interspeech 2008 incorporating SST 08

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

Mumtaz Begum [1], *Raja N. Ainon* [1], *Roziati Zainuddin* [2], *Zuraidah M. Don* [3], *Gerry Knowles* [4]

[1] Department of Software Engineering, University of Malaya, Malaysia
[2] Department of Artificial Intelligence, University of Malaya, Malaysia
[3] Department of Linguistics, University of Malaya, Malaysia
[4] Miquest Worldwide Sdn. Bhd., Malaysia

mumshaka4@perdana.um.edu.my, ainon@um.edu.my, roziati@um.edu.my, zuraida@um.edu.my, g.knowles@lancaster.ac.uk

10.21437/Interspeech.2008-91

Abstract

The manipulation of prosody, including pitch, duration and intensity, is one of the leading approaches in synthesizing emotion. This paper reports work on the development of a Malay Emotional synthesizer capable of expressing four basic emotions, namely happiness, anger, sadness and fear for any form of text input with various intonation patterns using the prosody manipulation principle. The synthesizer makes use of prosody templates and prosody parametric manipulation for different types of sentence structure.

Index Terms: concatenated emotion synthesis, prosody templates, prosody parametric manipulation

1. Introduction

Prosody includes a combination of pitch, duration and intensity, and in natural speech prosodic features can be manipulated to express different emotions [1]. It is typically described in terms of fundamental frequency contour (F0), duration of speech segments (phoneme or syllable) and intensity (loudness) [2]. These features are also important in speech synthesis, and they can be manipulated to produce the required emotion. Different methods can be used to generate the necessary prosody, but the most commonly used approach is the rule-based approach. This approach was used to generate prosody for synthesizers that produced voice output digitally in formant and diphone based synthesizers [3, 4].

This paper aims to enhance the quality of the rule-based approach by integrating prosody parametric manipulation (a form of rule-based approach) with prosody database templates [5, 6]. It is hoped that the integrated approach will improve the output prosody for different forms of text input.

2. Malay Speech

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 mass media [7]. 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 [8]. 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 [9].

3. Emotional Expression in Malay Speech

Although there are differences, the expression of emotion in Malay is generally similar to the expression of emotion in English [10, 11]. Expressions of emotion in speech are determined by variation in prosodic features such as pitch, duration intensity and voice quality. 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 [12]. This study analyzed four basic emotions, namely anger, happiness, sadness and fear. These emotions were analyzed by taking into account the changes in the F0 contour, particularly changes in pitch level and pitch range. The view taken here is that changes in emotion in speech are reflected mainly in changes in the F0 contour [13].

4. Acoustic Features Correlation in Malay Speech

The correlation between the acoustic features and emotions in Malay speech were determined by a detailed analysis of the recorded speech corpora. This study actually makes use of two different corpora. For the prosody templates approach, a set of 16 sentences comprising four words in a sentence with combinations of two or three syllables in a CVC arrangement (Consonant Vowel Consonant) were constructed and 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 parametric prosody manipulation approach, 40 sentences for each emotion were constructed with different lengths and different combinations of syllable structure (2, 3, 4, 5 or 6 syllables in a word) as described in table 1 below. A single professional actor was used for the recording of the emotions.

Table 1. *The syllable pattern for each approach.*

Approaches	No of words in a sentence (x)	Number of syllable in a word (y)
Prosody templates	**$x = 4$**	**$y = 2$, $y = 3$**
Parametric Prosody	**$x < 4$**	**$2 \leqslant y \leqslant 6$**
	$x = 4$	**$y \neq 2$, $y \neq 3$**
	$x > 4$	**$2 \leqslant y \leqslant 6$**

4.1. Subjective Evaluation of Recorded Speech

A subjective evaluation of the recorded speech was conducted to ensure the accuracy of the recorded emotional expression. 20 evaluators equally divided into two groups: native speakers and non-native speakers of Malay took part in this listening test. For each of the recorded sentence, the listener is required to choose one of the four basic emotions that represent the emotion expressed in the sentence.

Despite the lack of meaning in some recorded speech, the listening test revealed that a majority of listeners can accurately identify the intended emotive elements in the recorded speech out of which native Malay speaker have better recognition than non native Malay speaker. The overall result of the perception test indicates that anger has the highest rate of recognition at 91.15% followed by happiness at the rate of 88.25%. Sadness and fear had much lower rates at 80.35% and 77.25% respectively. Table 2 below shows the result of the confusion matrix of the prosody template sentences. Table 3 shows the perception result of sentences designed for parametric prosody manipulation. Generally the recognition rate for parametric prosody manipulation was higher in view of the emotional meaning associated with the syntax.

