Journal of Computer Science 2012, 8 (12), 1961-1969
ISSN 1549-3636

doi:10.3844/jcssp.2012.1961.1969 Published Online 8 (12) 2012 (http://www.thescipub.com/jcs.toc)

Restricted Domain Malay Speech Synthesizer Using Syntax-Prosody Representation

[1]Sabrina Tiun, [2]Rosni Abdullah and [2]Tang Enya Kong

[1]Faculty of Technology and Information Science, University Kebangsaan Malaysia, Selangor, Malaysia
[2]School of Computer Sciences, University Sains Malaysia, Selangor, Malaysia

Received 2012-08-29, Revised 2012-10-02; Accepted 2012-11-13

ABSTRACT

The speech synthesis approach required in restricted domain speech application is a synthesizer that has high quality like the speech output of 'slot-filler' approach but have at least the least flexibility of the 'genuine' speech synthesizer. Thus, in this research study, we propose an alternative approach of creating a speech synthesizer to be used in a restricted domain speech application. In our approach, we use word unit as the primary unit and our speech corpus is represented by syntax-prosody tree structures. Speech synthesis is performed by constructing a syntax-prosody tree of a target input sentence. The construction of the tree is by done by adapting an example-based syntactic parsing approach and the concatenated of synthesis units from the constructed tree nodes will be the synthesized utterance. For evaluation, we performed MOS subjective evaluation on our speech synthesizer with natural speech and two other Malay TTS system. Based on an ANOVA and T-Tests analysis, we found the overall MOS scores of our speech synthesizer output, sound B was (mean = 3.34, sd = 1.10), the other two Malay TTS system; C (mean = 1.95, sd = 0.72) and D (mean = 1.80, sd = 1.04) and the natural speech, A (mean = 4.71, sd = 0.21). We conclude that our Malay speech synthesizer sounded more natural, easier to listen, more pleasant and more fluent compared to the sounds of the other two Malay TTS systems. As expected, the recorded speech was perceived more natural than the output of our Malay speech synthesizer.

Keywords: Malay Speech Synthesis, Restricted Domain Speech Synthesis, Syntax-Prosody Representation

1. INTRODUCTION

In limited domain speech synthesis, the voice synthesis is expected to be highly natural sounding, which mimicking human's voice. In limited domain speech synthesizer, it is possible to deliver such expectation due to the limited vocabulary of the limited domain application. This limited domain application usually requires less number of new words and has small number of vocabulary. Thus, it does not really need a very intelligent speech synthesizer or a big size of speech corpus. Thus, it is then possible to have large chunk size of synthesis units like words and phrases or even sentences.

According to Taylor (2000), the approaches of speech synthesizer in limited domain speech application are divided into two types; (1) slot-filler approach and (2) 'genuine' speech synthesizer or also known as Text-to-Speech (TTS). The slot-filler approach is an approach in speech synthesis that uses templates of pre-recorded utterance. The 'slot' is defined as the space in the pre-recorded template that will be filled by 'fillers'. Fillers are the infrequent speech chunks, normally in words or phrases form. Example of infrequent speech chunks will be the names of people or places, or date and times. The words in the pre-recorded template of slot-filler are usually the frequent words. For a 'genuine' speech synthesizer, or the typical TTS system, there will be no pre-recorded template. Given any target text sentence, the synthesizer will be able utter the sentence. This capability is known as intelligent or flexibility in speech synthesis system. However, the drawback of 'genuine' speech synthesizer is usually its unnatural speech output.

In a very limited domain of speech application like weather broadcast or travel information broadcast, slot-filler approach is feasible since the number of infrequent words is small and vocabulary is limited. However, for limited domain that has larger vocabulary and higher number of possible new words, or also known as a

Corresponding Author: Sabrina Tiun, Faculty of Technology and Information Science, University Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia Tel: (+603) -3921-6730

restricted domain, using the slot-filler approach for its speech synthesizer is not suitable. Using the 'genuine' speech synthesis approach may also unsuitable due to its unnatural sounding output. Thus, the speech synthesis approach required in restricted domain speech application is a synthesizer that has high quality like the speech output of 'slot-filler' approach but have at least the least flexibility of the 'genuine' speech synthesizer.

