Template-driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment

Syaheerah L. Lutfi  
Raja Noor Ainon  
Salimah Mokhtar  
Software Engineering Department  
Faculty of Computer Science and IT  
University Malaya  
Kuala Lumpur, Malaysia  
syaheerah@perdana.um.edu.my, ainon, smazz@um.edu.my

Zuraidah Mohd Don  
Faculty of Language and Linguistics  
University Malaya  
Kuala Lumpur, Malaysia  
zuraida@um.edu.my

Abstract - This paper describes the pilot experiment conducted for the purpose of adding an affective component to the first Malay Text-to-Speech (TTS) system, Fasih. The aim is to test a new method of generating an expressive speech via a template-driven system based on diphones as the basic sound. The synthesized expressive speech can express four types of emotion. However, as an initial test the pilot experiment focused on anger and sadness. The results from this test show an impressive recognition rate of over 60% for the synthesized speech of both emotions. The pilot experiment has paved the way for the development of an emotions filter to be embedded into Fasih, thus allowing for the possibility of generating an unrestricted Malay expressive speech.

Keywords: Text-to-Speech, emotions, template-driven, diphones, synthesized speech, MBROLA, vocal affect, Malay.

1 Introduction

Adding emotions to a synthesized speech means that the latter can verbalise language with the kind of emotion appropriate for a particular occasion (e.g. announcing bad news in a sad voice). Speech articulated with the appropriate prosodic cues can sound more convincing and may catch the listener’s attention, and in extreme cases, it can even avoid tragedies [1]. An improved synthesized speech can also benefit from other speech-based human-machine interaction systems that perform specific tasks like read-aloud texts (especially material from the newspaper) for the blind, weather information over the telephone, auditory presentation of instructions for complex hand free tasks [2], story book narration and proof reading.

This paper gives an account of the findings of a preliminary experiment that attempts to produce two human emotions, namely anger and sadness. The purpose here is to test the viability of adding an emotions filter to FASIH\(^1\), the first Malay TTS system.

---

\(^1\) Fasih is developed by MIMOS Berhad, and is continually upgraded with the help of linguistic information provided by Gerry Knowles and Zuraidah Mohd Don.

2 Background

2.1 Vocal affect

Emotions can be conveyed verbally or non-verbally in communication. In this article we examine the verbal expression of anger. For example, slower, lower-pitched speech, with little high frequency energy, generally conveys sadness, while louder and faster speech, with strong high-frequency energy and more explicit enunciation, typically accompanies joy [3]. Interest in the mechanism of human speech has led to increased research on speech and findings from these studies have now been applied to a synthesized speech using electronic devices.

There are two approaches to the study of vocal expressions of emotions: perception-oriented and acoustic-oriented [4,5]. Perception-oriented is listener centered, and is concerned with how listeners perceive emotions, whereas acoustic-oriented is speaker centered and is concerned with analysing the vocal parameters of expressed speech that is linked to emotions. Events in the waveform include both vowels and consonants and also prosody. Prosody includes such parameters as pitch, duration and loudness. Intonation refers in the first instance to pitch. Accent is stress associated with a pitch movement in connected speech. [18]

2.2 Vocal affect in relation to anger and sadness in Malay perspective.

The expression and perception of emotions may vary from one culture to another [8]. In an attempt to develop emotions in a synthesized speech which are indistinguishable from human emotions, as outlined in the project goal, characteristics of anger and sadness in Malay culture are studied. However, there is not much literature on vocal emotions in Malay, and so far we have not come across any study on anger and sadness from a Malay perspective.

In an attempt to understand emotions from the Malay perspective especially with regard to anger and sadness,
we refer quite substantially to the work by Wazir Jahan [9]. According to her the description of the emotions in Malay was not based on empirical research but based on passing observations and intuitive reasoning. She concedes that many studies have been carried out on *latah* (for women) and *amuk* (for men, English *amok*), since these two expressions of emotion are closely related to the understanding of the ‘Malay mind’ then brought about by rebellious reactions against colonization. Wazir Jahan examines the observations of the Malay mind by several western anthropologists who believe that the Malay people look ‘externally impassive’ but are actually very sensitive even to something as normal as ‘the accidents of every day life’. Evidence gathered from past observations seem to show that the Malays are inclined to keep their emotions in check until the time when they cannot contain them anymore and that is when they explode. These observations seem to be in line with what is expressed by the former Prime Minister, Tun Dr. Mahathir in his book *The Malay Dilemma*, “the transition from the self-effacing courteous Malay to the amok is always a slow process. It is so slow that it may never come about at all. He may go to his grave before the turmoil in him explodes” [10]. In this article we are not interested in the phenomenon of amok in itself but in its expression since it bears elements of a culturally specific form of anger.

