Prosody Generation by Integrating Rule and Template-based Approaches for Emotional Malay Speech Synthesis

Mumtaz Begum  
Department of Software Engineering, University of Malaya, Malaysia

Raja N. Ainon  
Department of Software Engineering, University of Malaya, Malaysia

Roziati Zainuddin  
Department of Artificial Intelligence, University of Malaya, Malaysia

Zuraidah M. Don  
Department of English Language, University of Malaya, Malaysia

Gerry Knowles  
Miquest Worldwide Sdn. Bhd., Malaysia

Abstract—This paper presents a hybrid technique to enhance the quality of the rule-based approach to generate prosody for Malay speech synthesis by integrating prosody parametric manipulation with template parametric manipulation so as to increase the intonation variability of the synthesized output. Basically the prosodic features of the neutral synthesized speech are manipulated in an attempt to express the four basic emotions, namely happiness, anger, sadness and fear. We also present an objective methodology to evaluate the effectiveness of the synthesized output to generate the appropriate prosody in order to confirm the subjective perception tests.

Key Words—Concatenated emotion synthesis, parametric prosody manipulation, template parametric manipulation, rule-based approach

I. INTRODUCTION

In the typical speech synthesizer, prosody information affects the pitch contours and duration factors of the sounds being generated in response to text input. To a greater or lesser degree, all of the current synthesis techniques sound unnatural unless prosody information is added. Prosody refers to the rhythmic and intonation aspects of a spoken language which includes a combination of pitch, duration and intensity. In natural speech, prosodic features can be manipulated to express different emotions [1, 2].

There are many approaches to generate the required prosody in synthesized emotional speech, such as corpus-based, rule-based and template-based methods [3,4,5]. However, the most commonly used approach is the rule-based approach as it is most computationally efficient, even though the output may sound unnatural. Our previous work on the rule-based approach to add emotion filter to a Malay text-to-speech system reported reasonably good recognition rates from perception tests [6,7]. In [8] we employed the template-based approach and also reported good results. In this paper we present a hybrid technique to enhance the quality of the rule-based approach to generate prosody for Malay speech synthesis by integrating prosody parametric manipulation (PPM) (a form of rule-based approach) with template parametric manipulation (TPM) so as to increase the intonation variability of the synthesized output. In particular, in this paper we present an objective methodology to evaluate the effectiveness of the synthesized output to generate the appropriate prosody in order to confirm the subjective perception tests.

The rest of the paper is outlined as follows. Section II briefly describes standard Malay speech and the correlation between prosodic features and emotions in Malay speech is explained in section III. The prosodic feature extraction and analysis of recorded Malay speech samples are given in Section IV. Section V describes the overall system including the PPM and the TPM approaches that are integrated in the system. Section VI describes the evaluation of the effectiveness of the synthesized output. Finally in Section VII the results of the perception tests conducted on the synthesized emotional speech are presented.

II. STANDARD MALAY SPEECH

Malaysians use many different languages in their daily communication mainly Malay, English, the various Chinese dialects, Indian languages and other languages. Malay is the most widely spoken language in Malaysia. The language has many dialects, but the focus in this study is on Standard Malay (SM), which is the language used in formal contexts such as in education and the mass media [9]. SM has a total of six vowels and the syllables structure of a Malay word can have a combination of CVC arrangement (consonant vowel consonant), CV (consonant vowel) and V (vowel) only [10]. The most common form of syllable structure in SM is bi-syllabic and tri-syllabic which in total makes up 97.52% of Malay words [11,12].
III. CORRELATION BETWEEN PROSODIC FEATURES AND EMOTION

Many studies have shown that there is a correlation between emotional dimensions in speech and prosodic features such as pitch, duration and intensity [13,14]. Generally, if the prosodic features were to be ranked according to their contribution, changes in pitch contribute the most to the expression of emotion while intensity is least important. This study analyzed four basic emotions, namely anger, happiness, sadness and fear which were analyzed by taking into account the changes in the F0 contour, in view of the fact that changes in emotion in speech are reflected mainly by changes in the F0 contour [15].

