12-2005

A Grapheme to Phoneme Converter for Standard Malay

Haizhou LI
Institute of Infocomm Research

Mahani Aljunied
Institute of Infocomm Research

Boon Seong Teoh
Singapore Management University, bsteoh@smu.edu.sg

A Grapheme to Phoneme Converter for Standard Malay: A Rule-Based Approach

Li Haizhou[1] hli@i2r.a-star.edu.sg
Mahani Aljunied[1] vismas@i2r.a-star.edu.sg
Teoh Boon Seong[2] bsteoh@smu.edu.sg
[1]Media Division (Human Centric)
Speech & Dialogue Processing Lab
Institute of Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
[2]Lee Kong Chian School of Business
Singapore Management University
#4076, 50 Stamford Road
Singapore 178899

Keywords: Speech recognition, Speech synthesis, Grapheme-to-Phoneme, Malay language

Abstract

This paper describes the process of creating a grapheme-to-phoneme (G2P) converter for Standard Malay (SM). A fundamental step to building TTS and ASR engines, is to build a good G2P system that can automatically generate accurate phonemic representations for words. Our goal is to generate phonemes that reflect real speech, thereby facilitating more accurate phoneme alignment with actual wave-forms (obtained from voice-data collection), keeping human intervention to the minimum. Here we discuss the key areas in SM that require considerable phonemic alterations including letter elisions, consonant insertions, multiple ways of uttering a letter/diagraph – areas that any good G2P system for SM should address. The application of these rules to two sets of corpus will also be discussed, and their generated phonemes examined for both accuracy measurement as well as for further rule refinements.

1. Introduction

In the area of speech technology, namely text-to-speech (TTS) and automatic speech recognition (ASR), an important concern is the ability to align speech waveforms, obtained from voice data collection, to the phonemic representations in a particular language as accurately as possible. These sound wave forms, or spectograms gathered, would require phonemic representation before any TTS or ASR language model can be built. There are several ways to derive the phonemic or phonetic representations for words or utterances in a language:

One is to utilize a machine readable phonetic dictionary that contains phonemic representation for all (usually root) words in that language, as well as other linguistic information that affect the pronunciation of the words in some way – like part of speech, stress information (for example, the verb *RE*bel vs. re*BEL*). These pronunciation dictionaries may be available commercially, or shared by linguistic resource institutions and individuals. Some systems complement their lexicon with morphophonemic rules that handle the more subtle aspects of pronunciation of words like assimilation, and consonant cluster simplification – both of which are relevant to Malay.

Another way to derive phonemic representation for words is to build a grapheme-to-phoneme (G2P) converter that generates phonemes from words (in the form of text) automatically. These systems 'translate' orthographic word forms into their phonemic representations. Most G2P programs can be classified into 2 types – a statistically-based and a rule-based one. The former relies almost exclusively on training and 'recognizing' phonemes from word-orthography based on a sufficiently large amount of corpus or training data. In this approach, a key component is the need for a large corpora.

Rule-based G2P rely on existing linguistic and phonological knowledge, or linguistic generalisations based on analytic observation of actual speech. Using this body of knowledge, rules are written that convert the orthography of a word into its phonemic form. The set of basic sounds used in a particular language will have to be identified, and each of these sound units would then have its unique phoneme representation (i.e. one phoneme symbol for one sound). These rules are typically ordered according to the phonological rules of that language. This approach works better with languages that have more regular, less exceptional relationship between its letters and sounds, for example Spanish or Finnish. The less predictable the relationship between the graphemes and phonemes in that language is, the more difficult it is to write these rules, and this approach may not be suitable for languages that have a relatively deep orthography.

With regards to SM, which is the language of concern in this research, we find a rule-based approach a rather attractive option. Its relatively shallow orthography, compared to a language like English, makes it easier for G2P rules to be written.

In the next section we briefly discuss the

suitability of this apporach for Malay, and following that, in section 3, we highlight the key issues in SM that need most attention when creating any G2P converter for SM. Section 4 describes the bulk of our research, from defining the SM phoneme set, to a description of the linguistic rules that were written, as well as examining the issues not handled by much of prior research in this area. A discussion of our methodology and results can be found in section 5 and a concluding section follows that.

2. Suitability for Standard Malay

A comprehensive phonetic dictionary for Standard Malay (SM) is not available, so obtaining one is not possible, and creating one would require considerable time and monetary resources. Such a dictionary also entails having a closed set of words for the system which isn't very conducive to the rather 'open' nature of SM where many non-native or borrowed words – with orthography adapted (e.g. "aerobik" meaning *aerobics*) or unadapated (e.g. "Bill") –exist and will continue to enter SM (not unlike in Thai). If used, such dictionaries will have to be tremendously huge, and thereby extremely laborious to maintain, not to mention computationally intensive. As mentioned above, some dictionary-based systems include a number of rules to handle unknown words, or words which are not already in their lexicon. Some researchers have noted that a G2P system "has to handle more then it does today so that one can get rid of the huge lexicons. There has to be more rules implemented" (Lindh, 2001). It has been noted that there is a potentially infinite class of personal, company and product names to be spoken correctly, and this is difficult for any TTS system (Henton, 2003).

The statistical approach has been adopted for SM some researchers (Tan 2004, and El-Imam 2000) with varying results. These studies also include some linguistic rules to complement this method which positively affected the accuracy of their systems.

We have pursued a rule-based G2P for this task bearing in mind the above considerations, and also due to our understanding of the fairly direct relationship between the orthography of SM and its sound structure. However, like other languages, there are of course exceptions to the way in which many words are pronounced. We had anticipated to build a dictionary that contains the phonemic transcriptions of these atypically pronounced words.

