eXTRA: A Culturally Enriched Malay Text to Speech System

Syaheerah L. Lutfi[1], **Juan M. Montero**[1], **Raja N. Ainon**[2] **and Zuraida M.Don**[3]

Abstract. This paper concerns the incorporation of naturalness into Malay Text-to-Speech (TTS) systems through the addition of a culturally-localized affective component. Previous studies on emotion theories were examined to draw up assumptions about emotions. These studies also include the findings from observations by anthropologists and researchers on cultural-specific emotions, particularly, the Malay culture. These findings were used to elicit the requirements for modeling affect in the TTS that conforms to the people of the Malay culture in Malaysia. The goal is to introduce a novel method for generating Malay expressive speech by embedding a localized 'emotion layer' called eXpressive Text Reader Automation Layer, abbreviated as eXTRA. 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 in four emotions: anger, sadness, happiness and fear. In this paper however, concentration is given to the first two emotions. eXTRA is evaluated through open perception tests by both native and non-native listeners. The results show more than sixty percent of recognition rate, which confirmed the satisfactory performance of the approaches.

1 INTRODUCTION

Intelligent systems have been shown to increase their effectiveness by adapting to their individual users. In recent years, it has been recognized that affective factors can play an important role in this adaptation. One of the core conveyers of affect is speech. However, most of the research within the field of intelligent systems uses synthetic characters that are usually based on a full-blown simulated individual personalities whose behaviours are psychologically [1]}or biologically-driven (e.g., Blumberg, 1994 in [2]. This includes having conversational styles that are limited to a certain content and manner of speech, which reduces the believability and trustworthiness of such characters. It is believed that the reason for this is because such character's design ignores the pivotal mask of affect, which is the *socio-cultural grounding* [3];[4];[2];[5];[6]. Little research within the emotion-oriented technology field aims at understanding cultural differences which influence vocal affect.

While some aspects of emotion are universal to all cultures, other aspects may differ across cultures [4];[2]. For example, Americans and Asians have slightly different conceptions of self. American culture promotes a view of the self as independent. On the other hand, most Asian cultures, such as those of Japan and China, promote a view of the self as interdependent (collectivist culture). People from these cultures tend to describe themselves in terms of which groups they belong to. They learn to rely on others, to be modest about achievements, and to fit into groups. Maldonado and Hayes Roth [2] pointed out that biological or psychological model of synthetic characters do not express "the essence of humanity of their constructions", for example, particularly referring to speech, the intonation in the utterances in a particular topic addressed may not conform to a certain culture in terms of *what*, *how* and *when* it is said. This is more obvious when the topic being addressed is within the realm of persuasion. Taking into considerations the findings of these studies, we therefore proposed that the modeling of affective speech for a TTS to be expanded to include pursuits of cultural variability, producing a culturally-specific synthetic speech

2 BACKGROUND STUDIES

2.1 Culturally-dependent Vocal Affect

The expression and perception of emotions may vary from one culture to another [7]. Recent studies [8];[9] reveal that localized synthetic speech of software agents, for example, from the same ethnic background as the interactor are perceived to be more socially attractive and trustworthy than those from different backgrounds. The participants in Nass's [9] experiments conformed more to the decisions of the ethnically matched characters and perceived the characters' arguments to be better than those of the ethnically divergent agents. Just as with real people, users prefer expressive characters that "sound" like them; that they can better relate with, because it's easier to understand and predict. Therefore, cultural localization is critical even in the effort of actively matching the user's ethnicity, and perhaps also central psychological tendency.

Particularly in speech, the acoustic characteristics such as intonation, pronunciation, timbre, and range of vocal expressions that are localized to certain extent of variability, are constantly used in everyday activities to differentiate individuals across cultures [2]. These conversational aids can be used to determine not only the geographical origin of a particular person or character, but even their cultural influences and places of residences.

Based on these studies, we realized 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

Additionally, by concentrating on the culturally-specific manner of speaking and choices of words when in a certain

[1] Language Technology Group, Technical University of Madrid, email: {syaheerah,juancho}@die.upm.es
[2] Language Engineering Lab, University Malaya, email: ainon@um.edu.my
[3] Faculty of Language and Linguistics, University Malaya, email: zuraida@um.edu.my

emotional state, the risk of evoking confusions or negative emotions such as annoyance, offense or aggravation from the user is minimized, other than establishing a localized TTS.