In this research study, we propose an alternative approach of speech synthesizer to be used in the restricted domain speech application. In our approach, we use word unit as the primary unit to synthesize the target text if phrases or whole sentences units are not available. The approach to select the suitable word synthesis units for concatenation is by using a speech corpus represented by syntax-prosody trees. We do not use the standard unit selection approach in choosing the most suitable candidates units. Instead, we adapted the example-based parsing used in a machine translation. Our speech synthesizer also has a better flexibility quality than the slot-filler approach since the speech synthesizer will also have syllable-like synthesis, which we have discussed much detail in Sabrina *et al*. (2011).

2. MATERIAL AND METHODS

2.1. Syntax-Prosody Representation

Our mini syntax-prosody speech corpus consists of 422 sentences (trees), 1720 phrases (sub-trees), 145 word vocabulary, 6978 word counts and 2858 sub-words (Sabrina *et al*., 2011). We represent our speech corpus using a syntax-prosody representation. Each of a sentence in our speech corpus corresponds to a single syntax-prosody tree structure. The tree structure is a dependency syntactic tree, with each of its nodes annotated with Part-Of-Speech (POS), prosodic features of prominent marks and phrasal breaks and aligned with a speech unit. The dependency tree structure is built based on String Structured Tree Correspondences (SSTC) structure, where each word corresponds to each node and each phrase corresponds to each sub-tree, or also known as subSSTC. **Figure 1** shows a syntax-prosody tree structure corresponding to a sentence of wave file (**Fig. 2**). In both the tree and the wave file, prosodic features are annotated. Symbol '$' is annotated to the word located at the beginning of a sentence. The symbol '*' is prominent symbol, indicating that the annotated word with such symbol contains prominent syllable or syllables. Word with symbol '1' signifies the word is located at the end of a phrase (the phrasal break). Such word is suspected to have obvious duration and pitch curve or energy compared to the rest of the words. Finally, word located at the end of a sentence will be annotated with symbol '2'. For more detail on the construction of the speech corpus, one can refer to an online documentation at Sabrina *et al*. (2011).

2.2. Word Unit Selection

Our speech synthesizer or we named it as the UTMK-MSS has four steps in order to parse an input sentence into a syntax-prosody tree: (1) Tagging, (2) lexical matching, (3) structural matching and (4) recombination. A synthesizer module is use to synthesize the utterance of the input sentence. **Figure 3** shows the simplified diagram of our UTMK-MSS system. The shaded boxes (except the box with text 'build new word') are the four steps.

2.3. Tagging

Prior to the lexical matching process, the target words are tagged with POS and prosodic features. The Malay POS Tagger is an adapted tagger from a portable probabilistic language-independent POS tagger named *Qtaq* (Mason, 2009). Target words are also tagged with prosodic features based on punctuation symbols; e.g. comma, semicolon, period. The words before the punctuation symbols will be tagged with break types by assuming that those words have different degree of speech properties; longer duration, declining pitch and lower energy, compared to the rest of the words in the target sentence. Besides period symbol, which is tagged with break type '2', the rest of the symbols; comma and semicolon, are tagged with symbol of break type '1'. The word at the beginning of the sentence is marked with symbol '$'. This is to ensure that the lexical matching only retrieves sub-trees that are indexed with the word located at the beginning of the sub-trees string, if matching based on word with symbol '$'. It is assumed that word at the beginning and the end will cause audible distortion when they are concatenated at any location besides their respective locations and this is due to the occurrence of prosodic mismatch.

2.4. Lexical Matching

The lexical matching process mainly involves with word matching, if a whole sentence matching or phrases are not found in the indexed speech corpus. The word matching is particularly concerned with certain positions of words in the target sentence; (1) the word at the beginning position, (2) at the phrase break and (3) at the end of a sentence. This is because word at the beginning and end of sentence and at the end of a phrase has distinct speech characteristics, which is, if it is replaced by the same word but originated from a different positions, it is highly possible that prosodic mismatch will occur. In the word matching, POS will be least important than prosodic feature. Thus, if the process unable to retrieve the exact matches of target POS and prosodic feature, POS will be ignored. In the **Fig. 4** the word agak ('*maybe*') with POS of Verb (V) was chosen instead word agak ('*maybe*') with POS of Adverb (ADV). This is because word matching prioritizes word string and the prosodic feature (in this case, word position is included as prosodic feature as well). The output of the word matching process will be a pool of sub-trees (or subSSTCs).

