Subword Unit Concatenation for Malay Speech Synthesis

Sabrina Tiun[1], Rosni Abdullah[2] and Enyakong Tang[3]

**[1] Pusat pengajian Sains Komputer, Universiti Sains Malaysia
Minden, 11800, Malaysia**

**[2] Pusat Pengajian Sains Komputer, Universiti Sains Malaysia
Minden, 11800, Malaysia**

**[3] Multimedia University,
Cyberjaya, 63100 , Malaysia**

Abstract

In this paper, we propose an idea to use subword synthesis units to generate novel word sound of Malay language. The subword synthesis units were carefully created based on Malay phoneme adjacency analysis. The phoneme adjacency analysis was used to find out which adjacent phonemes that would not introduce audible distortion if the phonemes were concatenated. Besides phoneme adjacency factor, we also include prosodic features in selecting the subword synthesis unit. Thus, by considering both phoneme adjacency and prosodic feature in selecting the subword units, a smooth sound of novel word can be obtained.

Keywords: *Subword Concatenation, Phoneme Adjacency Analysis, Concatenative synthesis, Speech Synthesis, Malay language.*

1. Introduction

The work presented in this paper is part of work done for Malay speech synthesizer pilot system called UTMK-MSS system. UTMK-MSS system was developed based on the approach that prioritizes naturalness before flexibility aspect. The reason why naturalness prioritized before flexibility is because most Malay Speech synthesizer systems are flexible but their naturalness quality needs to be improved.

Naturalness quality can be achieved in speech synthesizer system if both segmental quality and prosodic quality of the voice systems are high. Synthesis voice with high prosodic quality is when there are no mismatched of duration, pitch and intensity between the joined speech units. In the other hand, high segmental quality is when the joined units do not have audible discontinuity that is caused by spectral mismatched [1].

Thus, in order for the UTMK-MSS system to speak naturally, the system need to choose speech units that once concatenated will introduce less or none audible distortion caused by the prosodic and segmental mismatched. In UTMK-MSS system, prosodic mismatches are avoided by selecting the most appropriate speech units using NLP approach (see Fig.1). Whereas to avoid segmental mismatched, only larger chunks are considered: Word and subword.

Subword unit is the synthesis unit that will be used to generate novel word unit. However, the concatenation of this unit potentially can degrade the naturalness of generated speech that caused by the segmental (spectral) mismatched. Thus, we avoid the possible of subword segmental mismatched by only synthesizing novel word from the list of subword unit that do not cause audible distortion.

These subword units were created based on the result of phonemes adjacency analysis. It was an analysis that listed out which Malay phonemes can be adjacent with other Malay phonemes without audible discontinuities. We gathered the findings of other researchers about non-audible phonemes concatenation and used it as a guide for non-audible Malay phonemes adjacency, which then used to build the subword unit lookup as in Fig. 1.

The phoneme adjacency analysis procedure and result, and also how we build the subword lookup are being described in section 2. In section 3, we explain how novel word sounds are created using subword concatenation. Afterwards, the evaluation on the subword concatenation will be in section 4 and 5. Finally in section 6, we end our paper with conclusion and discussion.

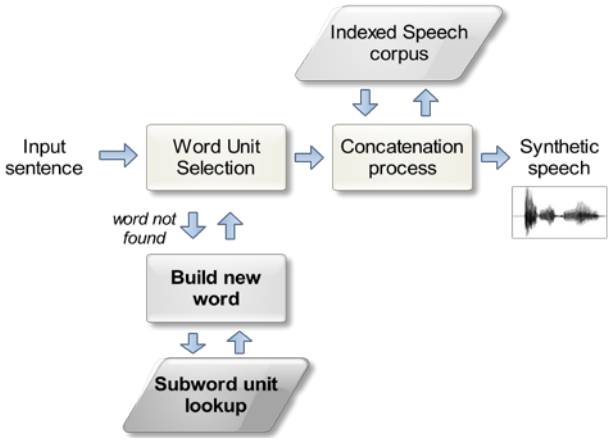

Fig. 1: The overall of UTMK-MSS process.