A study carried out by Silzer [8] illustrates that the expression of human emotions are cultural specific, e.g., how anger is expressed in English is different from how ‘marah’ is expressed in Malay. He explains that the causal component of marah is more specific such that marah “is the result of intentional personal offence, where the offender knowingly committed the "bad" act, while realizing it would cause the other person unpleasant feeling”. This causes the offended party to inform the offender in a certain tone of voice that he or she has done something wrong, and should rectify the problem. It is also observed that when expressing anger, Malays are inclined to shout. This way of expressing anger could probably be caused by the accumulation of negative feelings which when released manifest in the form of shouting or yelling.

From our preliminary studies we find that Malay utterances when uttered in anger tend to have a slightly higher overall pitch while sadness is accompanied by lower overall pitch when compared to English utterances [4].

### 3 Emotions Filter

Figure 1 shows the proposed architecture to "emotionalize" Fasih. Fasih uses an MBROLA synthesizer [11] which uses a diphone concatenation method to synthesize speech. Diphone concatenation systems use mainly two parameters of artificial voice to produce synthetic speech, namely pitch and duration.
tells the synthesizer to produce the sound for phone $s$ of 122 ms, and to put a pitch point of 320 Hz at 50% of 122 ms and a pitch pattern point of 324 Hz at the end of the sound. Pitch pattern points define a piecewise linear pitch curve. Notice that the pitch pattern defined is continuous, since the program automatically drops pitch information when synthesizing unvoiced phones. For pause, underscore is inserted followed by the duration of pause in ms, for example “_1000” is a pause for 1000 ms.

The synthesizer outputs chunks of synthetic speech determined as sections of the piecewise linear pitch curve. Phones inside a section of this curve are synthesized in one go. The last one of each chunk, however, cannot be properly synthesized while the next phone is not known (since the program uses diphones as base speech units).

The complete accurate parameters inserted for the whole utterance could produce speech that has the correct tones for the correct type of emotional state, and we regard each complete input as an emotion template.

### 3.3 Recording of speech

First, we constructed a few semantically neutral sentences; i.e. sentences that could be uttered in any emotion. But for the purpose of the pilot test, we only used two declarative sentences: “Saya nak yang itu” (I want that one) and “Itu kepunyaan aku” (That is mine), which both convey anger and sadness. Recording was conducted in the Cultural Centre drama training room in the University of Science Malaysia (USM). We recorded the speech using the voices of two semi-professional actors using the SOMIC SM-350 head worn microphone. The recorded sound samples were directly stored in Praat [17].

### 3.4 Segmentation and labeling of data

The purpose of annotating (segmenting and labelling) the speech samples is to extract the prosodic features which are involved in the expression of emotions. The speech samples were first pre-processed, which involved cleaning them from noise using Speech Filing System (SFS) [12]. Then, the phonetic segmentation and alignment was performed manually using Praat. The recorded speech samples were annotated at three levels – words, syllables and phonemes in the Praat text grid editor, as shown in Figure 3, to ensure accuracy during the extraction of emotion features.

![Figure 3: Three-layer annotation of the utterance “saya nak yang itu”](image)

Specifically, each segmented and labelled chunks of the speech sample was analyzed to extract their duration and significant pitch points. The information was later inserted as the parameters of each phone following the SAMPA phonetic transcription in Mbrola. When the speech is played, the phones were read at one go (concatenated) which produced synthesized speeches that have similar tones to the original speech recorded.

