Syllabification Algorithm based on Syllable Rules Matching for Malay Language

HAFIZ MUSA, RABIAH A.KADIR, AZREEN AZMAN, M.TAUFIK ABDULLAH
Department of Multimedia, Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang. Selangor Darul Ehsan, Malaysia
hafizmusa@gmail.com, {rabiah, azreen, taufik}@fsktm.upm.edu.my

Abstract: In this paper, we present a new syllabification algorithm for Malay language. Syllabification is the process to extract or divide syllable from words. Syllabification process is language dependent where each language can have its own set of syllable structure. Syllabication is an important component in speech synthesizer, speech recognition and transliteration system. Syllabification algorithms have been proposed in many languages including English, Spanish, Myanmar, Singhala, Chinese and etc. Unfortunately, there are not many information regarding evaluation of syllabification scheme for Malay. In this paper, we propose an efficient algorithm based on syllable rules matching. In order to evaluate the algorithm, a prototype has been developed to measure the accuracy of syllabification. We evaluate our method using Bernama, Kamus Dewan and Overlap data collection. The syllable rules matching achieved 60.7% accuracy on BERNAMA collection, 77.4% on Kamus Dewan Collection and 71.6% on Overlap collection.

Key-words: - Syllabification, Text-to-Speech, Syllable Matching, Speech Synthesizer, Elicitation

1 Introduction
Syllabification is an essential component of text to speech system[1]. For example syllable unit is use to produce natural sounding and unrestricted speech synthesis system for Hindi. [2]. In unit selection text to speech system, syllable produce acceptable result if is in the same context for certain language. Raghavendra and Yegnanarayana [3] shown that syllable based synthesizers produce better sounding speech than diphone and phone. Syllable and other sound unit is also use to exploit acoustic representations of speech for synthesis, together with linguistic analyses of text to produce natural sound text-to-speech system.

In other experiments, it’s shown that syllable produce promising result in Telegu, Bangla, Romanian, Turkish text-to-speech system .[2][3][4][5]. In the area of automatic speech recognition, syllabic-like units has been proposed as a building unit in speech recognition system as alternative to phonemes unit. Syllabic-like units is that by modeling perceptually more meaningful units, the better modeling of speech can be achieved. [7].

1.1 Syllabification Theory
Onset-Rhyme (OR) models of syllable structure were developed by Fudge in the 1969 If we look vowel and consonant from phonological perspective, a syllable is often made up of a consonant plus a vowel or a single vowel. This follows the principle of maximal onset – minimal coda. The maximal onset principle states that the maximum number of consonants possible to attach a syllable onset [8]. In the Fig.1, syllable is made of rhyme and onset. Within rhyme (or core) we find peak (or nucleus) and coda.

![Fig.1: Syllable Structure](image_url)

We choose consonant and vowel (CV) sequence as phonological unit because almost all language has CV or CVC word. Different languages have different structure of syllable such as V, CV, VC, CVC, etc. A syllable is a unit of speech that consists of a vocalic center (nucleus) surrounded by a consonantal onset and a consonantal coda,
one or both of which may be empty. If C stands for Consonant, and V for Vowel, the structure of Malay syllable is [C]V[C], where the two C’s are optional (indicated by square bracket). The consonant C can be represented by one, two, or three characters. The vowel V can be represented by one or two characters [10].

In Malay language, onset and coda for syllable which have more than one consonant are limited to load word only. For example, word psikologi which are loan word from English. It also note that Malay syllable structure are different from English even for loan word.

1.2 Syllabification Approach
Until now, diverse syllabification approach has been presented in different papers. Shankar Ananthakrishnan used statistical syllabification approach, which apply maximum likelihood estimation (MLE) and expectation maximization for English language [10]; Karin Müller use probabilistic context-free grammar for multilingual syllabification [5]. Ouellet and Dumouchel introduced Heuristic Syllabification using N-gram statistical method for English language. Heriberto Cuay’ahuitl developed Spanish syllabification algorithm based on grammatical rules [11]. George Anton Kiraz and Bernd M’Obius implemented multilingual syllabification using weighted Finite-State Model. Other approaches that deal with syllable structure are Maimaitimin Saimaiti1 and Zhiwei Feng [9], applied a rule-based approach that uses the Principle of Maximum Onset for Urgur language. Ruvan Weerasinghe, Asanka Wasala, and Kumudu Gamage [11] described rule based syllabification algorithm for Sinhala after study the syllable structure and linguistic rules for syllabification of Sinhala words. Yousif A. El-Imam and Zuraidah Md. Don[12] proposed algorithm for syllabification based on Maximum Onset principle for text to speech system for Arabic and Malay. Another example are that algorithm was proposed by Nur-Hana Samsudin and Tang Enya Kong [13, 19]. The system used four syllable structure (CV), (VC), (CVC) (V) cluster with a few sub-models proposed for exception such as loan words pronunciation such as Bali Ranaivo-Malançon [14, 15].