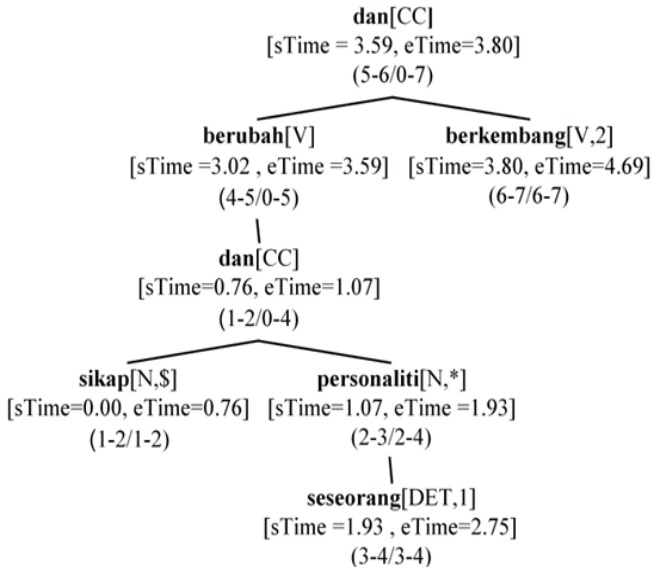

Fig. 1. Syntax-prosody tree structure of string sikap$_1$ dan$_2$ personaliti$_3$ seseorang$_4$ berubah$_5$ dan$_6$ berkembang$_7$ ('the attitude and the personality of someone are changing and evolving')

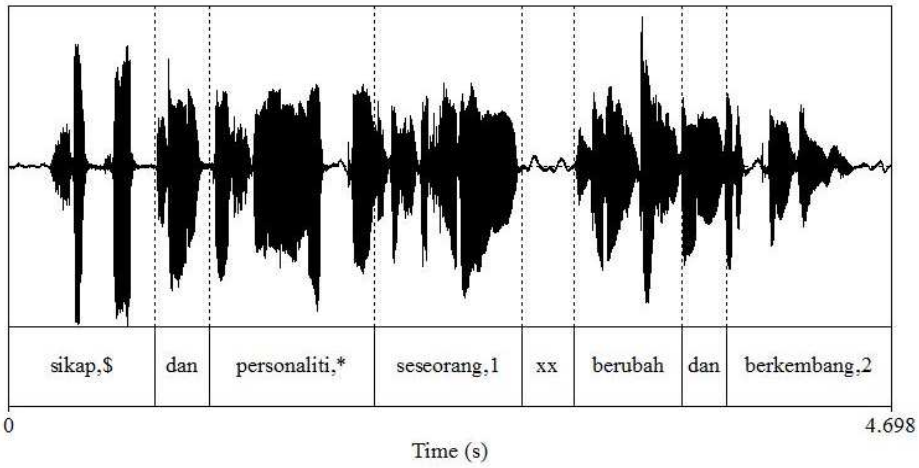

Fig. 2. Wave file that has been segmented, labeled and annotated with prosodic features corresponds to the sentence and syntax-prosody tree structure in **Fig. 1**

After the lexical matching, the rest of the unmatched word will be handled by combining sub-word strings. Since every sub-word is aligned with sub-word synthesis unit, therefore synthesizing the sound of unmatched word is by concatenating the sound of the combined sub-words strings. Detail on sub-word unit matching and concatenation can be found in Sabrina *et al*. (2011). At the end of the lexical matching process is a pool of relevant sub-trees. However, not all retrieved sub-trees will be used for the final construction of the parsed tree (of the input sentence), since, only the best candidates will be chosen. Thus, the criteria of best set of sub-trees are based on the co-occurrence and frequency. Co-occurrence is defined as when an example contains the highest number of sub-trees and if the condition does not exist, the retrieved sub-trees with the highest frequencies in the database will be selected instead. In order to combine these sub-trees into a well-formed parsed tree structure, the structural matching and recombination process are needed.