3. Key Linguistic Issues for any SM G2P

A main goal while building the G2P system is to try to match actual speakers' pronunciation as closely as possible. SM is the variety more widely used in non-academic contexts, in Singapore, Malaysia and Brunei. It is also the variety that people will use when making inquiries or giving information over the phone or in person. This was the motive for the selection of SM as a pronunciation model. We also believe that a good system should, from the onset, incorporate as much of real speech elements as possible. So in our rule-writing process, we incorporated rules based on prior Malay phonological work, did our own study of the lexicon and corpus available to us, as well as paying attention to the way speakers actually speak and the items that need to be spoken about. Having our ears systematically on the ground, is a kind of "analytic listening" (Dutoit, 1997). So in designing the rules for our G2P, we tried to make this our target, such that when voice data is gathered, wave-forms to generated phonemes alignment errors would be minimised.

Based on our existing linguistic knowledge regarding SM, as well as our own observations about the way SM is spoken by its speakers, we have identified several areas which are relevant to capture this variety. Most of these issues would need to and have been handled in other research work, but we find it necessary to incorporate some other regular, but often neglected, features present in local speech.

3.1 SM vs *Baku* Malay

First, it is useful to distinguish the variety of Malay we are concerned with (SM) from Baku Malay (BM). As Malay is used not just in one country, the Malay-speaking nations (Malaysia, Indonesia, Brunei, Singapore) found it beneficial to come to some kind of consensus about terms and spelling so that linguistic and literary resources can be shared more easily (Asmah, 1989). This eventually led to the formation of a language council that presides over standardisation issues. So BM is actually a variety created to facilitate inter-country Malay communication. It is also the variety used in schools in Singapore, and it can be more easily understood by Indonesians. But out of the classrooms, in homes and with friends, this variety is not used. Even in Malaysia, the media does not use the BM variety, but instead uses SM which is more akin to the Riau-Johor dialect of Malay. Malay native speakers in Singapore also use this variety. BM in general, is even more phonetic than SM, as pronunciation is less far off from spelling. For instance, final vowel lowering is not an issue

in BM, but a main one in SM. Glide insertions matter for both, as do glottal stop insertions. BM also does not involve final consonant deletions, while SM does.

3.2 Generating Accurate Phonemes for SM Words: What Needs to be Done

The following are some of the main areas that need to be dealt with while writing G2P rules for SM:

3.2.1 Transition from One Vowel to another- Diphthongs, Glides and Glottal Stops

There are sequences of vowel letters like "ai" "au" and "oi" which are pronounced as single diphthongs, represented as diphthongs /ai/ /au/ and /oi/ to use our transcription conventions, in the same syllable. Any description of SM will need to include these. However "oi" is sometimes pronounced as /oy/ the initiating vowel position being a low-mid back vowel, as opposed to /oi/ which starts off from a low-mid central tongue position. /oy/ is usually found in words borrowed from English.

Two glide consonants, the voiced palatal and velar approximants (/y/ and /w/ respectively, see Table 1 below) assist in the transition in articulation across abutting vowels, or vowels across syllable boundaries in SM. /y/ needs to be generated when the vowel sequence begins with a high front vowel, moving on to a vowel of another position. For example “liat” would need to have another phonemic segment not indicated in the orthography of the word /l ii y aa t>/, making the word bi-syllabic. This consonant is also articulated in the sequence between “a” and “i” like in the word “permainan”, /p er r m aa y ii n aa n/. Vowel combinations that begin with the high back vowel /uu/ followed by a vowel of another position require the insertion of /w/. So does the abutting sequence “au”, found in words like “paut”. The approximant /w/ is articulated in abutting vowel contexts that begin from a high back vowel, to another vowel position, as in "tua" (meaning *old)* transcribed as /t uu w aa/.

The irregularity is that in SM, the sequence /ii aa/ like in the word "liat" above, isn't always articulated with a glide in between. When a prefix ending with a vowel (say, "se-" "ke-", “berke-“ or "di-") is attached to a word beginning with another vowel, like in the word "diambil" (meaning *was taken*, transcribed /d ii ? aa m b ei l/) or "seindah" (meaning *as beautiful as* transcribed /s er ? ii n d aa/), a glottal stop will be uttered instead of a glide. For abutting vowels across suffix boundaries, it is the two glides that are usually used across any combination of stem-final and suffix-initial vowels (namely suffixes "-an" and "-i"), and the only one context where the glottal stop can be heard at the suffix boundary is when the two abutting vowels are of the same quality (like in “kehampaan” /k er h aa m p aa ? aa n/ meaning *disappointment*) , and when the first is a non-high, vowel, followed by a high-front vowel like in the word "mencintai" (meaning *to love,* transcribed /m er n ch ii n t aa ? ii/)

So a good G2P system must be able to automatically generate these sound units that actually do not have orthographical clues to their presence, as well as distinguish them from diphthongs which require different phoneme generation.

3. 2. 2. Unreleased Plosives and Final Stop Devoicing

A less complicated area is the need to generate unreleased stops in word-final and in abutting consonant clusters in SM words. This is similar to the English case of released/unreleased plosives. An added issue with plosives for SM is that word-final ones are unvoiced, so word like "sebab" gets the transcription /s er b aa p>/.

3.2.3. Final Vowel Reductions

An obvious difference between SM and *Baku* Malay is that in the former, many instances of "a"' are pronounced as the schwa in stem-final open syllable positions, like "masa" (*time*) as /m aa s er/. However not all instances of "a" in this context would be reduced. There are many words in SM which do not display this pattern, like "wanita" (meaning *women*, transcribed as /w aa n ii t aa/) without the final schwa. It is often described in linguistic descriptions of Malay that the vowel lowering of “a” doesn’t apply to Malay words of certain origins say, Sanskrit. A good system should be able to predict when the "a" will be reduced and when it would not, requiring a more detailed study of the linguistic data.

In fact, if we look at the inventory of sounds for "a" in actual spoken SM, there are at least 3 sound correspondences to the letter. The are the phonemes /aa/, /er/ (schwa) and /ae/, the low-front vowel which is a borrowed sound significant enough to be noticed.

It must also be noted that upon observation of actual speech, SM speakers maintain the vowel reduction in both stem-final position affixed and unaffixed words (“ketiadaan”, meaning *absence,* /k er t ii y aa d er ? aa n/, where "ke…an" is a circumfix for the stem, and the root-word, "tiada" /t ii y aa d er/). This is not merely a simple case of word final-"a" vowel lowering. So

as much as possible, these variations must be dealt with. Prior work on Malay speech tended to take this into account but do not handle these 'exceptional' words that do not undergo vowel lowering.

