eXpressive Text Reader Automation Layer (eXTRA): Template-driven 'Emotion Layer' in Malay Concatenated Synthesized Speech

Syaheerah L. LUTFI,
Raja NOOR AINON,
Language Engineering Lab,
Software Engineering Department,
Faculty of Computer Science and Information Technology (FCSIT)
University Malaya, Kuala Lumpur.
syaheerah@perdana.um.edu.my,
ainon@um.edu.my

Zuraida M. DON

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

Abstract

This paper concerns the infusion of naturalness into Malay Text-to-Speech (TTS) systems through the addition of an affective component. The goal is to introduce a novel method for generating expressive speech by embedding an 'emotion layer' called **eX**pressive **T**ext **R**eader **A**utomation Layer, abbreviated as eXTRA. The emotion generation method is template-driven. The templates are diphone-based and each template carries unique affective data. These emotion templates apply affectionate intonations to synthesized speech in two types of emotions in this prototype: anger and sadness. The module is an independent component that can serve as an extension to any TTS system that uses Multiband Resynthesis Overlap Add (MBROLA) engine. In a pilot project, the prototype is used with Fasih, the first Malay Text-to-Speech system developed by MIMOS Berhad, which can read unrestricted Malay text. eXTRA is evaluated through perception tests by which the results show more than sixty percent of recognition rate, which confirmed the satisfactory performance of the approaches.

1 Introduction

When communicating emotions, enhanced meaning and patterning are perceived and an atmosphere harmonious to natural learning is created. In particular, vocal emotions supply significant cues to delivering accurate, effective messages. However, the ability to express emotions distinguishes human speech from synthetic speech. The robotic and rather unnatural output quality of current Text-to-Speech (TTS) systems can be a reason to restrict the application of this technology. Therefore, there is a calling factor to improve output of TTS to optimize its use.

In recent years, there have been an emerging number of studies focusing on Malay text-to-speech conversion (El-Imam and Don, 2000; Razak, Abidin and Komiya; 2003 Tiun and Kong, 2005; Syaheerah *et al* 2005, 2005a). These are *concatenative* speech conversion systems, which mostly apply phonological rule-based approach for prosody modification in order to invoke imitation of humans' pronunciation. Nonetheless, though these prosodic models were introduced in the hope of providing a high degree of naturalness, it is still insufficient to make the output sound less mechanized.

Three major issues that contribute to this problem have been identified; firstly, there are various linguistic features that interactively affect the phonological characteristics, making it difficult to gather complete rules to describe the prosody diversity (Wu and Chen, 2002). The second challenge in modelling an affective component is the variability in speech. A speaker may not repeat what he says in the same way; he may not use the same words to say the same thing twice knowingly or not (even in read speech) (Murray and Arnott, 1996). One can also say the same word in many different ways depending on the context. Therefore, the instances of the same word will not be acoustically identical. This is quite difficult to map in a TTS system, especially when using qualitative rules, which causes the repetition of the same set of prosody when reading long sentences.

The usual practise is that, the linguistic features and speaking styles will be translated into prosodic patterns, which are repeatedly applied to speech. While this may be good for a small amount of sentences, repeated tones become boring and tedious for reading whole paragraphs of text. Apart from that, the same sets of tones do not fit different types of sentences with varying contents and lengths. (Syaheerah *et al*, 2005a). Therefore, applying fixed qualitative rules to prosodic variation patterns or ranges comes with great limitations. Lastly, there is a dearth of prerequisite studies on various human emotions. Consequently, to find a solution for these issues, a novel approach using emotion templates to apply expressiveness to the output of TTS system

was investigated. This paper presents the completed work and the prototype.

2 Background Studies

2.1 Emotions in Speech

Emotions play a significant role as an affective influence, especially in capturing attention (Brave and Nass, 2003). In fact, the significance is so great that it leads brain-based researchers such as Sprenger (2002) to assert that emotions mediates just about anything, as she puts it, "emotions drive attention, and attention drives learning, memory and just about everything else". Thus, anything that does not have emotions is often perceived as boring or unconvincing. When confronted with an interface, users actually constantly (non-consciously) monitor cues to the affective state of their interaction partner. (Reeves and Nass, 1996 as cited in Brave and Nass, 2003). This obviously shows that a natural and efficient interface requires not only recognizing emotion in users, but also to *express* emotion.