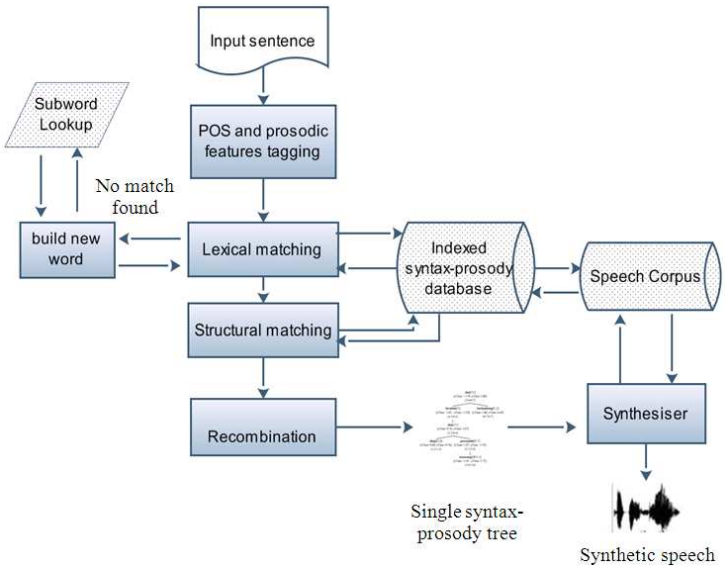

Fig. 3. The simplified diagram of UTMK-MSS processes

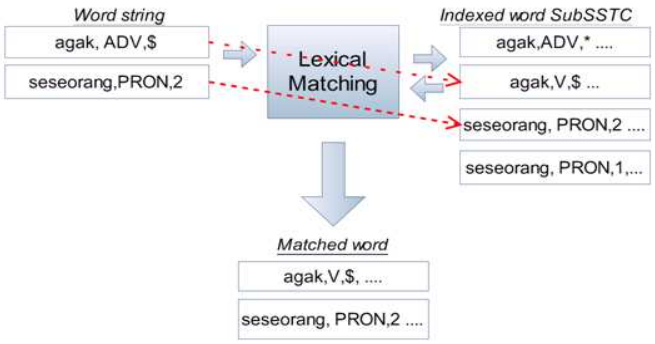

Fig. 4. The above figure shows the word agak (‘*maybe*’) with POS of verb (V) was chosen instead word agak (‘*maybe*’) with POS of Adverb (ADV)

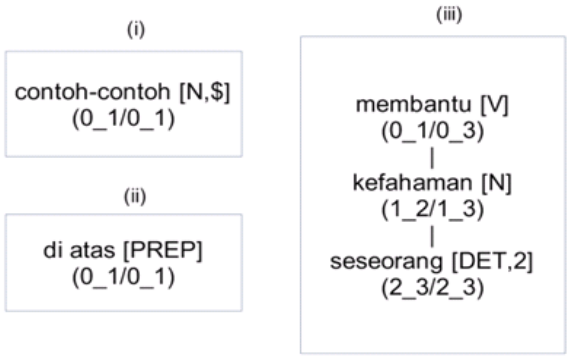

Fig. 5. List of sub-trees for sentence of contoh-contoh di atas membantu kefahaman seseorang (‘the above examples helped anybody’s understanding’)