The correlation between acoustic features and emotion in Malay speech was ascertained by a detailed analysis of a recorded human speech corpus. This study actually makes use of two different corpora. For the TPM approach, a set of 16 sentences each consisting of four words with combinations of two or three syllables in a CVC arrangement were used to record the four basic emotions. In view of the syllable structure of Malay words, it was unfortunately not always possible to come up with meaningful emotional sentences. For the PPM approach, 40 sentences for each emotion were constructed with different lengths and different combinations of syllable structure as described in Table I. A single professional actor was used for the recording of the emotions.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>No of words in a sentence (x)</th>
<th>Number of syllable in a word (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPM</td>
<td>4 words</td>
<td>2 and 3 syllables</td>
</tr>
<tr>
<td>PPM</td>
<td>Less than 4 words</td>
<td>any</td>
</tr>
<tr>
<td></td>
<td>4 words</td>
<td>Not 2 and 3</td>
</tr>
<tr>
<td></td>
<td>More than 4 words</td>
<td>any</td>
</tr>
</tbody>
</table>

The recorded speech was subjected to a listening test in order to ensure the accuracy of the emotional expression produced. The evaluators were divided into two groups consisting of native speakers and non-native speakers of Malay. From the four possible emotions given for each sentence, the listener was required to choose one to represent the emotion being expressed in the sentence. The listening test revealed that a majority of listeners can accurately identify the intended emotive elements in recorded speech. The results of the listening test for the four emotions for TPM and PPM approaches are displayed in a confusion matrix as given in Table II where the non-italicized and italicized numbers represent utterances employing the TPM and PPM approaches respectively.

IV. PROSODIC FEATURE EXTRACTION AND ANALYSIS

The recorded speech with an acceptable degree of recognition was extracted for its acoustic features. First, the recorded speech was manually segmented at phoneme level. This task can be carried out automatically, but it is known to be less accurate than manual segmentation [16]. The task was carried out using PRAAT, a feature extraction tool with segmentation capability [17]. The segmentation of individual phonemes was carried out by listening carefully to the recording and marking the boundary of each phoneme. Each individual phoneme is measured for its duration and F0 contour. Different emotions have different associated F0 contours and duration.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Speech</th>
<th>Happy</th>
<th>Angry</th>
<th>Fear</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
</tr>
<tr>
<td>Happy</td>
<td>87.20</td>
<td>89.30</td>
<td>87.00</td>
<td>5.80</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>89.30</td>
<td>89.30</td>
<td>5.80</td>
<td>2.50</td>
<td>2.40</td>
</tr>
<tr>
<td>Angry</td>
<td>3.80</td>
<td>89.80</td>
<td>2.80</td>
<td>2.50</td>
<td>4.40</td>
</tr>
<tr>
<td></td>
<td>3.10</td>
<td>92.50</td>
<td>1.20</td>
<td>1.20</td>
<td>3.20</td>
</tr>
<tr>
<td>Fear</td>
<td>3.90</td>
<td>2.00</td>
<td>78.60</td>
<td>15.50</td>
<td>11.40</td>
</tr>
<tr>
<td></td>
<td>2.80</td>
<td>3.70</td>
<td>82.10</td>
<td>11.40</td>
<td></td>
</tr>
<tr>
<td>Sad</td>
<td>3.80</td>
<td>2.50</td>
<td>19.70</td>
<td>74.80</td>
<td>79.70</td>
</tr>
</tbody>
</table>

Anger utterances tend to have shorter duration and F0 contours that rise and fall at the beginning and end of the sentence. The F0 contours for happy utterances generally rise and fall at the beginning of the sentences and rise at the end of the sentence. Fear utterances tend to have shorter duration and F0 contours that fall at the end of the sentence, whereas sadness is associated with longer duration and falling F0 contours at the beginning of the sentences and rising contours at the end of the sentence.

The duration of each phoneme in utterances was measured in milliseconds. Vowels were not only measured for duration but also for pitch (measured in Hz) at certain significant points marking major changes in the pitch factor. Only the vowels were analyzed for pitch because vowels are the most pitch-significant phonemes within a word [18]. The extracted prosody parameters of the speech utterances were directly compared with emotionless voice concatenated using Fasih, a Malay text-to-speech synthesizer that uses MBROLA, a diphone-concatenative synthesizer [19]. The differences in prosodic features between the emotional speech extract and the neutral output of Fasih were then expressed as a percentage increase or decrease to formulate the dpmr (duration for parametric manipulation rate) and ppmr (pitch for parametric manipulation rate) factors using the following equation:


\[ dp_{mr} = 1 + \frac{(d_s - d_r)}{d_s}, \quad (1) \]

\[ pp_{mr} = 1 + \frac{(p_s - p_r)}{p_s}, \quad (2) \]

where

- \( d_s \): neutral duration
- \( d_r \): recorded duration
- \( p_s \): neutral pitch
- \( p_r \): recorded pitch

Table III illustrates how the percentage difference was derived for the word ‘Saya’ which means ‘I’. In the table, ‘d’ denotes duration and ‘p’ represents pitch.