Emotions can be conveyed verbally or non-verbally in communications. Vocal emotions contain strong prosodic cues. 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 (Picard, 1997). In fact, many studies (Murray and Arnott, 1993,1996; Mullenix, 2002;Mozziconacci, 2002) demonstrated that it is possible to identify the various aspects of speaker's physical and emotional state, including age, sex, appearance, intelligence and personality by voice alone.

Studies on vocal expressions of emotions are developed from two approaches, perception-oriented and acoustic-oriented (Razak, Abidin and Komiya 2003; Cahn 1990; Scherer (1978) as quoted by Hofer 2004). The perception-oriented approach is listener centered and is concerned with how the listeners *perceive* the emotions, whereas the acoustic-oriented approach is speaker centered and is concerned with the analysis of vocal parameters of *expressed* speech that links to emotion. Acoustic features in speech are further divided into *prosodic* and *phonetic* classes. Phonetic features deal with basic sounds such as sounds produced by vowels or consonants in speech and their pronunciations. Prosodic features are composed of pitch, temporal (duration) and amplitude structure that correlate to the intonation or rhythmic aspects of speech such as stress word in utterance, the raising and falling of pitch and accents (Razak, Abidin and Komiya, 2003; Murray and Arnott, 1993). Prosodic features of speech contribute more to emotional expressions in speech than phonetic features.

Prosody is a combination of such parameters as pitch, duration and loudness (Zuraida, 1996). In synthetic speech, a prosodic pattern is provided by inserting variation of the combination mentioned. More importantly, the variation of *pitch* (or fundamental frequency, F_0) and *duration* is one of the main elements of a prosodic pattern that could avoid synthetic speech from sounding monotonous and hence, become more natural (Chen et. al, 2002; Cahn, 1990). Therefore, an emotion layer should have the information of both these core conveyers of affect. This can be done by extracting the acoustic correlates of anger and sadness, in terms of pitch and duration from real human voice samples.

2.2 Culturally-specific Vocal Affect

The outcome of studies that involve emotions that are socially constructed (Silzer, 2001; Wazir-Jahan, 1990) reveals that it is crucial to infuse a more familiarized set of emotions to a TTS system whereby the users are natives. This is because; a TTS system that produces affective output that is better 'recognized' would have a reduced artificiality and increased spontaneousness, hence offering users more comfort when interacting with the TTS system. The finding from an experiment by Nass and Lee (2002) also supports this statement. This leads us into building the emotion layer in such a way that it generates emotions that are more 'agreeable' to the Malay people. This is done by directing the speaker to record her speech by speaking them in the two emotional states in relation with the Malay culture. In order to motivate and focus the speaker, each of the sentences was accompanied by a scenario that elicits the related emotion. For example "Kamu sungguh kurang ajar" (You are so rude) and "Reaksi terhadap anak murid yang menendang kerusi guru dengan sengaja" (Reaction towards a student of yours who kicked your chair on purpose) were sentence and scenario, respectively. Having such elicitation scenario helps to reduce the interpretation variations.

To ensure that the intended emotions elicited in the speech samples are recognized by listeners, a series of perceptual tests was conducted. Results show that the speech samples are highly recognized with minimum effort, in other words, these samples are perceived as intended by the native participants.

The chart in Figure 1 below shows the summary of the result:

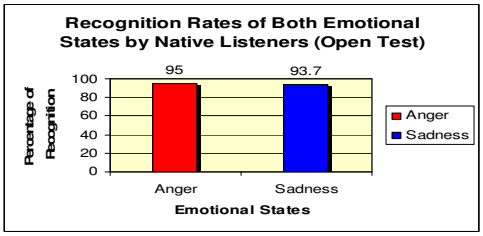


Figure 1: Test result from actor's speech samples

The detailed results can be found in Syaheerah *et al* (2005b).

2.3 The Malay Language Syllable Structure

It is observed that in Malay language, the structure of syllables is straightforward. In addition, the intonational or prosodic relationship between syllables within a word is more obvious than between two words. 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 (Syaheerah *et al*, 2005c). In Malay, the syllable structure units are as follows:

- **CVC** (Consonant-Vowel-Consonant)
- **CV** (Consonant-Vowel)
- **VC** (Vowel-Consonant)

2.4 Proposed Method

A prototype by Wu and Chen (1999) on template-driven generation of prosodic information for their concatenative Chinese TTS system has inspired the use of templates to generate emotions for Malay synthesized speech. Additionally, the findings in the interdisciplinary studies discussed in previous sections shaped the idea to propose a hybrid technique for building an effective emotion component to be used with concatenative Malay TTS system. The rationales are listed below:

i. Since the hosting system uses diphone concatenative synthesis (MBROLA), the employment of this technique is compulsory.

ii. The facts about Malay language syllable structure discussed section 2.3, added with the restrictions of phonological rule-based systems mentioned in section 2.1, shaped the idea to create a *syllable-sensitive* rule-based system.

iii. The effectiveness of the template-driven method proposed by Wu and Chen (1999) has brought the idea to adapt this method and combine it with the techniques in (i) and (ii).