2. Malay Phoneme Adjacency Analysis

The subword lookup (see Fig. 1) is built based on the result of phonemes adjacency analysis. Building the subword unit lookup required two steps; (1) to find out what phonemes adjacent to other phonemes that will not introduce audible distortion, and (2) to build a list of subwords from the existing word synthesis units based on the list of phonemes in step (1).

2.1 Phoneme Adjacency Analysis

Table 1: Malay consonants and vowels.

Consonants	***Stops (7):*** /g/, /d/, /b/, /p/, /t/, /k/, /ʔ/ ***Affricates (2)*** : /ʤ/, /ʧ/ ***Fricatives (9):*** /s/,/f/,/v/,/z/,/ʃ/,/θ/, /ð/, /x/, /ɣ/, /h/ ***Liquid (2):*** /r/, /l/ ***Nasal (4):*** /m/, /n/, /ɲ/,/ŋ/ ***Semi-vowel/glides (3):*** /w/,/j/
Vowels	/a/, /e/, /i/, /o/, /u/, /ə/
Dipthongs	/ai/, /ao/, /au/

The phonemes in Malay consists of 27 consonants, 6 vowels and 3 dipthongs [2][3] (see Table 1).

In [4]'s study, the stop, fricatives and affricates would not introduce discontinuities if spliced is happened between these consonants. Therefore, we perceptually tested the joined of the inter-categories and intra-categories of those consonants for Malay. Two samples of phonemes were took out from each category and tested for their combination (see Table 2 and Table 3). The perceptual test was done by concatenating subwords from bisyllabic or polysyllabic word and had been recorded in isolated manners.

In perceptual test 1 (Table 2), we purposely chose /h/ to represent fricatives consonant. This is because /h/ is identified as glide in [3], but [5] and [2] stated that /h/ belongs to fricative consonants. Since we are interested in fricative consonants, therefore /h/ has to be considered in the test.

Table 2: Perceptual test I.

	Stop	***Fricative***	***Affricate***
Stop	/p/ and k/	/k/ and /z/	/k/ and /ʧ /
Fricative	/ɣ/ and /t/	/f/ and /z/ /ɣ/ and /s/ /h/ and /z/	/h/ and /ʧ/ /f/ and /ʧ/
Affricate	/ʤ/ and /t/	/ʤ/ and /s/ / ʤ / and /h/	/ʧ/ and /ʧ/ /ʤ/ and /ʧ/

Table 3: Perceptual test II.

	Alveolar	***Dental***	***Palatal***
Alveolar	/n/ and /z/ /s/ and /z/ /t/ and /z/ /l/ and /z/	/k/ and / ð /	/z/ and /j /
Dental	/ð/ and /t/	/ð / and /s/	/ð / and /j/
Palatal	/j/ and /z/	/ ð/ and /j/	/ɲ/ and /j/

Based on the perceptual test 1, the joint of phonemes either within the inner or between the outer categories did not produce audible discontinuities. As for affricate /ʧ/, the result obtained was similar to [4]'s finding, and therefore, the suggestion made by [5] was disregarded. In the other hand, perceptual test on /h/ turned out to be as smooth as the other tested fricatives.

Tests were also carried out to find whether it is also true for consonants from alveolar, dental and palatal, if were replaced with each other, will not produce audible distortion. This information is useful if one wants to obtain

subword ended with vowel, but, unable to find the matched adjacent consonant of the next subword.

For an example, if the word to be synthesised is *diazan (*'uttering the moslem call prayer'*)*, but, the speech corpus only contains the words *azan* ('the moslem call for prayer') and *dialas* ('putting a cover'). The subword unit /*dia*/ can be extracted from the word *dialas* ('putting a cover') and concatenates it with the subword unit /zan/ from the word '*azan*' ('the moslem call for prayer').

In Malay, the consonants under these three categories are: (1) alveolar - /t/,/d/,/s/, /z/, /n/, /l/, (2) dental - /θ/, /ð/ and (3) palatal - /ɲ/, /j/. Again, two samples of phonemes from each category are chosen for the perceptual test. The joints of the inter-categories and intra-categories of those consonants were tested out perceptually (see Table 3). The combination of all of the phonemes sounded natural except a 'click' happened between the joint points. Since by using fading-out and fading-in effects, the 'click' sound can be removed or reduced, the phonemes of alveolar, dental and palatal are included in the subword lookup list.