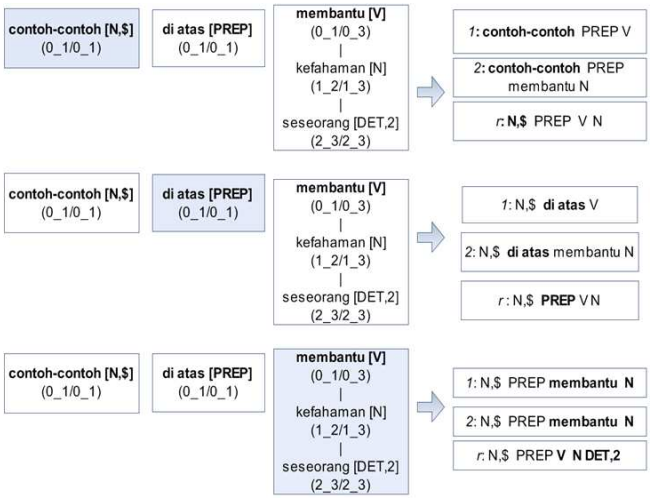

Fig. **6.** Example of generalized sub-trees

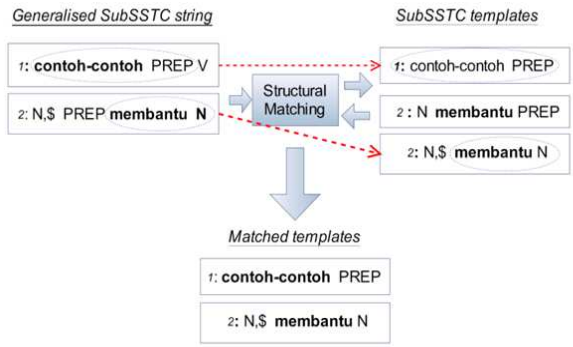

Fig. 7. Examples of matched sub-trees (or subSSTC)

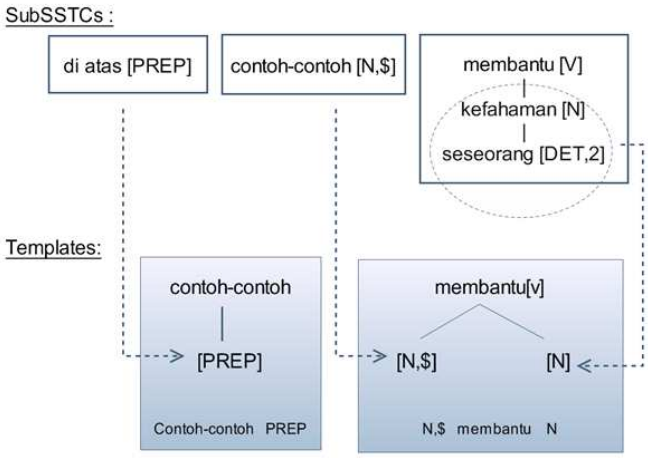

Fig. 8. Recombination process

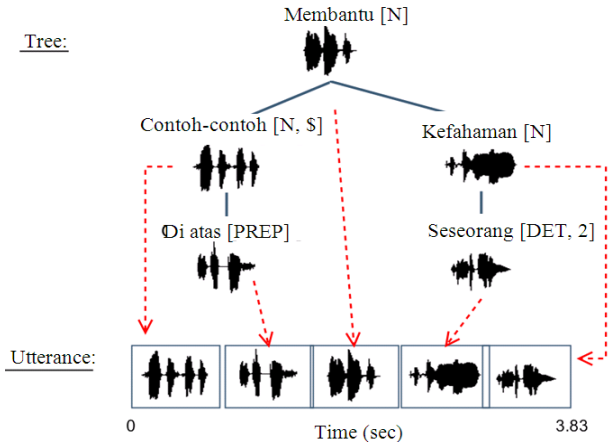

Fig. 9. Concatenating aligned speech units from a parsed syntax-prosody tree structure

2.5. Structural and Recombination

In order to construct a single parsed tree from the pool of sub-trees, the structural matching and recombination processes are performed. Prosodic features are included as one of the main features in the matching and recombining sub-trees. Thus, let us say we have a string of contoh-contoh diatas membantu kefahaman seseorang (‘the above examples helped anybody’s understanding’) as an input into the lexical matching. Based on the tagging process and lexical matching, the matched sub-trees are retrieved as listed in **Fig. 5**.