3 Template-driven Emotion Generation

The combination of the techniques in (i), (ii) and (iii) above derives the **eX**pressive **T**ext **R**eader **A**utomation Layer, or eXTRA. Figure 2 below shows the block diagram depicting eXTRA being the affective component added to the host TTS.

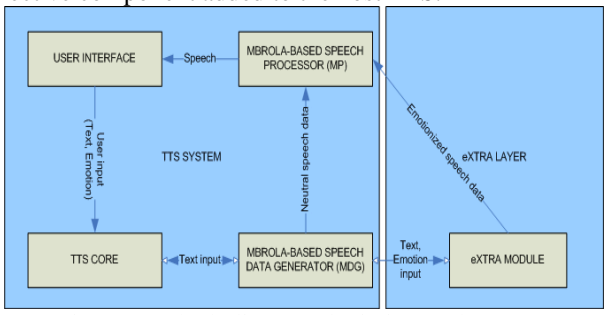


Figure 2: Block diagram depicting on how eXTRA is hooked into the host TTS

3.1 Emotion Templates

For the generation of highly natural synthetic speech, the control of prosodic parameters is of primary importance. Diphone synthesis allows maximum control of prosodic parameters. Therefore, attempts to model the emotions in eXTRA took advantage of model-base mapping or "copy synthesis" to build the emotion templates. In other words, the emotional prosodic parameters from the actor's speech samples are 'copied' into the templates. First, the actor's speech data is annotated on phoneme level using Praat (Boersma and Weenink, 2005). Then, the exact pitch and duration information from each phoneme is extracted and transferred into acoustical data in templates, which ensures more natural emotional-blended speech when the target template is applied to the speech. Figure 3 below summarizes the process of emotional prosodic information extraction process.

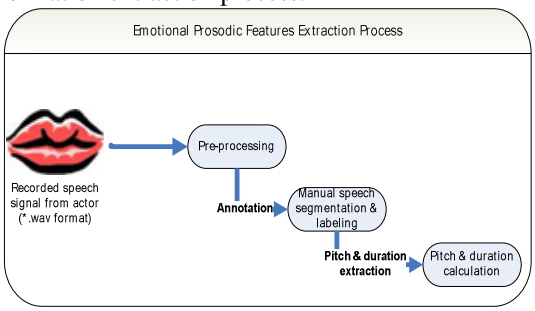


Figure 3

The extracted emotional prosodic information is directly stored in MBROLA player according to MBROLA input format (figure 4) in *.pho format. Each file describes different emotional tones and functions as an emotion template. There are sixteen emotion templates describing sixteen emotional tones for each emotional state. Figure 4 shows a few samples of the emotion templates. Two samples depict anger and the other two depicts sadness.

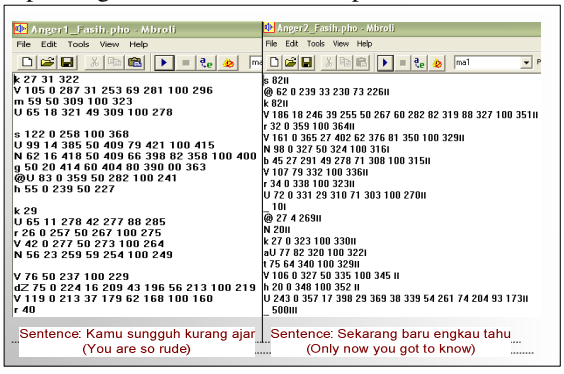
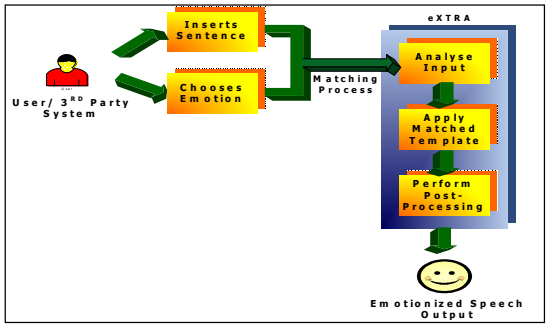


Figure 4: A sample of anger emotion templates

The phonemes (listed vertically) in the templates' sentences are converted to Speech Assessment Methods Phonetic Alphabet (SAMPA) symbols. Each phoneme consists of its own emotional prosodic parameter values (listed horizontally).