2.2 Building the Subword Lookup

Using the phonemes adjacency result in section 2.1, we built a list of subword that will not introduce audible discontinuities when the subwords are joined together. Since the subwords units will be used to synthesize new word, therefore, the list of subwords unit must be created from the existed words in the speech corpus.

Creating the subword unit lookup was done according to these steps: Firstly, all words (strings) from the speech corpus were extracted out. Secondly, automatically using [7]'s program, letters were converted into phonemes and syllabified into phonemized words. Finally, based on the result of phoneme adjacency analysis, syllables were united with their adjacent syllables, if their adjacent phonemes were not any of the tested consonants (see Table 2 and Table 3). This is how the subword unit lookup was built. Fig. 2 presents the overall process of building the subword unit lookup based on the phoneme adjacency analysis.

The words units in the speech corpus are annotated with Part-of-Speech (POS), types of phrase break and prominent marks (or none if the word is not prominent). Therefore, the subwords will also inherit the same prosodic features from word which the subwords were obtained.

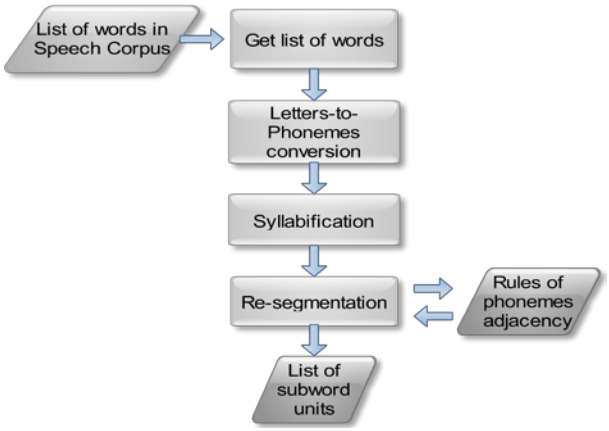

Fig. 2 The overview process of subword unit lookup.

Table 4 shows the example of indexed subword unit lookup. In Table 4, subword string that contains numerical of '0' or '1', indicating the position of the syllables of the subword unit within the word. Numerical '0' signifies that the subword located at the first syllable of a word, and '1' indicates the subword unit is a final syllable(s). The subword without numerical '0' and '1' means that the subword is a middle position syllable(s). For example, in Table 4, the subword /*sil1*/ is from the final syllable of the word *komunikasi* ('communication'), inherits noun as POS, with the starting time of 7.54 ms and ended at 7.98 of the wave file (with node ID 2879).

Table 4: The example of subword unit lookup entries.

Subword	***POS & Prosodic***	***Start Time***	***End Time***	***String***	***ID***
0komu	N,1	7.07	7.26	komunikasi	2879
nika	N,1	7.26	7.54	komunikasi	2879
si1	N,1	7.54	7.98	komunikasi	2879

3. Subword Units Concatenation

The subword concatenation is a concatenation of synthesis unit which is smaller than word units, or a process known as 'build new word' in Fig.1.

As mentioned before, in order not to degrade a synthesized utterance, both segmental and prosodic must not have any mismatches. Segmental mismatch will be avoided by joining subword units that will not introduce audible distortion, whereas for prosodic mismatch, we avoid it by

selecting subword based on the subword prosodic features and its position in a sentence.

As pointed out by [6] in their syllable re-combining rules that the position of syllable within a word is not important except when the word located at the start and the end of a phrase. The rules can be also applied to subword. However, we want to be very particular on the location of subword not only at the phrase level but also at the word level, especially on the first and last position of subword segment in a word. Subword located at the first and the last position in a word must be replaced with the same subword with similar positions in order to avoid audible distortion.

The subword matching steps are similar to the steps of building the subword unit lookup. In the subword matching, the first step is to convert the letters in the unmatched word into phonemes. Then, the phonemized words are segmented into syllables using the same syllabification rules of [7]'s program. Symbol of '0' and '1' are appended to the first syllable and the last syllable respectively. Using the phoneme adjacency rules in Table 2 and Table 3, the syllables segment in the word are re-segmented.