**Table III**

<table>
<thead>
<tr>
<th>Standard (ms)</th>
<th>Recorded (Hz)</th>
<th>Percentage Differences</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>p</td>
<td>d</td>
<td>P</td>
</tr>
<tr>
<td>s</td>
<td>92</td>
<td>nil</td>
<td>99</td>
</tr>
<tr>
<td>a</td>
<td>82</td>
<td>230</td>
<td>168</td>
</tr>
<tr>
<td>y</td>
<td>92</td>
<td>nil</td>
<td>94</td>
</tr>
<tr>
<td>a</td>
<td>101</td>
<td>220</td>
<td>176</td>
</tr>
</tbody>
</table>

The synthesizer only generates pitch values for vowels because Malay speech utterances focus on the vowel for the proper articulation of emotions. All recorded sentences for each type of emotion were subjected to the same process of extraction, comparison and formulation of differential rate.

V. System Overview

The emotional synthesizer reported in this paper combines two approaches. The TPM approach is used for text input made up of four words with a combination of two and three syllables. Any other forms of input texts are synthesized by the PPM approach. When users input text to the emotion synthesizer and select the type of emotion to be synthesized, the synthesizer chooses the approach by counting the number of words and syllable structure of the input text. Fig. 1 shows the functional diagram of the emotions prosody generator.

**A. Template Parametric Manipulation (TPM) Approach**

This approach replaces the entire standard pitch and duration parameters with emotional parameters by applying the dp_{mr} and pp_{mr} value available for each phoneme in the prosody database. Each type of syllable combination has its own parameters. Altogether, there are sixteen possible combinations of two and three syllables for a four words sentence, meaning that there are sixteen templates for each type of emotion. The overall function of this approach is illustrated in Fig. 2. For synthesizing emotional speech using the TPM approach, the first step is to determine the syllable structure for selecting the appropriate prosody template to be used.

Next, the neutral prosody parameter is multiplied with the available dp_{mr} and pp_{mr} factor in the prosody database template to generate emotional prosody. The replacement of the prosodic parameter is carried out at the level of the individual phoneme. The syllable format of the prosody database consists of a full-CVC arrangement. The text input, however, can have a combination of CVC, V or CV. If the input text has CV structure, the emotional synthesizer only replaces the phonemes available in the template and ignores any phonemes that do not have a match. For example, the word “ayah” (a-yah) has V structure followed by CVC. The replacement for the first syllable is based only on the vowel, and the second syllable is based on the entire CVC prosody. Since there is no third syllable, the process ends at the second syllable and then moves on to the next word. This process is illustrated in Fig. 3.
B. The Parametric Prosody Manipulation (PPM) Approach

For the other form of text input, the emotion synthesizer uses the PPM approach. For synthesizing emotions using this approach, the first step is to isolate and identify individual phonemes as consonants or vowels. The synthesizer manipulates the duration of each phoneme according to the dpmr factor available for each type of emotion. Vowels that were identified were manipulated not only for duration but also for pitch using the available ppmr factors, as illustrated in Fig. 4.

![Diagram of the Parametric Prosody Manipulation approach]

VI. EVALUATION OF THE SYNTHESIZED OUTPUT

The effectiveness of the emotions synthesizer depends on its ability to generate the appropriate prosody (pitch and duration). An evaluation by comparing the prosodic pattern of the recorded utterances with the prosodic pattern of the synthesized output for both the full CVC structure and other forms of syllable structure were carried out. Three types of utterance were used for the evaluation. For the recorded and full CVC utterance, the same sentence *Borhan ternampak bintang ganjil* was used. The F0 movements of the recorded and full CVC synthesized sentences were illustrated in Fig. 5 and Fig. 6. It shows that the synthesizer can produce F0 movements similar to those of the recorded utterances. The duration of the full CVC synthesized speech is equivalent to 97.6% of the recorded speech utterance.

For evaluating the effectiveness of the synthesizer to synthesize non uniform CVC structure, the sentence *Lina melawat rumah jamil* with its CV-CV CV-CV-CVC CV-CVC arrangement was synthesized using the emotion synthesizer. When synthesizing sentences with syllabic structures other than CVC, the F0 pattern was still the same despite a shorter duration in view of the smaller number of consonants in the synthesized sentence.