4 How eXTRA Works

Figure 5 provides the visual illustrations of



eXTRA's framework.

Figure 5: A simplified framework of eXTRA

Using the syllable-sensitive algorithm, each word from user input is analyzed and chunked into syllables in reverse order (stack) to determine syllable count; the input sentence is processed from the last word to the first. The result is then matched against the emotion template that contains the sentence with the same syllable-count and sequence. In other words, the template selection is done by identifying the integers that represent the syllable sequence of the template-sentence – "2222", "2332" etc. This is done by using a template selector module. When matched, the prosodic information from the template will be transferred to input at the level of phoneme. To ensure a more natural tune, the post-processing is done. It involves assigning silence and default parameters to additional phonemes correlating to each word wherever necessary. Figure 6 below presents a screenshot of the Fasih extended with eXTRA.

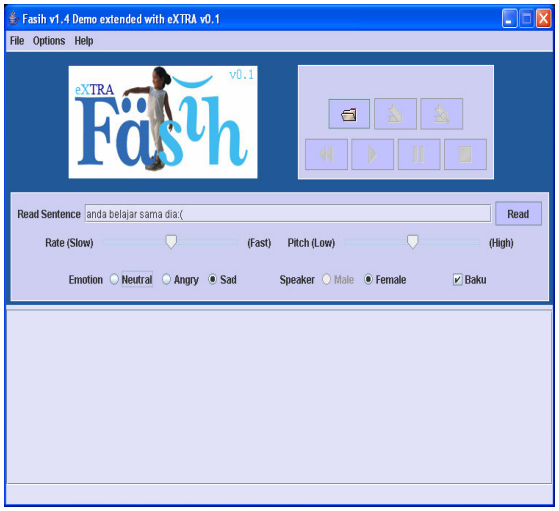


Figure 6: A screenshot of Fasih extended with eXTRA

5 Evaluation

The prototype is evaluated in a perceptual test participated by 10 native listeners who were not aware of the test stimuli. They were asked to listen to a series of neutral and emotionally inherent sentences. The results are presented in the chart of Figure 7.

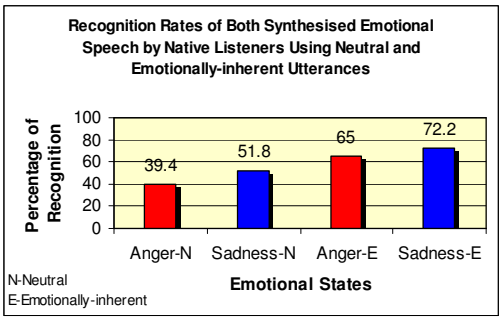


Figure 7: Results for synthesized emotional speech using netural and emotionally inherent content.

5.1 Summary of Results and Conclusion

Recognition rates for synthesized utterances that use neutral content were quite low, while for emotionally-inherent content, the recognition accuracy was significantly higher. It is observed that participants tend to focus on the contents rather than the tones elicited despite repeated reminders. Nevertheless, this kind of response is expected, because in real life situations, meaning and context are a bigger clue to the emotional state of the speaker. There were also significant differences in recognition accuracy among the two emotions using emotionally inherent contents: recognition rates observed for Anger set were significantly lower than Sadness sets; either the speaker was relatively less successful in expressing anger in some utterances, or anger is more difficult to recognize in certain utterances. Apart from that, the significant differences shown in the results from the experiments between neutral and emotionally-inherent contents proved that utterances that have no conflicting content is more suitable for use in building templates.

Overall, the recognition rates show higher figures compared to previous research work (Bulut, Narayanan and Syrdal, 2002; Nass *et al.*, 2000; Murray and Arnott, 1993, 1996). Basically, these results indicated over sixty percent recognition rates for both intended emotions expressed in the synthesized utterances, which are encouraging, considering that people recognize only sixty percent emotion in *human* voice (Shrerer, 1981 in Nass *et al.*, 2000).