A program's output below shows the output of every step to create subwords strings of the unmatched word for subword matching.

```
UNMATCHED WORD     : menjelaskan
PHONEMISED WORD    : m e n ʤ ə l a s k a n
SYLLABIFICATION    : men.ʤə.las.kan
SYLLABLE POSITION  : 0men.ʤə.las.kan1
RE-SEGMENTATION    : 0men.ʤəlas.kan1
```

The string delimited by "." of the RE-SEGMENTATION output in output source code above will be used to be matched against the subword unit lookup. In the matching, both subword string and prosodic features are used to retrieve the matched subword unit. The reason to use the prosodic features instead of POS is because we assume matched prosodic features will ensure no or less prosodic mismatch occurs between the joined subword. We do not prioritize POS since we want a higher number of subword unit candidates to be retrieved.

Besides the prosodic features, subword units that come from the same wave are given priority compared to subword unit from other wave file with similar POS and prosodic features.

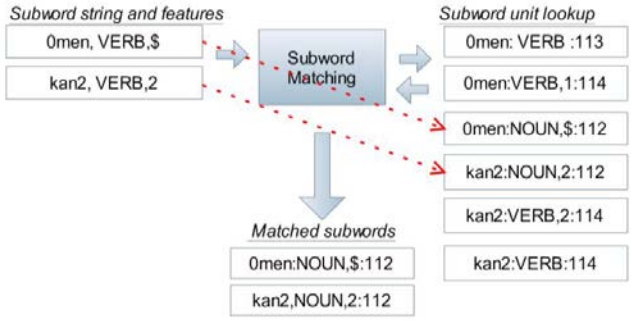

Figure 3 An example of which prosodic features and subword originated from the same wave file are given higher priority.

For example, in Fig. 3, the subword unit of */0men/, NOUN,$*' was chosen based on the prosodic value of '*$*' although its type of POS is different from the target subword. The subword unit of /*kan*2/ with SSTC-ID of '*112*' was chosen instead of the other subword /*ka*n2/, since it comes from the same wave file as the other chosen subword unit.

4. The Smoothness Test

In order to evaluate whether the joint between the subword units is smoothly concatenated, we ran a smoothness test. The smoothness test is a subjective test where subjects will listen and choose which subword joints are smooth and which are not. In this paper, the smoothness test was conducated by replicating the same smoothness test conducted by [8] and [9].

We use the smoothness test setup of [8] [9] because our objective and their experiment objective are the same; that is to assess the joint of small speech units (the size of speech unit smaller than words). In the next subsections, we will describe in detail the data test use, description of participants and how the test conducted. The result obtained will be discussed in next section, which is section 5.

4.1 Participants, Test Data and Procedure

4.1.1 The Participants

The participant of our smoothness test was about 37 people that were invited through phone calls, meeting-in person and e-mails. All of the participants were Malay native speakers with no hearing problem. The reason why only

native Malay was chosen was because native speaker's judgment is more competent than the non-native speaker, due to the native speaker better language knowledge than non-native speakers [12]. The gender distribution of male and female was also balanced with 51% were female and 49% male. We also only invited participants who were not working as language technologist. A language technologist defined here is a person who is working in speech technology, natural language processing or understanding, computational linguistics and other related fields in speech and language. This is to ensure the test is done based on the people who has no idea about the state-of-arts of speech synthesis and thus, real user's judgement on smoothness quality in speech synthesis can be captured. Since aged people might have hearing problems, and young people might not sensitive enough to express their judgement, the smoothness test were only taken by people who are aged above 20 but not over 50 years old.

4.1.2 Test Data

Since, we unable to have all the words in the target sentences to be synthesised with subwords unit, the sound test for the smoothness test were made up by combining the concatenation of word and subword synthesis units in total of only five sentences (.wav files).

Table 5: The test data for smoothness test.