Just as important, is another context where SM vowels gets lowered is stem-final, closed syllables that contain vowel letters "u" and "i" in nucleus positions. Unlike in other contexts, "u" is articulated as /oh/ while the letter "i" as /ei/ in words like "batik" is /b aa t ei ?/ (instead of /b aa t ii ?/), "pantun" is /p aa n t oh n/ (instead of /p aa n t uu n/), "tapis" is /t aa p ei s/, (instead of /t aa p ii s/). The closing consonants in these syllables are elided when the stems end with "h" and "r". So we get /l er b ei/ and /t aa r oh/, instead of /l er b ii h/ and /t aa r uu h/, for the words "lebih" and "taruh". With other closing consonants, they are pronounced. Based on the speech of five SM native speakers' readings of sentences containing words of this nature, as well as listening out for the use of SM in the media, we find this "u" and "i" lowering rather prevalent and necessary for us to generate the right phonemic representations for this pronunciation.

3.2.4. Multiple Pronunciations of "e"

Like the letter "a", "e" has several corresponding sounds as well. In most instances it is pronounced as schwa, and in other instances as the mid-front vowel /ei/. This results in a handful of homographs like "bela" (pronounced /b er l aa/ meaning *to rear*, or /b ei l aa/, meaning *defend*) which poses a problem for speech engines, but these are "few and far between" (Asmah 1989). Besides these homographic words, there are also words that use the *e-taling* and any G2P system should attempt to handle this significant set of words like "Bedah" (/b ei d aa h/), a name, and "perak" (/p ei r aa ?/) both having the pure vowel /ei/ represented by the orthograph "e". In a word like "geletek", both e-types exist /g er l ei t ei ?/. Deciding how to pronounce the Malay "e" is one of the causes of pronunciation errors among new second language learners of SM.

Another issue that needs to be included is the glottal stop presence in word-initial vowel segments. Also the alternation of "k" between the velar plosive /k/ and the glottal stop when word finally as in "tidak" /t ii d aa ?/ (meaning *not*). Again here, not all stem-final "k"'s get pronounced as /?/. Many remain as the unreleasd plosive /k>/.

The gemination of "k" at suffix boundaries in words like "kedudukan" /k er d uu d oh ? k aa n/ (meaning *position*), root-word "duduk" is also another tricky area, since one grapheme "k" in this context needs a generation of two phonemes. These are common enough to be noticed in SM speech and handling them was considered necessary.

4. Resolving these Issues: Via Rule-Based G2P

4.1. Identifying Phoneme Sets

To generate phonemes from letters, or combinations of letters, we first have to define what the letters are. SM is written in the roman script or *Rumi,* with 26 letters of the alphabet, and some non-alphabet characters used in SM words, namely the hyphen, found in reduplicated words, and the apostrophe found in some Arabic terms and names. The roman writing was a British introduction to the region during the colonial days. Prior to that, the Arabic script was used to write Malay. So all the letters of the alphabet can be found in written SM, some more commonly than others. The minimal sound units used in SM are listed in tables 1 and 2 below.

Table 1 reflects the consonants identified. All the consonants listed here are used by Malay speakers, some of which originate from other languages, and have been described as secondary consonants. These are /f/, /v/, /kh/, /q/, /gh/, /th/, /dh/. We have also included /l~/, the dark or palatalised alveolar lateral-approximant as it is found in many Malay names like "Abdullah". Although commonly represented by the letters "ll", not all cases of this sequence of letters gets uttered as /l~/. This palatalisation is evident in Arabic, and we have observed that in some names more than others, this consonant gets articulated consistently. In all, in our consonant set, there are 30 sound units, or phonemes.

There are 8 vowels in Table 2 that lists the

Place & Manner of Articulation	Bilabial	Labio-dental	Dental	Alveolar	Post-Alveolar	Palato-Alveolar	Palatal	Velar	Uvular	Glottal
Oral Stop	/p/ /b/			/t/ /d/				/k/ /g/	/q/	/?/
Nasal (stop)	/m/			/n/			/ny/	/ng/		
Affricate						/ch/ /jh/	/dy/			
Fricative		/f/ /v/	/th/ /dh/	/s/ /z/		/sy/			/kh/ /gh/	
Lateral				/l/		/l~/				
Approximant					/r/		/y/	/w/	/h/	

Table 1: Consonants of SM (For cells with 2 phonemes, left phonemes are voiceless, the right ones, voiced)

inventory of vowel sounds used in SM. 6 of them are basic Malay vowels. We have included 2 more secondary vowel phonemes, which are spelt using

the same set of vowel letters, to cater for borrowed words. The are the low front monophthong (or pure vowel) /ae/, and the low-mid, back vowel /or/. Both are found mostly in borrowed English words like "faks" /f ae k> s/ (*fax*), "aerobik". /ae r oh b ii k>/ (*aerobics*), and "blok", pronounced /b l or k>/, (*block*). Company names and foreign names abound with these 2 vowel sounds as well. The letters representing /or/ is typically "o", while /ae/ is represented by "a", "e" and the digraph "ae".

Tongue position	Front	Central	Back
High or closed	ii		
High-mid or half-closed	ei	er	
Low or open	ae	aa	or
High or closed			uu
High-mid or half closed			oh

Table 2: Vowel System for SM

We have identified 4 diphthongs in SM, all of them rising (or closing) diphthongs as they end with high vowels. The 3 basic ones are : 1) /ai/ - as in "lambai"(pronounced /l aa m b ai/, meaning *wave*). 2) /au/ - as in "kalau", (pronounced /k aa l au/, meaning the conjunction *if*), and 3) /oi/ - as in "baloi" (pronounced /b aa l oi/ meaning *fitting*). The additional diphthong is /oy/ found mainly in words borrowed from English, like "boikot" (meaning *boycott*, pronounced /b oy k or t/). The orthographic representations for these diphthongs in SM words are more consistent than the other vowels, namely "ai", "au" and "oi", although the last digraph "oi" can be ambiguous between /oi/ and /oy/.