2.2 Vocal Affect in Malay Culture In Relation To Anger and Sadness

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 [10]. 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 a*mok*), 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 examined 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*" [11]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 [7] illustrates that the expression of human emotions are cultural specific, e.g. how anger is expressed in English is different from how 'marah' (anger) 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.

Preliminary studies show 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 [12]

2.3 Issues in Affective Speech Modelling

In recent years, there have been an emerging number of studies focusing on Malay text-to-speech conversion [12-16]. 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 localize the output (to make it culture-dependent), hence, limiting its naturalness.

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 [17]. The second challenge in modeling 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 convey the same message twice knowingly or not (even in read speech) [18]. 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 [14]. 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 of the prototype.

3 THE MALAY SPEECH DATABASE

The findings from the studies above lead us into building a database that consists of speech data with 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 *suitable* with the Malay identity. There are two sets of utterances: one with neutral contents, and the other with emotionally-inherent contents. Each set contains thirty two utterances. For each of the utterances with emotionally-inherent contents, an accompanying scenario that elicits the corresponding emotion is given. 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 open perceptual tests was conducted. Results show that the speech samples with neutral content set have a low recognition rate while the samples with emotionally-inherent content are highly recognized with minimum effort, in other words, these samples are perceived as intended by the native participants. The results are shown in section 6.

4 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 [15]. In Malay, the syllable structure units are as follows:

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

5 IMPLEMENTATION

A prototype by Wu and Chen [17] 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 justifications are listed below:

i. Since the hosting system uses diphone concatenative synthesis (Multi-Band Resynthesis OverLap Add or MBROLA), the employment of this technique is compulsory.
ii. The facts about Malay language syllable structure discussed section 6, added with the restrictions of phonological rule-based systems mentioned in section 4, shaped the idea to create a *syllable-sensitive* rule-based system.
iii. The effectiveness of the template-driven method proposed by Wu and Chen [17] has brought the idea to adapt this method and combine it with the techniques in (i) and (ii).

Table 1: Detailed Information on the Child Components

Child Components	Responsibility
of Template Selector: SyllableReader TemplateMatcher	This component reads and analyses syllables This component matches the results of the analysis, with data from the database in order to select the correct emotion template.
of Merger: Composer Emotionizer	This component provides for navigable structures of phonemes (input and template are separately handled). This component applies a rule-based algorithm to Composer data in order to merge input-derived data with template data.

5.1 Template-driven Emotion Generation

The combination of the techniques in (i), (ii) and (iii) above derives the eXpressive Text Reader Automation Layer, or eXTRA. Figure 1 exposes eXTRA module's detailed internal architecture, while Table 1 explains the responsibilities of each of the child component.

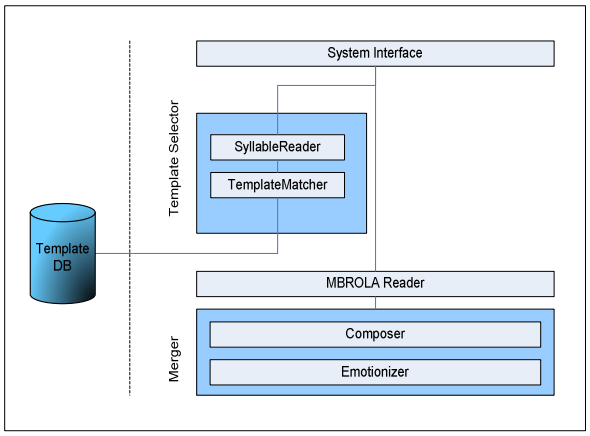


Figure 1: Low-level Architecture of eXTRA

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 speech analysis software, Praat [19]. 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. The next section explains how the prototype works in more detail.

5.1 How eXTRA Works

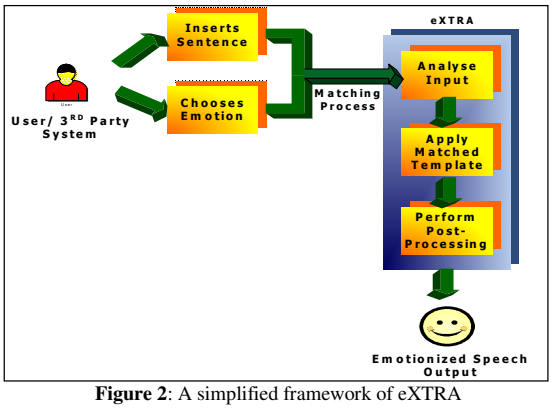


Figure 2: A simplified framework of eXTRA

Figure 2 provides the visual illustrations of eXTRA's framework 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 2 shows the framework of eXTRA while Figure 3 below presents a screenshot of the Fasih extended with eXTRA.