At the structural matching, the sub-trees listed in **Fig. 5** will be generalized into POS except the root node. Sub-tree generalization is a process where all the nodes of sub-tree are generalized into POS, except for the targeted root node of the sub-tree. For example, in **Fig. 6** when the sub-tree of contoh-contoh [N,$] (‘examples’) is the target sub-tree, its root node will not be generalized into POS like the rest of the sub-trees. The generalized sub-tree will be used to retrieve sub-tree templates. In the example-based parsing of [8], there are four types of templates; type 1, type 2, type 3 and rule. For the synthesis unit selection, we only use type 1, type 2 and rule template since type 3 is a partial tree structure template that is purposely made for handling complex translation process like idiom expression. The other node structure templates are defined as follows; type 1 is a template for structure tree with one level depth, type 2 is a two level depth of node structure template and rule template is one level depth node structure with all the nodes are generalized into POS. **Figure 6** shows that at each of the generalization process, the shaded box indicates the sub-tree which is assumed to be the potential root node for the combination of all the retrieved sub-trees. Boxes after the arrows are the generalized strings based on template types; type 1 = 1, type 2 = 2 and rule = r. Afterwards, the generalized sub-tree strings will be matched against the indexed templates (from a template database), **Fig. 7**.

The next step is to combine the templates from the structural matching with sub-tree from lexical matching. This recombination process is done by replacing the nodes in the templates that contain only POS and prosodic with lexicalized nodes. The end result will be the parsed tree of the target input sentence. In **Fig. 8** the nodes [PREP] and [N, $] in the template tree are replaced by nodes di atas [PREP] (‘above’) and contoh-contoh [N, $] (‘examples’) respectively. Whereas, the nodes kefahaman [N] seseorang [DET, 2] (‘anyone's understanding') replaces the Node [N] in the other template tree. Since the tree nodes are aligned with speech units, thus, to produce the utterance of the input sentence is simply by concatenating the aligned speech units.

2.6. Concatenating Synthesis Units

The recombination process generates a single tree, in which its nodes are aligned with speech units. The aligned speech units are extracted out based on the node ID and the start-time and end-time of particular speech segments from targeted .wav files. Using a simple concatenation process, without applying any signal processing, those synthesis units are concatenated. For example, in the **Fig. 9** all the speech units aligned with the nodes of the constructed parsed tree will be concatenated. The dot lines show the corresponding speech units with the speech segments in the generated utterance. If the node is tagged with phrasal break of ’1’, a silence is inserted after its speech segment. In order to avoid the synthesis units being concatenated overlap, a fade-out and fade-in are applied in every synthesis unit. Based on the assumption that the synthesis units are selected with correct prosody using the syntactic parser together with the prosodic features, inserting the correct position of silence and applying fading effect to smoothen the edges of the synthesis unit, it is then assumed that UTMK-MSS be able to generate natural-sounding of Malay synthetic utterance.

3. RESULTS

We evaluated the output of the UTMK-MSS using the Mean Opinion Scores (MOS) test of Viswanathan and Viswanathan (2005). The objective of the MOS test is to find out how natural our speech output compare to natural speech (playback speech) and the other Malay TTS systems. Viswanathan and Viswanathan (2005) MOS test on naturalness contains four items; (i) Voice of naturalness, (ii) ease of listening, (iii) voice pleasantness and (iv) voice of continuity. Each of the items has the scale of 1 to 5 points. In order to assist the participants in making decisions, each of the score point is given a description; for example, 5-*Excellent* 4-*Good* 3-*Fair* 2-*Poor* 1-*Bad*.

3.1. Data and Procedure

For MOS test, we prepared synthesized voice of ten sentences consisting of 9 to 11 lengths of words from UTMK-MSS system and two others Malay TTS systems and also a recorded speech (natural speech). The ten sentences were made up by combining the high frequent words in the mini speech corpus. The made up sentences are syntactically and semantically correct, yet, they are not existed in the speech corpus. In the MOS test, the natural speech was recorded using the voice of an experienced Malay female native speaker and we named the test data as sound. (A) The output from our Malay speech synthesizer, UTMK-MSS, was named as sound. (B) The Malay TTS output produced by using unit selection approach was named as sound (C) and a Malay TTS using fixed diphone unit concatenation approach was labeled as sound. (D) The total number of participants participating in the MOS test was 37. The participants did the evaluation test voluntarily and were invited through phone calls, meeting-in person and e-mails. All of the participants were Malay native speakers with no hearing problem. The gender distribution of male and female was balanced with 51% were female and 49% male. We only invited participants who were not working as language technologist and within the range of age 20 to 50 years old. A simple GUI program was developed for the evaluation test. The participants used headphones or speakers to listen to the test sounds, in which, would only being played once they clicked the corresponding buttons. Participants can replay the sentences as many times as they want. However, they were only allowed to go to the next test if they had completed the current test.