The phonemic names, or symbols decided upon and shown in the tables above were selected on the basis of their similarity to the sounds as well as the grapheme that often represents that sound. For example, we have /ei/ and /er/ phonemes which are often written by the letter "e", and also /sy/ for the voiceless, palato-alveolar fricative, that is always indicated by the letters "sy" in Malay words. The same reasoning applies to the selection of /y/ as the phoneme for the grapheme "y" (as opposed to the more commonly used IPA option /j/). We also tried to reflect similarities in the phoneme symbols we used, to capture voiced and voiceless consonant pairs (like /ch/ and /jh/, and /th/ and /dh/, both pairs having "h" indicative of their likeness). Not all of these sounds are pronounced by every speaker, especially for the very Arabic sounds like /dy/ and /gh, as we have observed. Often these consonants are adapted or simplified, and pronounced like other Malay sounds. For instance "maghrib" is often pronounced as /m aa g r ei p>/ rather than /m aa gh r ei p>/, with the letter sequence "gh" realised as a voiced stop, instead of the fricative. The same with the name "Ghazali" pronounced commonly as /g aa z aa l ii/, while "redha" is commonly said as /r ei d aa/, rather than /r ei dy aa/). This supports the view that Malay has fewer allophonic variants than Arabic (El-Imam 2000).

4.2. Letter-to-Phoneme Mapping

Figure 1 below illustrates the various stages of the G2P converter design. The first stage is a lookup table where a dictionary containing the list of exceptionally-pronounced words are stored together with their phonemic representation. If an input word is listed in this dictionary, no rules will be applied to that word, and its hard-coded phonemes in there is then used to generate the G2P output for that word. The size of this dictionary is kept as small as possible.

Input words not in the exceptions dictionary will go through the full conversion stages. First a mapping module that maps letter(s) into phonemes. Since each letter or group of letters may correspond to multiple phonemes, this module first maps them into a default phoneme. So the letter "a" is mapped into /aa/, without looking at its word context, "au" mapped into the diphthong /au/, "sy" to /sy/, and so on.

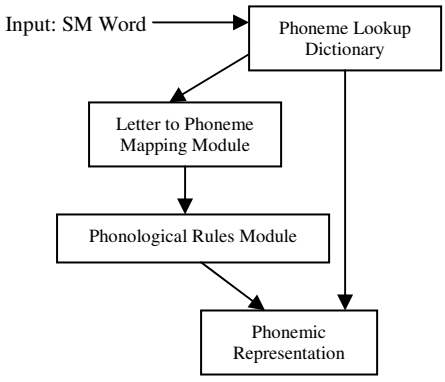

Figure 1: Flow Chart of SM Word from Graphemic to Phonemic Representation

4.3. Phoneme-to-Phoneme Substitution Rules

After each letter and *digraph* (2 letters indicating one basic sound unit) is converted to its default phonemic value, the phonemes then go through a phonological rule module where phoneme-to-phoneme substitution rules are applied. This is to 'correct' or refine the default phonemes obtained from the context-blind mapping stage. Each rule

consists of a grapheme or phoneme *pattern* to be matched against the input. The pattern searched for is defined in the *conditions* where the context of the occurrence of pattern A -- phoneme values, orthography, adjacent elements, and word boundaries -- can be specified. The *substitution* part of these rules then replaces, or reassigns new phonemes to the relevant sound segments in the pattern, and we can also assign any other feature assignments that can help in the application or suppression of other rules. These rules are carefully ordered so as to generate the correct phonemes at the end of the G2P process. Context-sensitive rules of this nature are somewhat in line with the tradition of generative phonology (Chomsky and Halle, 1968).

These rules can be categorised and ordered as follows, and sections 4.3.1-7 discuss what the rules do in terms of generating or refining the right phoneme for the graphemes in the words:

1. Non-native word identification
2. Glide-insertion rules (valid diphthongs not split up remain as diphthongs)
3. Glottal stop insertion
4. Devoicing and unreleasing of plosives
5. E-pepet/taling determination
6. Final “a” lowering
7. Final closed-syllable “i” and “u” nucleus lowering
8. Final consonant “h” and “r” deletion

4.3.1. Non-Native Word Identification

As mentioned in the previous section, one of the main problems observed in accurate phoneme generation for SM is the exclusion of the application of certain phonological rules (say final vowel reduction, and glottalisation of stem-final “k”) to non-native SM words. Many linguistic studies that encounter atypical treatment of non-native words tend to dismiss these words as exceptional, and do not attempt to pursue the discussion. From the onset of our research, we have based the first set of rules on a 50,000-word lexicon (see section 5 on Methodology and Results).

Some of these rules warrant more discussion. One of the main dependencies of variation was found to be the unusual, or exceptional way some SM words are being pronounced. As a result of that, we have included some 50 rules that detect potential foreign, or borrowed word patterns. These rules include checking for consonant clusters not following the typical SM structure like oral stops plus liquids /r/ or /l/ in onset combinations that would identify words like "klasik" (/k l aa s ii k>/), "trafik" (/t r aa f ii k>/) dan "drama" (/d r aa m aa/). Three-segment consonant onset clusters are also clear indicators of foreign words like "skrin" and "strategi". The presence of certain letters which are not used in native SM words like "x", "q", and "v", or onset letter combinations with silent letters like "ps" (as in "psikologi") and "pn" (as in "pneumonia"), signals foreign or borrowed words. The position of certain letters in the words also matter, like a stem-final "j" - found in words like "pakej" and "kolej" - is not native SM. Sometimes there are no orthography or phonemic clues to the foreign word, but we have observed from our initial lexicon study that words following certain syllable combinations are excluded from some SM-only phonological rules.

One such rule looks for *Consonant-“a”-Consonant-“i”-Consonant-“a”* type words - where "Consonant" refers to a slot potentially filled by any one or two consonant phonemes. The speculation is that this vowel harmony combination is found mostly in Sanskrit words that do no not need to undergo final /aa/ lowering. Words like "wanita "tadika" and "jantina" fit in to this pattern. Person names like "Farida", "Hasnita" that do not need to undergo vowel lowering, also benefit from these rules.