Sound	***Unit concatenation***
Estc1	{penggunaan}{0me}{dia1}{0ter} {diri1}{0da}{lam1}{0ben}{tuk1} {0komu}{nika}{si1}{#}{yang} {0men}{gunakan1}{radio}
Estc2	{agak}{0je}{las1{0un}{tuk1} {0me}{nerang}{kan1}{dakwah}{tertentu} {#}{0kera}{na1}{dapat}{memahami} {0ting}{kah1{laku}{0se}{seorang1}
Estc7	{bab}{0ter}{sebut1}{0men}{je}{laskan} {dan}{0meng}{huraikan1}{motif}{dan} {sikap}{mereka}
Estc8	{kehidupan}{yang}{0ber}{lumba1} {0un}{tuk1}{0men}{cari1}{kekayaan} {membuat}{0ki}{ta1}{jahil}
Estc10	{matlamat}{penyampaian}{dakwah} {0a}{da}{lah1} {supaya}{0ma}{nu}{sia1} {0men}{ja}{di1}{baik}

In Table 5, the curly bracket symbol shows the length of synthesis units, whether word or subword were used in the synthesised sentences. The symbol '#' represent a silence unit and the appended number '0' and '1' meaning part of the subwords sound are initial syllable or final syllable.

4.1.3 The Procedure

[8] and [9] conducted their smoothness tests by asking the participant to listen to a synthesised sentence and expressed their judgement on the smoothness of joint synthesis units. In the smoothness test of [8] and [9], the participants were asked to mark words that did not sound smoothly. In our smoothness evaluation test, we asked the participants to mark (by ticking) the empty box next to the word that they perceived as not smooth.

A simple GUI program was developed to conduct the evaluation test. The participants used headphones or speakers to listen to the test sound when they clicked to the corresponding buttons. The participants can replay the test sentences as many times as they want to.

5. Performance Measure for Smoothness Test

The performance measure we used for the smoothness test was by adapting the performance measures of phrase break in [10] and [11]. By replacing (see Table 6 formula acronym definition):

- Total junctures (N) as the total of words and joint subwords (TSWW).
- Total breaks (B) as total of joint subwords (TSW).
- The deletion error (D) as total of joint subwords perceived as word (SWW).
- Insertion error (I) as the total of word perceived as joint subwords (WSW).

The formula for BC and JC in [10] and [11] were used according to our assessment objective.

Table 6: Data collected from smoothness test

Abbreviation	***Description***	***Total words***
TSWW	*Total of all words and joint subwords*	1656
TSW	*Total of all joint subwords*	864
TW	*Total of all words*	792
SWSW	*Total of not smooth joint subwords*	364
WSW	*Total of not smooth words*	140
SWW	*Total of smooth joint subwords*	500
WW	*Total of smooth words*	652

The BC formula in [10] was adapted to measure the percentage of subword sounds that were perceived as not smooth by the participants. We named this formula as SWSW-CORRECT.

$$SWSW\text{-}CORRECT = \frac{TSW - SWW}{TSW} *100\%$$
$$= \frac{864 - 500}{864} *100\%$$
$$= \mathbf{42.13\%} \quad (1)$$

We adapted the JC formula in [10] for SWW-INCORRECT calculation that measures how many percent of subword and word that were not smoothly perceived by the participants. In JC formula, S value is included since it is referred as the type of phrase break predicted. However, we ignored S since we do not differentiate the joining of subword. We also ignore I, and instead we replaced it the value of WW. This is because we have to get rid of any word that was perceived as smooth.

$$SWW\text{-}CORRECT = \frac{TSWW - SWW - WW}{TSWW} *100\%$$
$$= \frac{1656 - 500 - 652}{1656} *100\%$$
$$= \mathbf{30.43\%} \quad (2)$$

We again adapted the BC formula to calculate the percentage of word sounds that were smoothly perceived by the participants. We named the modified BC formula as WW-CORRECT.

$$WW\text{-}CORRECT = \frac{TW - WSW}{TW} *100\%$$
$$= \frac{792 - 140}{792} *100\%$$
$$= \mathbf{82.32\%} \quad (3)$$

Table 7: The assessment scores result.

