the amount of computer memory requirements, thus making the system very light and embeddable.

A dictionary containing almost 50,000 SM words was built collecting common SM words including common phrases, common suffixed words, frequently used scientific terms and some loaned words. The construction of the dictionary would help in ruling out exceptions while detecting the pattern of phonotactical rules that exists in SM. LTS rules that is the most applicable and compatible to phoneme based SMaTTS was built where some revisions and modifications were made on the rule proposed by the previous authors as explained in Section IV to best fit our formant, phoneme based system.

Listening test is carried out on both Linux and Windows XP platform, using headphone to gain clear accuracy. This task is performed by testing out random SM words in the dictionary. The software in use is espeakedit and the SMaTTS itself.

VI. Graphical user interface

Although the speech engine allows the synthesis task to be performed using the command line, the GUI was designed in an attractive and user friendly design targeting on all users from different ages and educational backgrounds, whom most of them are non-programmers. Since SMaTTS is designed for the use in hands/eyes busy environments and should not require too much training, a simple GUI was developed where end-users can simply press the button to get the system to read a word, text, file or webpage aloud.

The advance mode button allows easy slider and button configuration. The objective of providing room for flexible parameters is to allow for a wider variation of speech output by changing the pitch and the amplitude of the synthesized speech and the speed of the utterance, to equip with different user preferences.

VII. Listening tests

To ensure the validity of the results, tests are conducted carefully and systematically, in the most similar approach as stated in the American National Standards Institute’s (ANSI) approved procedure (ANSI S3.2-1989 (R1999) American National Standard Method for Measuring the Intelligibility of Speech Over Communication Systems). The assessment for this system was made at four levels; phoneme, word, sentence and comprehensive reading and borrowed heavily from the testing method described in Meyer Sound Laboratories, Inc [18].

The test was carried out in actual implementation surrounding, involving ten adult correspondents. All these participants are SM native speakers without hearing or visual impairments. Five of the participants have never had any experience using any type of TTS and have least information about how a TTS system works whereas the other five participants have somehow exposed to the system and familiar with synthesized speech.

From the result, SMaTTS was proven to be highly intelligible in handling phoneme, single words and short sentences. Upon observation, the utterance of a long text or file reading is perceived well after listeners got exposed to the system. Sentence level accuracy test has shown a very encouraging result of 95.3% accuracy with six out of ten sentences were answered correctly by all participants. This prove the intelligibility of the system at sentence level.

The overall evaluation of the system employed CE method due to its comprehensiveness and strict evaluation system which is a good way to detect the

strong and weak points in the system. Ten participants would read and then listen to the text being read by the system before the text is assessed with the score from 1 to 5, where 5 is very good, 3 is fair and 1 is very bad.

A high score of 5 is recorded once for each distinctness and pronunciation attribute. Other than the naturalness and stress which scores 2.3 and 2.7 respectively, the average score for all other attributes are above 3.0, which imply that the synthetic speech is acceptable and fairly perceived. The pleasantness category scores 3.1 while pronunciation, distinctness and intelligibility attributes show a good score of 3.3 respectively. Comprehensibility marked the highest CE score of 3.6 with some response from listeners that they have no problem in perceiving almost every word in the text passage.

CE method is a subjective task and might be influenced by human factors and other distractions. For example, all participants were observed to be giving a score of 2 and above with many of them seem to be comfortable by marking 3 to most of the attributes. Some participants whom have never heard synthetic speech from any TTS systems claimed that they were giving out lower score due to their high expectation that the output speech would highly approximate a natural human speech. Although all these subjective influences were disregarded in the evaluation tasks, it is good to keep in mind the high expectation towards this technology for future improvement.

VIII. CONCLUSIONS

The constructions of vowels are relatively simpler compared to phonemes, thus, the test material focuses more on consonants evaluation. SMaTTS was proven to have a high intelligibility in handling phoneme, word and sentence level as well as producing rather good output speech for a long sentence and comprehensive reading. In fact, the rule based TTS system is also very flexible where it can utters almost any SM words including loaned words and scientific terms, as well as allowing enough possibility for future amendments. On the other hand, prosodic specification, number processing and homographs handling are the weakest aspects of this rule-based system and were major barriers in approximating a natural speech. Even though the summary should not be taken as a complete and thorough description, the results and discussions at each phase would provide a methodical and comprehensive guideline for quality assessment and future improvement. As for future works, since SMaTTS was using simple prosodic rules to predict the intonation, stress and duration, it is suggested that the implementation of Neural Network (NN) in this system in recognizing the sound pattern and synthesizing fundamental frequency (F0) contours would be a good solution to produce a more dynamic intonation.

REFERENCES

[1] A. Hussain S.A. Samad and K.T. Soon. (1999). *Theory, Methodology and Implementation of the Malay Text-To-Speech System*. Malaysian Journal of Computer Science, Vol. 12 No. 1, June 1999, pp. 28-37

[2] http://www.mimos.my/index.php?sub=6&ma=37&cat=about%20us&sub_id=18

[3] T. Dutoit. (1997). An Introduction to Text-To-Speech Synthesis. Kluwer Academics Publisher, The Netherlands.

[4] www.mosti.gov.my/MostePortal/website/images/PDF/ENG%20MOSTI-170106%20.pdf

[5] N.H. Samsudin and T.E. Kong. (2004). *A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach*. MMU International Symposium on Information and Communications Technologies 2004 (M2USIC 2004).

[6] Y.A. El-Imam and Z.M. Don. (2000). *Text-to-Speech Conversion of Standard Malay*. International Journal of Speech Technology 3, Kluwer Academic Publishers, pp. 129-146.

[7] http://www.nusuara.com/products/tts.php

[8] T.T. Swee. (2004). *The Design and Verification of Malay Text To Speech Synthesis System*. Master thesis. Dept. of Engineering (Electrical), University Technology Malaysia.

[9] http://www.cstr.ed.ac.uk/projects/festival

[10] W.K. Loo, S.H. and R. Zainuddin. (2007). *Building a Unit Selection Speech Synthesiser for Malay Language Using FESTVOX and Hidden Markov Model Toolkit (HTK)*. International Journal of Chiang Mai University, Vol. 6 No. 1, 2007, pp. 149-158.

[11] N. Chomsky and M. Halle. (1968). The Sound Patterns of English. New York: Harper & Row.

[12] www.karyanet.com.my/knet/ebook/pedoman/PanduanUmumEjaanBahasaMelayu.pdf

[13] T.B. Seong. (1994). *The Sound System of Malay Revisited*. Percetakan Dewan Bahasa Dan Pustaka. Kuala Lumpur.

[14] Y.K. Tan, T.B. Seong. and L.Haizhou. (2004). *Grapheme to Phoneme Conversion for Standard Malay*. Proceedings of the International Conference on Speech and Language Technology. (O-COCOSDA 2004), New Delhi, India.

[15] M. Donohue. (2007). *Conditions on Stress in Varieties of Malay/Indonesian*. Monash University. The Eleventh International Symposium on Malay/Indonesian Linguistics (ISMIL 11). 6-8 August 2007.

[16] A. Delilkan. (2002). *Stress Placement in Complex Words in Malay*. New York, University, New York. TLS VII: 2002 Proceedings.

[17] http://espeak.sourceforge.net/

[18] http://www.meyersound.com/support/papers/speech/