Another relevant vowel pattern we identified was : *-"o"-Consonant-"i"-Consonant*, where this combination pattern matches the ending portions of words like "telefonis" and "katolik". Consonant clusters in coda positions is also a non-SM characteristic of word formation; thus we wrote a rule that looks for stem-final clusters like /k s/ in words like "antraks" (/aa n t r ae k> s/). We tried to include as many of such patterns as possible found in words in our data sets.

In the rules that capture the above sequences, each of them will assign a "non-native" feature to the word which in turn prevents the application of certain rules that follow this set. Some of the rules that follow check for the condition that the word in question should not contain a non-native feature before applying any phoneme alteration. Besides assigning this feature, the assignment portions of some of these rules replace the default vowel phoneme values in that pattern, with other vowel phonemes. For instance, the vowel /oh/ (the default mapped phoneme for the letter "o") is replaced with /or/ in word patterns like in the above-mentioned *"o"-Consonant1-"i"-Consonant2*, when Consonant1 slot is filled with oral stops resulting in words like "optik" converted to /? or p> t ii k>/, instead of /? oh p> t ii k>/. In another rule, the default /er/ phoneme for the vowel letter "e" in "-eks" stem endings, also gets

replaced with the /ae/ vowel, in resulting in words like "teks" pronounced /t ae k> s/.

4.3.2. Glide and Glottal Stop Insertion Rules

We have earlier introduced the need to insert intervocalic glottal stops and intervocalic glides. The decision to insert either a glide or glottal stop was dependent on whether one of the vowels is part of a prefix or suffix (see section 3.2.1).

We have a total of 15 cross-morpheme boundary rules that look for sequences of letters that are likely to be prefixes that end with a vowel - like “ke-“, “se-“, “di-“, “kese-“, and “berke-“ – followed by another vowel (pure vowel or diphthong), which is assumed to be the first letter of the root word. This will generate correct phonemes for “keseimbangan” (/k er se ? ii m b aa ng aa n/), “diambil” (/d ii ? aa m b ei l/), “seakan” (/s er ? aa k aa n/). In the initial mapping stage, we have assigned diphthong phonemes to these letter sequences: /ai/ to “ai”, /au/ to “au” and /oi/ to “oi”. So our rules have to anticipate the existence of such diphthongs as we write them.

Firstly there are rules that look for a sequence of 3 vowel letters, namely "oia" "aia" "iaa" "aua" "aui". These are unambiguous contexts where the first two letters represent either a diphthong or part of a stem, while the last vowel constitute a suffix or part of a suffix. For pattern oia, we add /y/ between /oi/ and /a/. Some rules also look for potential prefix patterns like "ke-" "se-" "me-" in word-initial positions followed by vowels, as all stem words beginning with a pure vowel or diphthong which are prefixed by "ke-" "se-" "di-" need the glottal stops. Glide insertions are more common and they are assigned in other intervocalic contexts which are not diphthongs. Across vowels of he same orthography, like in "maaf", a glottal stop is also inserted.

4.3.3. Oral Stop Generations

We have written 4 rules to handle this area. Devoicing of stops at word-final positions is fairly easy to generate, as we look for word-end boundary and replace all instances of /b/, /d/ and /g/ to /p/, /t/ and /k/ respectively. However looking at the lexicon available to us, we found that we should expand the context for this alternation as we feel that some other contexts require this replacement as well. When hearing these words – “abstrak”, “dihadkan” “penabsahan” and “Habsyi” – spoken, we find a similar devoicing pattern. So instead of just looking word-finally, a rule was written to apply the same phoneme substitutions when these voiced stops are followed by certain voiceless consonants namely /t/, /s/, /k/, /ch/, /h/, /p/, and /sy/. So a word like “mendarabkan” gets the phonemically represented as /m er n d aa r aa p> k aa n/.

After devoicing, we applied the unreleased stop rule to convert oral stops /p/, /t/, /k/, /b/, /d/, /g/, and /q/ into their unreleased counterparts /p>/, /t>/, /k>/, /b>/, /d>/, /g>/, and /q>/ respectively. The contexts we specified for these application are post-vocalic word-final positions, and before some consonants including other oral and nasal stops, affricates, /sy/ and /s/.

4.3.4. *E-pepet* and *E-taling* Generation

This problem of /er/ versus /ei/ determination is also handled (to a large extent) in our G2P converter. Based on our lexicon study as well as what has been written about the regularity of Malay spelling (Asmah, 1989), there are clues in the orthographic combinations of vowels across syllables that helped us generate the right phonemes. A regularity of written SM, since the spelling reform of 1972, enabled us to identify which “e”’s should be pronounced as /ei/, or the e-taling, instead of the default schwa, also known as the *e-pepet*. We observed that in bi-syllabic root-words that contain two “e”’s in both nucleus positions, exemplified in words “tempel”, “bedek” and “senget”, the e-taling is articulated in both syllables, /t ei m p ei l/, /b ei d ei ?/ and /s ei ng ei t>/ respectively. In words with a sequence of 3 consecutive “e” nucleus slots, the first “e” is typically realised as the schwa while the other two as /ei/. So the word “geletek” gets the phonemically mapped into /g er l ei t ei ?/ after undergoing these rules.

Another pattern is in bi-syllabic words with one “e” nucleus in combination with “o” in the other nucleus. The “e” in “telor” (which was ambiguous without the e-taling diacritic ĕ) is an /ei/, and so are the “e”’s in “solek” (/s oh l ei ?/), “boleh” (/b oh l ei /) and “bengot” (b ei ng oh t>/). Other contexts we observed are stem-final syllables where “e” is closed with consonants “h” and “k”. A total of 6 rules were written handle the generation of e-taling, /ei/.