To further evaluate the effectiveness of the synthesizer, the complete listing of F0 at regular interval for the three utterances were extracted and analyzed. The F0 factor was tabulated and accuracy percentage (Ap) was determined using the following formula:

\[
Ap = \left(1 - \frac{|Pr - Ps|}{Pr}\right) \times 100\% ,
\]

where

\[
Pr = \text{Pitch of recorded utterance},
\]

\[
Ps = \text{Pitch of synthesized utterance}
\]

For example, the first vowel has starting pitch factor of 388 Hertz for the recorded utterance and 387 Hertz for the synthesized utterance. The difference between the recorded and synthesized is 0.026%. This mean the accuracy of the first vowel at the starting point is 99.74%. This shows that the synthesizer can replicate the prosodic pattern of recorded utterances with high accuracy to synthesis emotional speech similar to the recorded human utterances. Similarly, the comparison at regular intervals of F0 factors for the full CVC synthesized sentence and the non-CVC synthesized sentence shows an accuracy of 99.46%. This shows that the synthesizer can produce accurate prosodic factors for different syllable structures. Overall, the accuracy level for all 64 sentences is 97.43% between recorded and synthesized, and 99.05% between full CVC structure and random syllable structure.
Table IV shows the F0 factor for each of the three utterances. The table shows the extracted F0 factor for each vowel for each of the three utterances. The comparison between the recorded utterance and the synthesized utterance for the sentence “Borhan ternampak bintang ganjil” shows an accuracy level of 98.96%.

Fig. 7 and 8 illustrate the F0 movement for full CVC and random syllable structures at regular time intervals for each vowel in the respective sentences. Despite the timing differences of same sentence with different syllable structures, the F0 movement at each vowel interval remained the same.

The synthesized outputs for the two approaches were tested separately. For the TPM approach, a listening test identical to the earlier test was carried out. Sixteen new sentences for each type of emotion were constructed with different syllable arrangements and structures. These sentences were synthesized using the synthesizer and were converted to waveform files. All sound files were renamed and arranged randomly. Each participant listened to the voice extract and attempted to identify the emotion intended using forced choice answer sheets.

The perception tests carried out revealed that the recognition rate for anger was highest at 82.67%, followed by happiness at 80.33%. The recognition rate for sadness and fear was much lower at 68.33% and 65.69% respectively. Table V tabulates the result of the perception test.

For the PPM approach, the participants input their own sentences with a variety of lengths and syllable structures other than 16 sentence structures corresponding to the prosody templates.

The perception test conducted with 10 participants shows that the synthesizer can produce recognizable emotional synthesis for any given input text. Again, happiness and anger show the highest acceptability whereas sadness and fear have the lowest acceptability rate.

VII. TESTING AND RESULTS

![Figure 8. F0 factor- Synthesized (random structure).](image-url)

![Figure 7. F0 factor- Synthesized (full CVC structure).](image-url)
Table VI shows the confusion matrix for which the listeners perceived the synthesized output. Each type of emotions has 100 sentences being synthesized and the listeners determined whether they agree or disagree with the synthesized output. Anger has the highest rate of agreement in that 81% of input sentences were recognized as expressing anger while fear had the lowest recognition rate of 65% and was sometimes mistaken for sadness, anger, ecstasy or disgust.

**TABLE VI**
The confusion matrix of perception test for the PPM approach

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Perceived as</th>
<th>Synthesized as</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happy (%)</td>
<td>Angry (%)</td>
</tr>
<tr>
<td>Happy</td>
<td>77.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Angry</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Sad</td>
<td>1.00</td>
<td>4.00</td>
</tr>
<tr>
<td>Fear</td>
<td>19.00</td>
<td>14.00</td>
</tr>
</tbody>
</table>

**VIII. DISCUSSION AND CONCLUSIONS**

The prosodic parametric generation module developed in this research employs template database and rule-based prosodic generation to improve the intonation variability of the synthesized output. The testing of the generation module in both the subjective listening test and the objective accuracy evaluation shows that the module can generate the appropriate prosodic features to synthesize emotional expression for any given text input. The ability of the synthesizer to generate the intended emotion with accuracy was generally satisfactory, judging by the listening perception tests used in this research. Users were generally satisfied with the accuracy of the output as well as with the variation in speech style.

**ACKNOWLEDGMENT**

This work is supported by a Science Fund grant from Ministry of Science, Technology and Innovation of Malaysia.

**REFERENCES**