6 Acknowledgements

We are deeply grateful to Dr. Normaziah Nordin and Mr. Kow Weng Onn from the Pervasive Computing Lab, MIMOS for their fruitful suggestions and advise on improving this prototype. We are also greatly indebted to Mr. Imran Amin H. de Roode from Fullcontact, for providing professional guidance and assistance in technical effort. Finally, we would like to thank all whose

direct and indirect support helped us deliver this prototype in time.

7 References

Boersma, P., & Weenink, D. (2005). Praat (Version 4.3.17). Amsterdam, NL.

Brave, S., & Nass, C. (2003). Emotion in Human-Computer Interaction. In J. A. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 81-93). Mahwah, NJ: Laurence Elbaum Associates (LEA).

Bulut, M., Narayanan, S., & Syrdal, A. K. (2002). *Expressive Speech Synthesis Using a Concatenative Synthesizer.* **In** Proc. of ICSLP, Denver, CO.

Cahn, J. E. (1990). The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society, 8*, 1-19.

Chen, Y.*, et al.* (2000). Learning Prosodic Patterns for Mandarin Speech Synthesis. *Journal of Intelligent Information Systems,* 19(2), pp.95-109. Retrieved 18th Jan 2005, from Kluwer Academic Publishers Online, NL

El-Imam, Y. A., & Don, Z. M. (2000). Text-to-Speech Conversion of Standard Malay. *International Journal of Speech Technology, 3*(2), 129-146.

Hofer, G. O. (2004). *Emotional Speech Synthesis.* Unpublished Masters Degree Thesis, University of Edinburgh.

Mozziconacci, S. (2002). Prosody and Emotions, *Speech Prosody 2002*. Aix-en-Provence, France: ISCA.

Mullenix, J. W.*, et al.* (2002). Effects of Variation in Emotional Tone of Voice on Speech Perception. *Language and Speech, 45*(3), 255-283.

Murray, I. R., & Arnott, J. L. (1993). Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal Acoustical Society of America, 93*(2), 1097-1108.

Murray, I. R., & Arnott, J. L. (1996, October 3-6th). *Synthesizing Emotions in Speech: Is It Time to Get Excited?* **In** Proc. of 4th International Conference on Spoken Language Processing 1996, (pp.1816-1819).

Nass, C. & Lee, K. (2002). Does Computer-synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-attraction, and Consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3), 171-181.

Nass, C.*, et al.* (2000). *The Effects of Emotion of Voice in Synthesized and Recorded Speech*.Unpublished manuscript, Standford, CA.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: The MIT Press.

Razak, A. A., Abidin, M. I. Z., & Komiya, R. (2003, 21-24 Sept). *Emotion Pitch Variation Analysis in Malay and English Voice Samples.* **In** Proc. of The 9th Asia-Pacific Conference on Communications 2003, (pp.108-112).

Silzer, P. J. (2001, Oct 31st-Nov 3rd). *Miffed, Upset, Angry or Furious? Translating Emotion Words.* **In** Proc. of ATA 42nd Annual Conference, Lost Angeles, CA, (pp.1-6).

Sprenger ,M. (2002). Becoming a "Wiz" at Brain-based Teaching: How to Make Every Year Your Best Year (2nd ed.). California: Corwin Press.

Syaheerah, L. L.*, et al.* (2005, 19-21st September). *Adding Emotions to Malay Synthesized Speech Using Diphone-Based Templates.* **In** Proc. of 7th International Conference on Information and Web-based Applications & Services (iiWAS 05), Kuala Lumpur, Malaysia, (pp.269-276).

Syaheerah, L. L.*, et al.* (2005a, 12-15th December). *Template-Driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment.* **In** Proc. of 4th International Conference of Information Technology in Asia (CITA 05), Kuching, Sarawak, (pp.144-149).

Tiun, S., & Kong, T. E. (2005). Building a Speech Corpus for Malay TTS System, *National Computer Science Postgraduate Colloquium 2005 (NaCPS'05).*

Wazir-Jahan, K. (Ed.). (1990). *Emotions of Culture: A Malay Perspective*. NY: Oxford University Press.

Wu, C. H., & Chen, J. H. (1999). *Template-Driven Generation of Prosodic Information for Chinese Concatenate Synthesis*. **In** Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Phoenix, Arizona, (pp.65-68).

Zuraidah, M. D. (1996). *Prosody in Malay: An Analysis of Broadcast Interviews*. Unpublished PhD Thesis, University Malaya, Kuala Lumpur