4.3.5. Final “a” Lowering

This feature of SM is among the more apparent phonological features of this language variety. (7.6%) of the 50,000 words in the lexicon involve at least one application of this rule. Nonetheless, our handling of this issue didn’t merely depend on this rule. It is heavily dependent on prior rules that identify foreign word patterns (see 4.3.1). The actual subsitution rule captures root-words ending with an open syllable “a”, that is potentially followed by suffixes. The only condition we added to this rule is that the sequence in question should not be in a word that has been assigned the non-native feature so as to exclude the /aa/ to /er/

replacement from words that do not fit SM structures.

After studying the G2P output of the second set of corpus, we further refined these rules such that they also apply to words containing the third person possessive suffix "-nya" (as in "makanannya", meaning *food of third person,* to be uttered as /m aa k aa n aa n ny er/). This suffix "-nya" can also co-occur with the suffix "lah" to form "-nyalah" in which case the desired phoneme output for this sequence would be /ny er l aa/, with the vowel lowering occurring a non- word-final position. With this refinement, the word "pertamanyalah" (root: "pertama") would be accurately transcribed as /p er r t aa m er ny er l aa/. There are also rather common words like "setibanya", and "masanya" where this rule will need to apply more than once to the word so that we can accurately generate /s er t ii b er ny er/ (rather than /s er t ii b er ny aa/) and /m aa s er ny er/. Not unlike many other languages, the affixation rules for SM may apply to borrowed or foreign words as well. So even if a word has the non-native tag assigned, it may still need to undergo this rule is it contains the suffix "-nya". In this research, we have not applied this feature. Only with this will a borrowed word like "staminanya" (stem : "stamina", non-native) be accurately generated: /s t aa m ii n aa ny er/.

4.3.6. Final Syllable "u" and "i" Lowering

Another strictly SM feature which is easily distinguishable from other varieties of Malay is the reduction of high vowels /uu/ and /ii/ in stem final, closed syllable contexts. A study of the lexicon and the SM reading samples enabled us to determine the more specific environments where /uu/ needs to be substituted with /oh/, and /ii/ with /ei/. Our rules look for /uu/ and /ii/ nucleus slots which are followed by any of the following closing, or arresting, consonants: /n, /l/, /t/, /s/, /m/, /ng/, and /p/. So endings in words like "pantun" (*poem*) would be phonemically mapped to /p aa n t oh n/, from the pattern /p aa n t uu n/, "sambil" into /s aa m b ei l/ (from the original /s aa m b ii l/), and "harum" into /h aa r oh m/ (from /h aa r uu m/). We have observed among SM speakers that even with affixes, this vowel lowering will still apply to the root words. This results in "penampilan" (root : "tampil", /t aa m p ei l/) generated as /p er n aa m p ei l aa n/, and "caruman" (root: "carum", /ch aa r oh m/) as /ch aa r oh m aa n/, despite the non-word final position of the syllables concerned. This alternation context is the same as the above /aa/ to /er/ variation with suffix.

For closing consonants /r/, /h/, or /k/, there is more than simply a vowel phoneme replacement. Arresting /r/ and /h/ in stem final positions are also unpronounced, and thus dropped. So we delete these final consonants as well in the phonemic transcriptions of words like "patuh" to read /p aa t oh/ (instead of /p aa t uu h/), "lebih" to /l er b ei/ (instead of /l er b ii h/). This deletion is also applied to stem final syllables which contain non-high vowels including in the following sequences: /aa r/ found in /b aa h aa r/, "bahar" converted to /b aa h aa/. Similarly, "lemah" was accurately convereted to /l er m aa/ (from the original sequence /l er m aa h/) after the application of these rules. No lowering is required in these contexts since the nucleus is already filled by the low back vowel /aa/. This /r/ and /h/ deletion however, would not be applied if the word in question contains the suffix "-an" or "i", like "keseluruhan" (root: "seluruh") and "mematuhi" (root: "patuh"). The final /r/ and /h/ gets re-syllabified into the onset of the following suffix. This pronunciation feature is also reflected in our G2P converter.

For all stem-final syllables closed with "k", there has been much discussion in the linguistic community about the /k/ and /?/ alternation. It is generally accepted that Malay words with final "k" arresting consonants would end with the glottal stop /?/. This is true for both stem-final unaffixed and suffixed SM words. So a words like "balik" gets generated as /b aa l ei ?/, (from the original /b aa l ii k>/) with both vowel lowering and glottalisation of "k" implemented. When affixed, like in "membalikkan" (root: "balik"), the same substitution processes apply, giving us /m er m b aa l ei ? k aa n/ (from the unglottalised /m er m b aa l ii k> k aa n/).

There is one other context related to final "k" glottalisation that occur in "-an" and "-i" suffixed words. When "-an" is attached (either in combination with prefix "pe-" or "ke-") to a stem like "balik" (*return*) - we get "pembalikan" (*the return of*) - there is an additional phoneme that needs to be generated to mimic SM speech at the suffix boundaries of these contexts. There is a single "k" letter stem-finally that gets attached as an onset to the following syllable made of the vowel-initial suffix "-an" or "-i". At the same time, a glottal stop is also uttered preceding this /k/. So in the example "pembalikan", we needed to generate /p er m b aa l ei ? k aa n/, with two consonant sounds triggered off by just one "k" and the presence of this suffix. Many words fall into this category: "kebanyakan" (root: "banyak"; meaning *most*), and "menduduki" (root: "duduk"; meaning *to occupy*) were successfully converted to /k er b ny aa ? k aa n/ and /m er n d uu d oh ? k ii/ by our system. This has been been described as a

form of gemination in SM and in fact, a source of a common misspellings among SM speakers. Teachers find "kebanyakan" wrongly spelt as "kebanyakkan", with two "k"s at the suffix boundary. This could be attributed to the presence of both /?/ and /k/ in the pronunciation of such words. A simple web-search for the words "kebanyakkan" and "kedudukkan" will demonstrate how rife this is even in official government web sites and on-line newspapers.

However, once again, there are a lot of instances – significant enough for us to look at the details more closely – where /?/ is not realised from the letter "k". Only native Malay words in SM get this pronunciation, while borrowed words like "antik", "diagnostik", "sulfurik", and "mekanik" do not require it. The final "k" in these words remain as an unreleased velar plosive, /k>/. We have thus added a condition to this entire set of rules as not applicable to words identified as 'non-native' by earlier rules. So words that fit into foreign patterns like "transkrip", "klasik" and "hipokrit", would be excluded from vowel reduction.