Bali Ranaivo-Malançon proposed a structure of Malay pronunciation dictionary and set of phonological rules for Malay language.

The methods in essence can be divided into three broad techniques: Dictionary based, Rules based and Corpus (or statistical) based. Those three approaches have their advantages and disadvantages. The most simple approach to empty space is usually use in phonetic transcription to separate syllable. Carnegie Mellon University produced Pronunciation Dictionaries for North American English contains over 125,000 words [9]. Pronunciation dictionaries provide accurate syllabification but result in tedious process to add for new words. Rule-based syllabification usually apply universal principles, e.g.: sonority sequencing principle & Maximal Onset principle or by using template-matching and resyllabification [10]. Statistical approaches, on the other hand use machine learning technique. For example, supervised and unsupervised learning to predict the probability distribution over parameter set. Statistical techniques require a large quantity of training data before the prediction give a meaningful result.

In this paper, we use rule-based approach for syllabification Malay since there are no available training data. In section 2, we will study the morphology on phonology of Malay words. Followed by section 3, is an explanation on how we do a collection of Malay wordlist and corpus. The elaboration of the syllabification algorithm will be discussed in section 4. Section 5 is on how we evaluate and observe our algorithm. Finally we conclude our work in section 6.

2 Malay Word
Malay language or bahasa Melayu is an official language of Malaysia, Brunei and Singapore. Malay belongs to the Astronesian (Malayo-Polynesian) family group. Together with Indonesian language (Indonesian form of Malay), the Malay language is spoken by over 300 million people worldwide. Malay has two types of writing system, ie. Rumi (Latin alphabet) and Jawi (Arabic Script). However, Malay is normally written in Rumi. This paper discuss on how to develop rule based syllabification algorithm that based on phonological and morphological.

2.1 Malay Phonology
The sound system of Malay language is pretty simple. The phonemes of most words can be determined from the grapheme. Malay language have 25 consonants:

/b/,/d/,/dz/,/f/,/g/,/h/,/j/,/k/,/l/,/m/,/n/,/ŋ/,/p/,/r/,/s/,/ʃ/,/t/,/tʃ/,/w/, /x/, /z/, /ʔ/.

6 vowels
and 4 diphthongs /[/au],[ai],[oi],[ua]/.

2.2 Malay Morphology
Unlike many languages, Malay is agglutinative language meaning that new word are generally formed by adding prefix, suffix, circumfix, and infix to root word as shown in Fig.2.

![Fig.2: Malay Affix Word Structure](image)

2.3 Malay Syllable Structure
According to onset rhyme model the general structure of Malay Syllable consist of Onset, Rhyme, Peak and Coda. This syllable structures are displayed as a tree diagram. For example word tenteram is presented in Fig.3.

![Fig.3: Malay Syllable Structure for Word tenteram (peace)](image)

In the syllable of Malay’s word tenteram (peace), the nucleus is e, the onset is t, the coda is n, and the rhyme is en. This syllable can be abstracted as a consonant-vowel-consonant syllable, abbreviated CVC. There are several syllable structure in Malay language such as CV, VC, CVC, CCVC, CVCC, CCCV and CCCVC, which C stand for consonant and V for vowel. Table 1 shows the 8 syllables or more possible combination for Malay syllable structure.

![Table 1](image)

<table>
<thead>
<tr>
<th>Syllable Structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Ayam (chicken)</td>
</tr>
<tr>
<td>VC</td>
<td>Anda (you)</td>
</tr>
<tr>
<td>CV</td>
<td>Batu (stone)</td>
</tr>
<tr>
<td>CVC</td>
<td>Hantu (ghost)</td>
</tr>
<tr>
<td>CVCC</td>
<td>Insurans (infrastructure)</td>
</tr>
<tr>
<td>CCV</td>
<td>Praktikal (Practical)</td>
</tr>
<tr>
<td>CCVC</td>
<td>Strategi (Strategy)</td>
</tr>
<tr>
<td>CCCVC</td>
<td>Struktur (Structure)</td>
</tr>
</tbody>
</table>

The pattern of Malay syllable is straightforward and generated by combining syllable structure above. We use this seven syllable structure as our base to generate rules for our template matching syllable structure. From the table 1, the maximum example of Malay syllable structure is word of struktur (structure).

MSS=(C)³V(C)

In Malay language, the actual spoken syllables are the basis of syllabification in grapheme too. Our syllabification are based on etymological (ie. morphological) instead of phonetic principle. The following table, Table 2 present an example of Malay syllabification for a complete sentence ‘Saya suka makan nasi ayam yang sedap’.