### 3.5 Template-driven Emotion Generation

The templates store various tones that reflect a particular emotional state. We need an algorithm that allows an input sentence from a user to be read using the appropriate tones which convey the emotional state the user chooses (click on a particular ‘emotion’ button). This must be done so that the sentence matches the correct emotion template. To accomplish this, a syllable-sensitive rule-based algorithm is proposed. The simple syllable structure that the Malay language is based on allows for the use of an algorithm that focuses on the number of syllables rather than other linguistic features. In Malay, the syllable structure is as follows:

- CVC
- CV
- VC

Using this algorithm, each word from user input will be analyzed and chunked into syllables in reverse order (stack) to determine syllable count. In other words, the input sentence will be processed from the last word to the first. The result is then matched against the emotion template with the same syllable-count and sequence. Figure 4 provides a visual illustration of the algorithm.

![Figure 4: Simple scheme of the proposed Emotions Filter Module. Adapted from Syaheerah L.Lutfi et al [13]](image)
The program which is based on Java has two assumptions: Words where two consonants or vowels are next to each other which produce one sound are considered to have only one vowel or consonant;
1. $VV=V$, e.g.: the vowels ‘au’ “Pulau” (island), is considered as one vowel and not two.
2. $CC=C$, e.g.: the consonants ‘Ng’ in “Ngeri” (errie) is considered as one consonant.

Nonetheless, this cannot be applied to some Malay loan words that are of English origin, such as “drama” or “trauma” because the first two consonants in both words when combined, actually produce two sounds – ‘de | ra’ and ‘te | ra’.

However, the question that needs to be addressed is that whether this automation is suitable for unrestricted input. Since the parameters given correlate with each phoneme in a word, any words that contain extra or less phonemes in the syllables may cause the whole sentence to sound ‘strange’ when synthesized. For example, both the words “Saya” and “Aku” which both means “I” have two syllables, but ‘Sa | ya’ has two phonemes in the first syllable while ‘A | ku’ has only one. If input “Saya” is matched with a template that contains “Aku”, the extra phoneme /s/ in the first syllable will be left with standard parameters specified. Instead, if the matching is reverse, phoneme /s/ will be set to silence while information of phoneme /a/ will be given to /A/ in ‘Aku’, because both are vowels and assumed to have more or less the same duration.

This particular computational organization is also tested in the preliminary test to have an idea what the sound produced is like. The findings are stated in section 3.7 below.

The acoustic correlates in terms of pitch and duration of the speech samples were inserted and it was discovered that the synthesized speeches sound almost similar to the original speech samples. It appears that the synthesized speeches have varying frequency and “voices”, mimicking the original speaker voices. At this point we realised that to maintain voice consistency it is enough to have only one speaker and this will also ensure the consistency of the sounds stored in the emotions templates.

3.6 Perceptron test

To evaluate the synthesized speech samples, we conducted a simple listening test with 20 Malay speaking adults who were unaware of the identity of the test stimuli. The listeners were allowed to play the test files as many times as they wished and were asked to label the emotional state of the synthesized speech in a ‘forced choice’ option.

3.7 Results and discussions

Figures 5 and 6 below show respectively, the comparison of the original and synthesized recognition rates of the two speech samples.

4 Conclusions and Future Work

The fact that Malay has no stress or tone or any other intermediate level of prosodic organization [16] makes it a language suitable for computational linguistic research. Based on the high percentage of recognition rate, although done manually and without any computational
algorithm, we are quite confident that unrestricted, emotional Malay speech could be produced by Fasih. Nevertheless, we realised that the emotional tone elicited based on a template may be suitable for one input which satisfies the syllable count and sequence, but it may not be suitable for another which has inappropriate content. For example, these two declarative angry sentences “Awak memang sangat biadap” (You are very rude) and “Anak awak suka mencuri” (your child likes stealing) may both suit the syllable checks and the tone elicited by template may go well with the first sentence, but it may not fit the second because of its content.

Currently, we are working on producing unrestricted emotional Malay synthesized speech in a specified scope, by using only sentences with four words and each word only has two or three syllables. The findings mentioned above led the decision to choose utterances that have no conflicting content, and use the voice of one single subject.

In the future, it is hoped that a bigger corpus of Malay speech that contain all syllables and sentence types (declaratives, imperatives, interrogatives) is developed to assist in improving the naturalness of the artificial speech. Since Fasih does not only have the ability to count syllables, but also to recognize the category of words (noun, verb etc.), the corpus should also be expanded to cover all possibilities of the positions of word correlating to their categories. This would help eliminate the limitation mentioned. The right emotional tones are more likely to fit the sentences with similar positions of word categories.

5 References