Even with this exclusion, we observed while refining the phoneme results of the lexicon and corpus, that there is an over-application of these rules. The next thing we did was to zoom in on the relevant syllables that didn't require this vowel lowering and consonant deletion process, and looked at the onset consonants of these syllables. From this process we gathered that words with certain final syllable onset consonants consistently do not occur, or occur very infrequently with this alternation. This set of onset consonants are thus checked for in the rule conditions, and words that match this pattern do not undergo this vowel subsitution.

4.3.7. Other Rules

We have an additional 20 rules that handle to a certain extent non-SM words, with the intention of handling names of places, people and companies. Most of these unadapted foreign words appear in our final set of data from the newspaper corpus. English names like "Herbie", "Tracy", "Ladd", and "Sidney", Chinese names like "Chee", and even Arabic names like "Shariff" fall into a different phonic structure when compared to Malay. In the final round of rule refinement, we added a few rules in the beginning of our rule file that 'normalise' these patterns. For instance, in SM, the letter sequence "ch" doesn't occur. If a word like "Chan" were to be entered into the system, it would first be mapped to /ch h aa n/ interpreted as a consonant cluster. The second generated phoneme /h/ would need to be suppressed, so we have written rules for such purposes. Another rule handles a sequence of 2 "e"'s substituting it with the phoneme /ii/. In SM words, a sequence of 2 "e"'s typically belong to different syllables (like "seenak" /s er ? ei n aa ?/) , and a glottal stop is the dividing consonant. This is different from English reading rules.

In this section 4, we have tried to describe the way in which we incorporated as much of the word context as possible while SM phonological rules. Each of these sets of rules tried to handle each phenomenon together with as much of its variation as possible. We do not consider handling such variations as trivial because naturalness and accurate phoneme generation are our main concerns.

We would like to add that the sequence or ordering of some of the rules listed in section 4.3 do matter. Most obviously, non-native word detection rules must be applied first before they can exclude many non-SM words from undergoing certain subsitution rules. For vowel lowering rules and consonant deletion rules, it is the presence of the final consonant that triggers off "i" and "u" lowering. So for final syllables with closing /h/, and /r/, they will only be deleted after the vowel reduction happens. If the deletion happens first, the vowel lowering rules will not be applied because of the absence of the final consonants. "i" and "u" do not need to be lowered in final open syllables.

Let's take a look at a word like "pengasih" (*loving person*):

1)Pengasih > /p er ng aa s ii h/ > /p er ng aa s ei h/ > /p er ng aa s ei/
2)Pengasih > /p er ng aa s ii h / > */p er ng aa s ii /

In 1) and 2), the result of initial mapping module gives us :/p er ng aa s ii h/. Later, we had access to information about root-word grapheme (see section 5.1). If the root word ends with "ih", and the word-final phoneme sequence is /ii h/, we lower /ii/ to /ei/. Deletion of /h/ will be applied only after that to generate the correct output in 1). If the /h/ is not present, there will be no lowering, e.g. in the word "kasi" (*give*, /k aa s ii/), and as wrongly generated in 2) above */p er ng aa s ii/. In short the ordering of the of some of the rules does matter, but not for all of them.

5. Methodology and Results

After building the one-letter/digraph to one-sound mapping module (which supplies the default phoneme values for each letter or digraph, see Figure 1), we took a sample of a 5,000 words from

a set of 50,000 word lexicon of SM words. Every tenth word was extracted to make up this 'golden' set of words which was the basis of the first set of phonological rules. Made up of both affixed and bare (root) words, this set A was then run through the mapping module of our G2P system so as to look at the results of simple one-to-one, letter-to-phoneme output. This set consists of only pure Malay words, or adapted words. There are no person or place names in that list, quite like entries in a dictionary. The mapping output was manually studied by our linguists, and the main areas that needed to be handled (some of which were explained in section 3) were then identified.

The first set of SM phonological rules were written based on a study of the simple mapping output. Then once again, set A was run through the mapping module, and then through the first cut of our phonological rules module. The phonemic output of this was again manually verified, and the errors studied. Words which are truly exceptional and cannot be accounted for by any rule or refinement of these rules were placed in an exceptions dictionary ("Phoneme Lookup Dictionary" module in Figure 1) where the SM word and its phonemic representation were manually coded so that these words do not have to undergo the phonological rules. The other errors that could be improved with rule refinement were then grouped together, and existing rules were modified to better handle these errors. New rules were also added if necessary. This verification-feedback loop was been the development procedure for our rule-writing methodology. These rules were then applied to set A and we found 71% of the 5000 words in set A had correct phonetic representations (see Table 3).

Next, these rules were applied to another set of data (set B): a 5,000 word list derived from 4-years' daily editions of a local online Malay newspaper (cyBerita, 2001-2004). Obtained from this 23 million-word corpora, were 5,000 words of the most frequently occurring words in these articles. These newsreports cover a variety of topics including current affairs (domestic and international), recreation and sports, economy and business reports, fashion, religion, editorials and letters to the editors, as well as regular online edition columns. What is different in this second set of data is that there are many more current words relating to technology and politics which are not found in set A. A number of proper names (of people, countries, companies) also appear frequently in these articles. If these names are not SM words, their spellings are maintained and pronunciation-wise not very different from the way they are pronounced in the originating language. Thus we had the set of rules that aim to handle these foreign words. Because of resource constraints, we only have a small set of such rules. In some other systems, the originating language's G2P system are being used to generate the phonemes for these non-native words, particularly name-words, or *toponyms*. We did not implement this to our G2P system.

This set B was run through the G2P rules, and the results manually verified by a linguist. 77% of the words have accurately generated phonemic representation. Once again, the errors were analysed, with regards to our target SM pronunciation model. In this round, we find that little refinement can be made, and that to improve the performance significantly, we needed to accurately identify morpheme boundaries.