3.2. Test and Results

We ran an ANOVA test to find out whether the means of the A, B, C and D sounds were significantly different. If ANOVA test reveals there is a statistical difference, T-Test will be used to compare the MOS scores of sound B with the other sounds. We had conducted a MOS test on each of the naturalness qualities; voice naturalness, ease of listening, voice pleasantness and voice continuity and we present the result in **Table 1**. We also show the comparison of the B naturalness quality with the other sound in **Fig**. **10**.

We also ran ANOVA test for the overall MOS scores (total of all the items) and the result revealed that there was a significant difference among the sounds A, B, C and D at the $p<0.5$ level for the condition [$F(3,2956) = 1830.38$, $p = 0$]. Subsequent T-Tests analysis was done and the results can be seen in **Table 1** (at the last row). The comparisons of naturalness for recorded speech (sound A), sound B and the two Malay TTS systems speech (C and D) were plotted in **Fig. 11**.

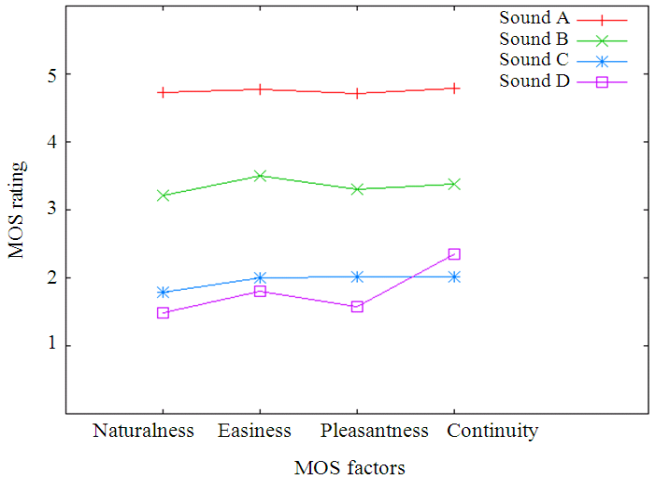

Fig. 10. The comparison line chart of sound A, B, C and D for the four items of naturalness tests; voice of naturalness, ease of listening, voice of pleasantness and voice continuity

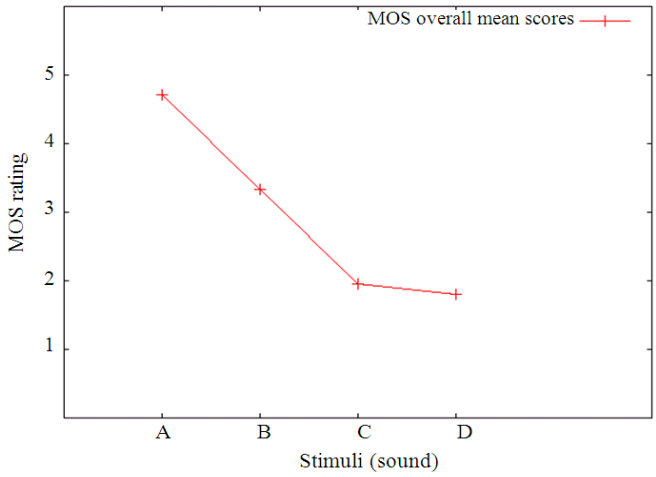

Fig. 11. The comparison line chart of sound A, B, C and D for the overall MOS scores test

Table 1. The T-tests results comparing sound B and sound A, C and D

	Sound							
	B		A		C		D	
MOS test	m	sd	m	sd	m	sd	m	sd
Voice naturalness	3.21	0.97	4.72	0.22	1.76	0.44	1.49	0.50
Ease of listening	3.50	1.15	4.75	0.20	2.00	0.50	1.81	0.98
Voice pleasantness	3.30	1.10	4.71	0.27	2.02	0.89	1.57	0.73
Voice continuity	3.38	1.16	4.79	0.16	2.00	0.74	2.35	1.51
Total of all MOS scores	3.34	1.10	4.71	0.21	1.95	0.72	1.80	1.04