Table 2. *The confusion matrix of template-based sentences.*

	Perceived as			
Emotion speech	Happy (%)	Angry (%)	Fear (%)	Sad (%)
Happy	**87.20**	8.70	2.50	1.60
Angry	3.80	**89.80**	4.40	2.00
Fear	3.90	2.00	**78.60**	15.50
Sad	3.00	2.50	19.70	**74.80**

Table 3. *The confusion matrix of rule-based sentences.*

	Perceived as			
Emotion speech	Happy (%)	Angry (%)	Sad (%)	Fear (%)
Happy	**89.30**	5.80	2.50	2.40
Angry	3.10	**92.50**	1.20	3.20
Sad	2.80	3.70	**82.10**	11.40
Fear	3.80	2.30	14.20	**79.70**

4.2. Acoustic Feature Extraction and Analysis

The recorded speech with 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 [14]. The task was carried out using PRAAT, a feature extraction tool with segmentation capability [15]. 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. 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 value. Only the vowels were analyzed for pitch because vowels are the most pitch-significant phonemes within a word [16]. For the purposes of the prosody template, the exact pitch and duration of each voiced phoneme were extracted and saved in the prosody templates. For the parametric prosody manipulation approach, extracted prosody parameters of speech utterances were directly compared with emotionless voice concatenated using the diphone concatenative synthesizer. The synthesis is based on MBROLA (Multi-Band Re-synthesis Overlap-add) that modifies the pitch and duration of input speech [17]. The differences in prosody between the emotional speech extract and the neutral output were then expressed as a percentage increase or decrease using the following formula:

$$\mathbf{dpmr} = \mathbf{b_o} + \frac{(\mathbf{d_s} - \mathbf{d_r})}{\mathbf{d_s}}$$

$$\mathbf{ppmr} = \mathbf{b_o} + \frac{(\mathbf{p_s} - \mathbf{p_r})}{\mathbf{p_s}}$$

Where
dpmr: duration for parametric manipulation rate
ppmr: pitch for parametric manipulation rate
b_o: base = 1 (100% of standard synthesizer parameters)
d_s: neutral duration
d_r: recorded duration
p_s: neutral pitch
p_r: recorded pitch

Table 4 and 5 below illustrate how the percentage difference was derived for the word *"saya"* which means *I*.

Table 4. *Comparing standard synthesizer's duration with the recorded utterance to derive dpmr value.*

	Standard duration (ms)	Recorded duration (ms)	Percentage Differences (%)	dpmr
S	92	99	8%	1.08
A	82	168	105%	2.05
Y	92	94	2%	1.02
A	101	176	74%	1.74

Table 5: *Comparing standard synthesizer's pitch with the recorded utterance to derive ppmr value.*

	Standard duration (ms)	Recorded duration (ms)	Percentage Differences (%)	ppmr
S	Nil	nil	nil	nil
A	230	370	161%	2.61
Y	Nil	nil	nil	nil
A	220	335	152%	2.52

The synthesizer only generates pitch value for vowels because Malay speech utterances focus on vowel for 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. An average dpmr and ppmr value was calculated for each type of emotion. The emotional synthesizer focuses the prosody manipulation at the beginning and at the end of the sentence because much of the F0 movement occurred at the beginning and end of a sentence.

5. System Architecture

The emotional synthesizer reported on in this paper combines two approaches. Prosody templates are 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 parametric prosody manipulation. When users input text into 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. Figure 1 below shows the flowchart of the emotional synthesizer.

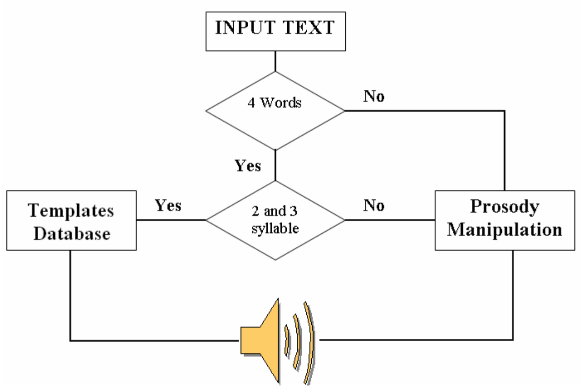

Figure 1: *Flowchart of the emotional synthesizer decision and synthesizing process.*

5.1. Prosody Templates Approach

This approach replaces the entire standard pitch and duration parameter with emotional parameters available in the prosody database. This differs from Syaheera's diphones-based template approach where the prosodic parameters are based on template sentence [6] whereas in this approach the prosodic parameters are defined at individual phoneme level. Each type of syllable combination has its own parameters. Altogether, there are sixteen possible combinations of two and three syllables, which mean there are sixteen templates for each type of emotion. The overall function of this approach is illustrated in figure 2 below.