Consider the input sentence "Awak tidak tahu malu" (you have no shame) is to be spoken in anger. This sentence has a syllable sequence set of "2222". Therefore, the anger template that will be selected from the database also comprises the syllable sequence set "2222". The sentence in this template is "Kamu sungguh kurang ajar" (You are so rude). Consequently, the anger template is applied to the input sentence to produce an emotionized output. This is done by matching the emotional prosodic parameters from the template-sentence to the input-sentence at the level of phonemes. The matching process is explained in the next section. 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.

In other words, the eXTRA module then enhances this speech data to become emotional speech data by applying an emotion template. Thus, the generating of emotional speech output requires three essential components: *input data* and *template data* (both representing speech data) and a *rule-based algorithm* that renders the data into output (Figure 3).

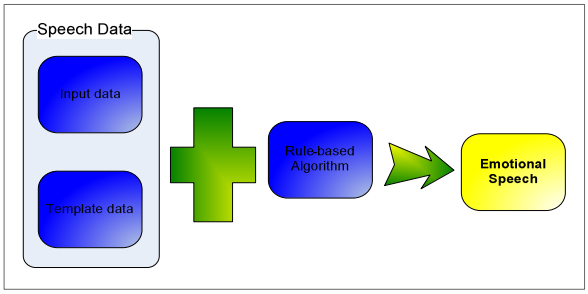


Figure 3: Major Components That Produce Emotional Speech

5.1.1 Matching Process

Consider the input sentence is "Awak tidak tahu malu" (you have no shame) and the selected emotional state is anger. This sentence has a syllable sequence set of 2222, therefore the matched anger template would be the template that has the same syllable sequence as well. From the template database, the particular matched template consists of the sentence "Kamu sungguh kurang ajar". Appendix A shows how the input phonemes are matched against the template phonemes. Vowels usually have longer duration than consonants, thus, contributing to more pitch points. However, vowel pitch points are not suitable to be transfered to consonants, since this may produce longer sound than expected. To solve this issue, syllabic and categorical matching are applied. Syllabic matching refers to the matching of phonemes between the input and template according to syllables. In other words, a pattern of syllables from the sentence is first identified in order to establish a match against another sentence's syllable pattern. Categorical matching refers to the matching of phonemes of the same type; vowels are matched against vowels while consonants are matched against consonants. This is illustrated in Table 2, where the vowels from the input sentence are matched against the vowels from the template sentence according to syllables. This also applies for consonants.

In the case where a phoneme is left without a match, a default duration value or silencing is assigned. A default duration value is assigned to the unmatched phonemes in the input sentence while the unmatched phonemes in the template are put to silence.

Table 2: The Organization of Matching Between the Template and the Input Contents

Line No.	Contents of Template Sentence			Contents of Input Sentence		
	SAMPA symbol	Parameters		SAMPA symbol	Parameters	
		Duration(ms)	Pitch Points pairs		Duration(ms)	Pitch Points pairs
1	k	silenced				
2	V	105	0 287 31 253 69 281 100 296	V	105	0 287 31 253 69 281 100 296
3	m	59 (<92)	50 309 100 323	w	92	50 309 100 323
4	U	65	18 321 49 309 100 278	V	65	18 321 49 309 100 278
				k	92 (default)	

Legend:
Unmatched phoneme

Table 2 shows that the relevant prosodic parameters from the phonemes in the template are transferred to the matched phonemes in the input. A post-processing is also done for the purpose of assigning silence and default values to the 'left-over', unmatched phonemes. The example shows that consonant /k/ in the template is put to silence while consonant /k/ in input is given a default value of 92 for the opposite reason. Such value is given so that the consonant produces a basic sound when concatenated. This value is copied from Fasih, which assigns *only* duration parameter to its consonants.