Note: m = mean, sd = standard deviation

4. DISCUSSION

By looking at the line charts in **Fig**. **10 and 11** together with **Table 1** we conclude that the our Malay speech synthesizer sounded more natural, easier to listen, more pleasant and more fluent compared to the sounds of the other two Malay TTS systems. As expected, the recorded speech was perceived more natural than the output of our Malay speech synthesizer. However, as mentioned in Huang *et al*. (2001) that synthetic speech MOS score using the standard MOS of speech coders (scaling 1 to 5 score) is not expected to be around 3.5 to 4.5, which is usually the quality for speech at highly natural and intelligible. In fact, that synthetic speech is typically scored at 2.5 to 3.5. Therefore, the overall mean MOS score of our Malay speech synthesizer at 3.34 shows that its output did not performed below par when compared to the typical synthetic speech quality.

Based on the observation on the four individual MOS test items, our Malay speech synthesizer has the highest MOS score for ease of listening test (mean at 3.5) and the other item tests mean MOS score were just around 3.3. This shows how the participants were willing to hear the voice of our speech synthesizer system for a long period of time despite of its less naturalness, pleasantness and fluency quality. Another point to ponder is the standard deviation (or the variance) of our Malay speech synthesizer MOS scores. Looking back at all the four MOS test items and the overall MOS test, the sound of our Malay speech synthesizer seemed to have wider range of standard deviation compared to the other stimuli. The wider variance of opinion suggested that there is a wide difference on what the participants think of our speech output. The wide gap of opinion could also mean that there is a possible inconsistency of naturalness quality among the synthesized sentence. We suspect that the inconsistence of naturalness quality probably occurs because of the weakness of the corpus-based approach. Since our system is based on a corpus-based synthesis approach, therefore, it may inherit the corpus-based strength as well as weakness. One of the weaknesses of corpus-based speech synthesis is when the least matched

instances of speech units are selected then a less desirable synthetic speech will be generated.

5. CONCLUSION

In this study, we propose an alternative approach in performing a speech synthesis which currently aimed for a restricted domain speech application. For a future work, besides the plan of seeing this research work implemented in a full-scale of restricted domain application like domain specific personal assistance in mobile application, we also want to see our Malay speech synthesizer expanded to be more flexible and more natural. Thus, future work will on flexibility, which is either; (i) we add a finer speech unit than the sub-word unit, yet will not jeopardize the naturalness quality, or (ii) we add more types of sub-words and syllables unit and create those unit recombination rules that can avoid audible distortion when those units are concatenated. For naturalness, enriching the syntactic-prosodic representation with semantic information will be a great help to make the prosody prediction more accurate. The accuracy of prosody prediction task subsequently will increase the naturalness aspect of our speech output

6. ACKNOWLEDGMENT

We would to thank Anuar Mansor for preparing the GUI MOS evaluation program and also to all the voluntarily participants in the MOS survey.

7. REFERENCES

1. Huang, X., A. Acero, A. Acero and H.W. Hon, 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. 1st Edn., Prentice Hall PTR, USA., ISBN-10: 0130226165, pp: 980.
2. Sabrina, T., R. Abdullah and E.K.Tang, 2011. Subword unit concatenation for malay speech synthesis. Int. J. Comput. Sci. 8. 69-74. http://www.ijcsi.org/papers/IJCSI-8-5-2-68-74.pdf
3. Taylor, P., 2000. Concept-to-speech synthesis by phonological structure matching. Philosophical Trans. Royal Soc. Series A, 358: 1403-1417. http://www.era.lib.ed.ac.uk/handle/1842/969
4. Viswanathan, M. and M. Viswanathan, 2005. Measuring speech quality for text-to-speech systems: Development and assessment of a Modified Mean Opinion Score (MOS) Scale. Comput. Speech Language, 19: 55-83. DOI: 10.1016/j.csl.2003.12.001