![Table 2](image)

<table>
<thead>
<tr>
<th>Syllable Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saya suka makan nasi ayam yang sedap.</td>
</tr>
<tr>
<td>Sa-ya # su-ka # ma-kan # na-si # a-yam # yang # se-dap</td>
</tr>
<tr>
<td>ev.ev # ev-ev # ev-eve # ev-eu # v-cve # evev # ev-eve</td>
</tr>
</tbody>
</table>

I like to eat chicken rich which are delicious

where - for syllable boundaries and # for word boundaries

The frequency of Malay syllable structure is given
in Fig.4 that is obtained from the Malay text prepared in a research of this domain.

Most syllable patterns in Malay are disyllabic. For example makan (eat) → CV+CV. Trisyllabic also common in Malay word for example siapa (who) → CV+V+CV. Only 5% have syllable number equal or greater than four. For example matahari (sun) → CV+CV+CV+CV. The longest syllable structure in Malay is twelve for word ketidakberperikemanusiaan (not having humanitarian feelings) → CV+CV+CV+CVC+CV+CV+CV+CV+CV+CV+V+VC. We need a well define syllable rules for Malay to provide correct syllabification of Malay word. The rule is defines at syllable level and word level.

3. Text Scopus Acquisitions

In this study we used Kamus Dewan (Malay for The Institute Dictionary) and BERNAMA (Malaysian National News Agency) corpus to evaluate our parser. Kamus Dewan (http://prpm.dbp.gov.my) is a Malay language dictionary published by Dewan Bahasa dan Pustaka and is the most authoritative Malay-to-Malay dictionary. We crawling website to collect text from Malaysian National News Agency, BERNAMA (http://www.bernama.com) which provides Malaysian news including politic, economy, sort from newspaper article. A third data (hereafter Overlap) collection was derived from both data collection. We presented different types of Malay words in Table 3.

<table>
<thead>
<tr>
<th>No. of files</th>
<th>Types</th>
<th>Source</th>
<th>No. of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>4155</td>
<td>Online Newspaper</td>
<td>Bernama</td>
<td>18,631</td>
</tr>
<tr>
<td>1</td>
<td>Electronic Dictionary</td>
<td>Kamus Dewan Edisi ke-4</td>
<td>67,233</td>
</tr>
<tr>
<td>1</td>
<td>Overlap</td>
<td>Combination of both source</td>
<td>79,795</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>165,659</strong></td>
</tr>
</tbody>
</table>

3.1 Syllable Dictionary

In addition of data collection, we require a dictionary with the syllable unit transcription. A word is manually segmented to each of syllable. Syllable dictionary will be used as our baseline syllabification.

4 Syllabification Process

A syllable parser is a program that takes an input text and cut it into segments called syllables. Syllabification is language-dependent where each language has its own syllables structure. Our approach is divided into four different phases. Firstly, the parser will read the text document, then it will normalize the text and remove punctuation. The normalize word is then converted into 1 for Vowels and 0 for Consonants. Then, it will try to match the converted CVC text with the syllable pattern rules.

In our syllable parser, the users are asked to extract the syllable and align the syllable. For example: word pandai (smart) will converted 010011 and align to 1,3. (010011 → 1,3) and store into syllable rules so that any word that have similar structure as the word pandai (smart) will automatically extracted using (010011 → 1,3) rules. Finally, the syllable parser prints the output of text document into syllable dictionary format. The process of syllabification is shown in Fig.5.
4.1 Syllable Rules Format

In our formulation of a syllable rules description to automatically extract rules that improve syllabification accuracy, the user are ask to align syllable CVC structure to its syllable variable length. This system will try to induce correct rule for syllable rules matching. The accuracy of syllabification is depends on the number of the syllable rules. We collect 188 of constructed syllable rules in this experiment. Table 4 shows the first 10 entries of syllable rules format to match to the correct syllable pattern.

<table>
<thead>
<tr>
<th>Syllable Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>101 → 1, 2</td>
</tr>
<tr>
<td>010 → 3</td>
</tr>
<tr>
<td>100 → 1, 1, 1</td>
</tr>
<tr>
<td>0110 → 2, 2</td>
</tr>
<tr>
<td>1010 → 1, 3</td>
</tr>
<tr>
<td>0101 → 2, 2</td>
</tr>
<tr>
<td>0010 → 4</td>
</tr>
<tr>
<td>0100 → 4</td>
</tr>
<tr>
<td>110101 → 1, 3</td>
</tr>
<tr>
<td>01001 → 3, 2</td>
</tr>
</tbody>
</table>