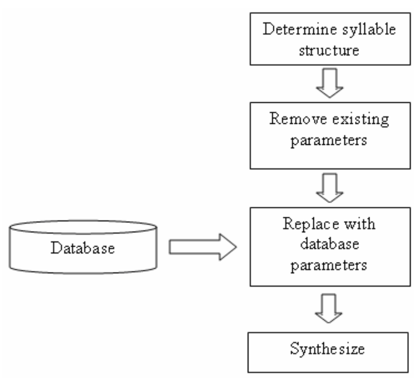

Figure 2: *Overall function of the prosody templates approach.*

For synthesizing emotional speech using the templates approach, the first step is to determine the syllable structure for selecting the appropriate prosody parameter to be used. Next, the standard prosody parameter is removed and replaced with available parameter in the prosody. The replacement of parameter is on individual phoneme level. The syllable format of the prosody consists of a 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 figure 3 below.

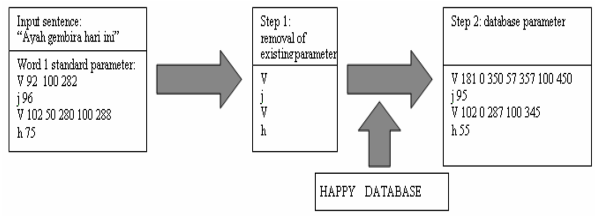

Figure 3: *The prosody template replacement process.*

5.2. Parametric Prosody Manipulation

For the other form of text input, the emotion synthesizer uses parametric prosody manipulation. For synthesizing emotion using this approach, the first step is to isolate and identify individual phonemes as consonants or vowels. The emotion 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, as illustrated in diagram 4 below.

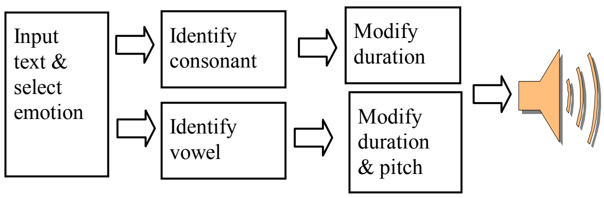

Figure 4: *The overall process of parametric prosody manipulation.*

6. Testing and Results

The synthesized outputs for the two approaches were tested separately. For the prosody template 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 snippet and determined the emotion intended through forced choice answer sheets. The perception test 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 6 below tabulates the result of the perception test.

Table 6: *The confusion matrix of perception test.*

	Perceived as			
Emotion speech	Happy (%)	Angry (%)	Sad (%)	Fear (%)
Happy	**80.33**	4.69	7.81	7.17
Angry	0.46	**82.67**	4.06	12.81
Sad	4.19	5.31	**68.31**	22.19
Fear	2.32	20.32	11.67	**65.69**

For the parametric approach, the listening test was conducted by use of acceptance test in which participants input their own sentences with a variety of lengths and syllable structures but not the word length and syllable structure of the PDT approach. Each participant input ten different sentences for each type of emotion and listened to the synthesized output. If the subjects judged the synthesized form to be acceptable emotional output, they marked the 'agree' column on the answer sheets. Otherwise, they marked the 'disagree' column and determine the type of emotion that they perceived when they listened to the synthesized output.

The acceptance test conducted with 20 participants shows that the synthesizer can produce recognizable emotional synthesis for any given input text. From total of 200 input sentences for each emotion, anger has the highest conformation rate at 81.00 % followed by happiness at 77.00%. Sad acceptance rate is at 69.00% and fear at 65.00% as illustrated in figure 5 below.

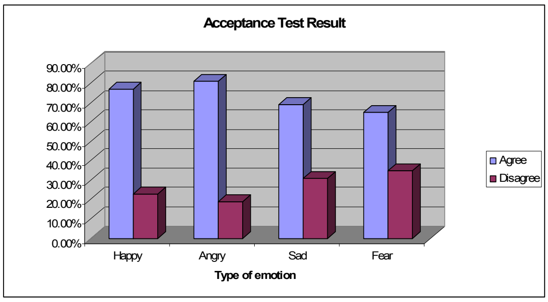

Figure 5*: The acceptability of synthesized output.*

To enable comparison between the two approaches, a confusion matrix was constructed for the acceptance test based on the interpretation of participants during the acceptance test. Table 7 below shows the confusion matrix for the PPM approach based on interpretation of participants.