5.1. Adding a Malay Morphological Analyser

As described in section 3, and as reflected in our rules, some SM phonemic alternation issues (like vowel lowering, consonant deletion, glottal stop and glide generation) lie in morpheme boundary regions. Up to this point (generating first output for set B), the rules that we have written looked for sequences of phonemes that can be suffixes or affixes. In each of these rules, we tried to 'anticipate' affixes that can occur by specifying the phonemic sequence of that suffix or affix. For instance, the glottal stop insertion rule across prefix boundary, we specified a pattern like /d ii/ or /k er/ at the beginning of a word, when followed by a vowel, generate a glottal stop /?/ in between the vowels. To be safe (i.e. to prevent over-application), we also checked for circumfixes, such that when there is a word-initial /d ii/ sequence present, we also ensured that there was also a closing sequence of either /k aa n/ or /ii/ (we took these to be suffixes "-kan" and "-i". Simpler, and more elegant rules can be written with the inclusion of a morphological analyser for SM. Not having a root-word dictionary available to us, together with other resource constraints, we did not build an affix stripping module that would be able identify true root words and their affixes.

Nevertheless, we were able to use an SM morphological analyser already developed in the Institute of Infocomm Research and incorporated that as a module into our G2P system. A detailed discussion of this system is beyond the scope of this paper. With the inclusion of this morpho-analyser (which has a root-word dictionary of about 15,000 root entries, and a rule-based affix-

stripping rules for Malay), we had access to 2 kinds of information that are relevant to our phonological rules: the actual spelling of the root word, and its list of suffixes and prefixes. The conditions of our rules were then modified to take into account these new and useful information. We could then look for phonemic patterns in words, as well as check that the phonemic substitutions are occurring at the right morpheme boundaries.

We ran both sets of data through the G2P again, this time with the inclusion of the morpho-analyser. The addition of this module resulted in an increase from 71% to 85% of words with correctly generated phonemes for set A. Set B saw an increase of from 77% to 88%. Most of the improvements lie in correctly reduced stem-final vowels, and correctly geminated single "k" at stem-final suffixed words. Looking at the frequency of the application of some rules – namely the /aa/ to /er/ stem final alternation, and the /uu/ to /oh/, and /ii/ to /ei/ in stem-final closed syllables – the number of times these rule applied to the 5000 words in set A increased from 1559 to 2429 times. Without the morphological analyser, we were not able to write /k/ to /? k/ gemination rules (to handle words like "kedudukan"). By being able to identify morpheme boundaries, we then were able to write new and more accurate glottal and glide insertion rules as well.

	Set A (5K dictionary-type words)	***Set B*** (5K of common words from news articles)
Without Morpho	71%	77%
With Morpho	85.4%	88%

Table 3: Percentage of words from corpus with accurately generated phonemes, with morphological analyser and without.

6. Conclusion

In this study we have tried to solve every phonological issue for accurate phoneme generation by the rule-based approach, complemented with a set of exceptions dictionary containing transcriptions of irregularly, pronounced SM words. In this process, 2 main points came into the picture: the need to better handle borrowed/foreign words found in SM texts, and the importance of identifying morpheme-boundaries. We have attempted to incorporate these issues into our G2P system to a certain extent, and the accuracy results show a significant increase with these considerations.

We have not incorporated syllable boundary identification, as we found morpheme boundary contexts more relevant to generating the right phonemes. Even with syllable identification, there is still the need to identify root-words. Another area not incorporated in this study is the duration of vowels, in Malay is non-contrastive although upon receipt of voice data vowel length patterns (as well as other prosodic elements) are likely to affect recognition results. Although we have tried to include foreign word detection rules, and a handful of letter-to-sound rules for non-Malay words, it is still a long way to go before being able to handle the multitude of non-Malay words in Malay speech, particularly in the context of multicultural societies like Singapore and Malaysia. The level of education of the speakers are also expected to influence their pronunciation of many words foreign words, either adapted into Malay or not. Final stop devoicing, for instance, is an interesting area to observe, to find out if in English words, the same phenomenon occurs in SM speakers' speech.

In anticipating actual speech, we expect variety in SM speech rather than rather uniform data. Speaker background attributes especially educational background will play a role in determining the way they speak, particularly with regards to unadapated borrowed words and foreign names. This is not a trivial issue in certain multi-cultural speaker contexts. Voice data collection needs to bear this in mind. Perhaps if such socio-variables is incorporated in some way into the G2P system, a more accurate phoneme generation model can be derived.

Acknowledgements

Our deepest gratitude goes to the Singapore Press Holdings and their information specialists for sharing their cyBerita archives with us in this research. Our thanks also go to the Machine Translation lab in the Institute of Infocomm Research for sharing their invaluable morphological analyser with us.

References

[1]Asmah Haji Omar. "The Malay Spelling Reform" in *Journal of the Simplified Spelling Society*, 1989-2 pp. 9-13.

[2]Boon Seong Teoh: 1994. The Sound System of Malay Revisited. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[3]C. Henton, "TheName Game: Pronunciation Puzzles for TTS" in *Speech Technology Magazine, September/October 2003*, 32-35.

[4] Jonas Lindh: 2001. *Introductory Evaluation of the Swedish RealSpeak System,* Term Paper for Speech Technology 1, Graduate School of Language Technology

[5] N. Chomsky, M. Halle: 1968. *The Sound Pattern of English*, New York: Harper & Row.

[6]Thierry Dutoit: 1997. “High Quality Text-to-Speech Synthesis: An Overview” in *Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis,* vol. 17 (1) pp. 25-37.

[7]Yeow Kee Tan, Boon Seong Teoh, Haizhou Li, “A Grapheme to Phoneme Conversion for Standard Malay”*, International Conference on Speech and Language Technology*, O-COCOSDA 2004, 19 Nov 2004, New Delhi, India.

[8]Yousif A. El-Imam, Zuraidah Mohammed Don, “Text-to-Speech Conversion of Standard Malay”, in *International Journal of Speech Technology 3,* 2000, Kluwer Academic Publishers, pp. 129-146, Netherlands.