Description	***Measurement***	***Words (%)***
Subword perceived as not smooth	SWSW-CORRECT	42.13
Subwords and words perceived as not smooth	SWW-INCORRECT	30.43
Word perceived as smooth	WW-CORRECT	82.32

6. Discussion and Conclusion

The objective of the perceptual smoothness test was to find out the percentage of non-smooth joint subwords. Since the joint subwords came from the list of subwords that were assumed would not create audible distortion, therefore, the value of SWSW-CORRECT should be lower and closer to zero percent. Based on the result presented in Table 7, we can conclude that the result of 42.13% is considered as bad result for our objective, which is to have none or less audible distortion for subwords concatenation. However, the result of word perceived as smooth (WW-CORRECT) with only at 82.32%, gives an impression that bad result of SWSW-CORRECT is not solely contributed by the subword concatenation. Since the word synthesis unit is not relevant for segmental mismatch (spectral mismatch), therefore, the audible distortion (non-smoothness) in the smoothness test was more likely caused by the prosodic mismatch. However, with only 30.43% of non-smooth word and joint subwords sounds perceived by participants, suggested that the approach of selecting word and subword unit for synthesis was not bad after all.

We would like to suggest that possibly, the culprit of the non-smoothness in the joint subwords was not only at the segmental mismatch but also at the prosodic mismatch. This was due to the fact that word unit was also perceived as non-smooth by the participants when word unit has nothing to do with segmental mismatch. In other words, the whole process of selecting speech unit for synthesis requires further investigation for an improvement, not only at the subwords units but also at the word units.

References

[1] Klabbers, E. and Veldhuis, R., "Reducing audible spectral discontinuities", IEEE Transcations On Speech and Audio Processing, Vol.9, No.1, 2001, pp.39–51.
[2] Maris, Y., The Malay Sound System, Kuala Lumpur: Fajar Bakti, 1980.
[3] Teoh, B. S., The Sound System of Malay Revisited, Kuala Lumpur: Dewan Bahasa dan Pustaka, Kuala Lumpur, 1994.
[4] Yi, J. R. and Glass, J. R. (1998). "Natural-sounding speech synthesis using variable-length units", Proceeding of ICSLP, 1998, pp.1167–1170.
[5] Samsudin, N., "Word selection characteristics based on the manner of articulation", Technical Report, Universiti Sains Malaysia, 2004.
[6] Lewis, E. and Tatham, M., "Word and syllable concatenation in text-to-speech synthesis", Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 615–618.
[7] Tan, T. P., Grapheme to Phoneme System, Unit Terjemahan Melalui Komputer (UTMK), Universiti Sains Malaysia, Penang, Malaysia, 2008.

[8] Tomokiyo, L. M., Black, A. and Lenzo, K. A., "Foreign accents in synthesis: Development and Evaluation", 2005, INTERSPEECH2005.
[9] Tomokiyo, L. M., Peterson, K., Black, A. and Lenzo, K. A. "Intelligibility of machine translation output in speech synthesis", 2006, INTERSPEECH2006.
[10] Taylor, P. and Black, A., "Assigning phrase breaks from part-of-speech sequences", Computer Speech and Language Vol.(12), 1998, pp: 99–117.
[11] Atterer, M. and Klein, E., "Integrating linguistic and performance-based constraints for assigning phrase breaks", Proceedings of the 19th international conference on Computational linguistics, 2002, pp. 1–7.
[12] Jones, C., Berry, L. and Stevens, C., Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners, Computer Speech and Language, Vol. (21), 2007, pp. 641–651.

First Author Sabrina Tiun received her Ph.D from Universiti Sains Malaysia recently. Her research work and interests range from Natural Language Processing, Computational Linguistics to Speech Processing. She is currently working as a lecturer in Universiti Sains Malaysia, Penang, Malaysia.

Second Author Rosni Abdullah is a professor and the Dean of the school of Computer Sciences in Universiti Sains Malaysia, Penang, Malaysia. She is known for her actively involvement in the research of Parallel Processing, Bioinformatics and Cloud Computing.

Third Author Tang Enya Kong is an Associate Professor at Multimedia University, Selangor, Malaysia. He is one of the prominent researchers in Natural Language Processing community in Malaysia. His works covers from Machine Translation, Natural Language Processing, Computational Linguistics, Knowledge Representation and Speech Processing.