Some participants interpreted the synthesized anger as fear or ecstatic. For happiness, it is sometimes perceived as excitement. Sadness is confused as an expression of fear, boredom, or disgust. Finally, fear is mistaken for sadness, anger, ecstasy or disgust.

Table 7. The confusion matrix of acceptance test.

	Synthesized as			
Perceived as	Happy	Angry	Sad	Fear
Happy (%)	77.00	0.00	1.00	1.00
Angry (%)	2.00	81.00	3.00	4.00
Sad (%)	1.00	1.00	69.00	5.00
Fear (%)	1.00	4.00	6.00	65.00
Others (%)	19.00	14.00	21.00	25.00

7. Discussion

Prosody manipulation has been given credit for its ability to generate emotional synthesized speech. The most commonly used manipulation technique is the rule-based approach, in which prosody values are generated by some sort of numerical processing, which can be applied either at phoneme level or at sentence level. Little emphasis has been placed on prosody generation using database templates, in view of the complexity of the input text. Database templates produce better variations in output than the rule-based approach. While each approach has its own advantages, the combination of the two approaches should improve the prosodic quality of the synthesized output.

8. Conclusions

The result shows that the ability of the synthesizer to accurately generate the right kind of emotion was generally satisfactory, judging by the listening and acceptance tests used in this research. Users were satisfied with the accuracy of the output as well as with the variation in speech style.

9. References

[1] Malfrère, F., Dutoit, T., Mertens, P., "Fully automatic prosody generator for text-to-speech", In: Proceedings of the International Conference on Spoken Language Processing 98, Sydney, Australia, paper 355, 1998.

[2] Mozziconacci, S.J.L., "Prosody and Emotions", in Proceedings of Prosody 2002, Aix-en-Provence, France, 2002.

[3] Mumtaz, B., "Adding an Emotions Filter to Malay Text-to-Speech System", Master Dissertation, Univ of Malaya, 2006.

[4] Burkhardt F.: "Emofilt: The Simulation of Emotional Speech by Prosody-Transformation," Proc. of Interspeech, 9th European Conference on Speech Communication and Technology, Lisboa, 2005.

[5] Kochanski, G. and Shih, C., "Prosody Modeling with Soft Template", Speech Comm, 39: 311-352, 2003.

[6] Syaheerah L.L., "Adding Emotions to Synthesized Malay Speech using Diphone-based Templates", Master Dissertation, Univ of Malaya, 2007.

[7] Knowles, G.O. and Zuraidah, M.D., "Word Class in Malay", Dewan Bahasa and Pustaka, Kuala Lumpur, Malaysia, 2006.

[8] Kee, T.Y., Seong, T.B. and Haizhou, L., ""Grapheme to Phoneme Conversion for Standard Malay", 2004.

[9] Farid, M.O., "Aspects of Malay Phonology and Morphology-Generative approach", Bangi, University Kebangsaan Malaysia, Press, 1980.

[10] Yunus, M., "The Malay Sound System", Kuala Lumpur, Fajar Bakti Sdn. Bhd. Malaysia, 1980.

[11] Chomsky, N. and Halle, M., "The Sound Pattern of English", Harper and Row, New York.

[12] Vroomen, J., Collier, R. and Mozziconacci, S., "Duration and Intonation in Emotional Speech", Proceedings of Eurospeech, Berlin, Germany, pp. 577-580, 1993.

[13] Wentao, G. and Lee, T., "Quantitative analysis of F0 contours of emotional speech of Mandarin", In *SSW6-2007*, Preservation, Identification and use of emotion in a TTS system, 228-233, 2007.

[14] Eide, E., "Preservation, identification and use of emotion in a text-to-speech system", IEEE Speech Synthesis Workshop, Sep, 2002.

[15] Boersma, P. and Weenink, D., Praat, doing phonetics by computer, http://www.fon.hum.uva.nl/praat/, Mar2004, Visited: Mar 2004.

[16] Onn, K.W., "Predicting Pitch Movement within Malay Sentences", at http://infocenter.mimos.my/admin/pdf/MS2005-08-KowWengOnn.pdf, 2005.

[17] MBROLA project homepage at http://tcts.fpms.ac.be/synthesis/mbrola.html