4.2 Syllable Rules Matching

Syllable rules matching process is part of syllabification. The following is the pseudo code for syllable rules matching.

```plaintext
sub syllabification_longgest_match {
    Load the set of syllables rules from syllable-file
    Tokenized sentences into word separated by space.
    while (char space is found) do for i = 1 to length of word
        for-each letter
            if consonant letter convert into 1 else convert into 0
            if not found in syllabification matching rules.. add
            else
                match syllable from rules files end if
            End if
        End for
    End while
    Print syllabified string
    End for
}

Sub Syllabification_template_match {
    Load the set of syllables rules from syllable-file
    Tokenized sentences into word separated by space.
    while (char space is found) do for i = 1 to length of word
        for-each letter
            if consonant letter convert into 1 else convert into 0
            if not found in syllabification matching rules.. add
            else
                match syllable from rules files end if
            End if
        End for
    End while
    Print syllabified string
    End for
}
```

Finally, after matching the syllable rules pattern, the correct and incorrect syllabifications of the words used in our prototype and produce an output as a plain text files as the following format:

- akad a-kad
- akademi a-ka-de-mi
- akah a-kah
- akaid a-ka-id
- akak a-kak
- akal a-kal
- akalkan a-kal-kan
- akan a-kan
- akang a-kang
- akanlah a-kan-lah
5 Evaluation and Observation

There is no validated reference for Malay syllabification algorithm in the light of maximal onset theory. Our experiment used the longest matching as a benchmark of evaluation approach to determine the correctness of an algorithm. Evaluation of syllabification algorithm concerns with the correctness of algorithm. We implement our experiment in two independent phases. First phase of experiment, we compared three approaches with the special collected syllable dictionary as discussed in section 3. In the 2nd phase, we analyzed an error on how our algorithm perform on larger dataset (Dewan bahasa, Bernama and the overlap data).

5.1 Comparison with a Syllable Dictionary

We proposed a method of automatic syllabification of Malay word by a comparison of different syllable extraction of each word in seven types of syllable dictionaries. This syllable dictionary contain selected word collection such as Vowel, Diphthong, Vowel Combinations, Consonant, Cluster Consonant, Derive Words, Loan Words and Schwa rules. Fig.7 shows the recognition rate using syllable dictionary for longest matching and syllable rules matching approaches.

![Fig.7: Recognition Rate using Syllable Dictionary](image)

The results of syllabification based on syllable rules matching and longest matching are shown in Table 4.

<table>
<thead>
<tr>
<th>Syllable Dictionary</th>
<th>Longest Matching</th>
<th>Syllable Rules Matching</th>
<th>Improvement Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>90.5</td>
<td>87.2</td>
<td>-3.3</td>
</tr>
<tr>
<td>Diphthong</td>
<td>99.7</td>
<td>76.9</td>
<td>-22.8</td>
</tr>
<tr>
<td>Vowel Combination</td>
<td>80.0</td>
<td>89.6</td>
<td>9.6</td>
</tr>
<tr>
<td>Consonant</td>
<td>90.2</td>
<td>94.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Cluster Consonant</td>
<td>95.3</td>
<td>62.5</td>
<td>-32.8</td>
</tr>
<tr>
<td>Derive Words</td>
<td>100.00</td>
<td>100.00</td>
<td>0</td>
</tr>
<tr>
<td>Swha Rules</td>
<td>100.00</td>
<td>100.00</td>
<td>0</td>
</tr>
</tbody>
</table>

There is an improvement on vowel combination and consonant syllable dictionary by using syllable rules matching approach, which are 9.6% and 4.6% respectively. Other syllables dictionary show decrement and produce the same percentage.

Further analysis is then conducted on the result of syllable rules matching to examine the influence of larger and random syllable of this technique. This dataset contain larger data and random wordlist collected from web. The data is manually collected from online newspaper, electronic dictionary and the overlap of these two sources. We have been evaluated and observed on how the data perform in data set. The result of this analysis is listed in Table 5.

| Syllabification Results (Percentage Correct) based on Kamus Dewan and BERNAMA |
|-------------------------|---------------------|-----------------|----------------|----------------|
|                        | Syllable Rules Matching | Recognize Word | Accuracy |
|                        |                        |                 |             |
| Bernama (18,631)       | 11,303                | 74.96           |
| Kamus Dewan (67,233)   | 52,020                | 76.68           |
| Overlap (79,795)       | 57,155                | 75.22           |

6 Conclusion

In this paper, a syllable rules matching for syllabification process for Malay language was presented. Even though there are there are not many
information regarding evaluation of syllabification scheme for Malay language, but in this experiment we found out that syllabification algorithm have been used for other languages. We decided to implemented the syllabification algorithm towards Malay language and the result shows that Malay language improved the recognition rate on vowel consonant and consonant syllable dictionary. It is also produced above 70% of recognition word.

For the further research, we will look at the elicitation approach in the improvement of syllabification process and the intelligibility of pronunciation for Malay language.

References
[16] (PALMA 2005), Kuala Lumpur, Malaysia, 2005