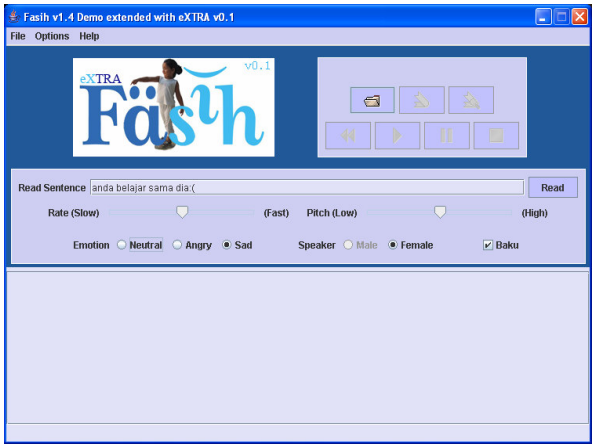
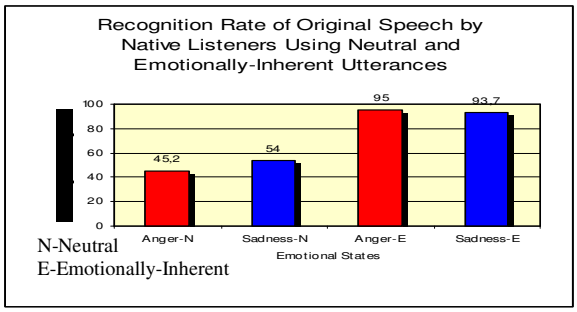


Figure 4: A screenshot of Fasih extended with eXTRA

6 EVALUATIONS

The prototype is evaluated in an open perceptual test participated by 20 native and 20 non-native listeners who were not aware of the test stimuli. They are Malaysian and international students of University Malaya, Kuala Lumpur. Native listeners were asked to listen to both sets of neutral and emotionally-inherent utterances (64 utterances) while non-native listeners only listened to emotionally-inherent utterances (32 sentences). A week earlier, they were asked to listen to the same set of utterances of the original samples (actress' speech). The results obtained with native listeners are presented in Figure 5 and Figure 6 respectively, while with non-native listeners, it is in Figure 7.



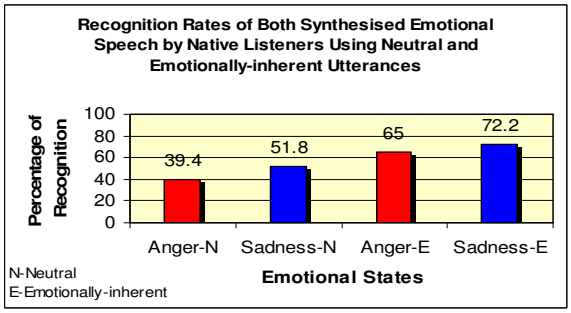


Figure 5 and 6: Recognition results for both original and synthesized speech samples using neutral and emotionally inherent content

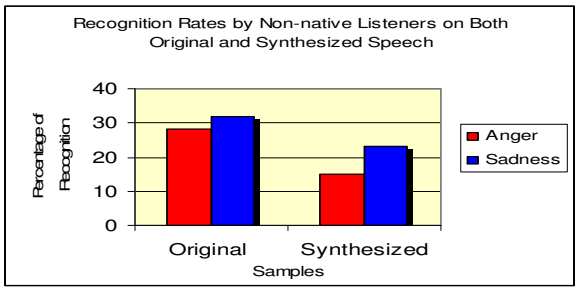


Figure 7: Results with non-native listeners

7 SUMMARIES OF RESULTS AND CONCLUSION

Comparing Figure 5 and Figure 6, the differences in recognition for samples of neutral contents between both charts are very small for both the emotions, which suggests that possibly the actress was relatively less successful in expressing the intended emotions using neutral contents. On the other hand, the difference of recognition for samples with emotionally-inherent contents shows that there is still room for improvement by approximately 25%, with regards to the modeling of the synthesized speech. The intonation and duration model for the emotional speech is used to generate the emotional utterances from neutral speech. In the original samples (Figure 5), the difference of recognition between the semantically-neutral and semantically-emotional speech bars is 45% on average while for the same comparison; the difference is 23.5% for synthesized speech. We suggest that these differences in recognition shown are due to content of the speech. 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. Lastly, the significant differences shown in the results from the experiments between neutral and emotionally-inherent contents proved that utterances that have no conflicting content and localized emotions are more suitable for use in building templates.

As for non-native listeners (Figure 7), there is a high difference of recognition between original and synthesized speech in anger compared to sadness. More participants are able to accurately detect sadness in synthesized speech compared to anger. This is possibly because of most sad speech samples exhibits lower F0 and longer duration and therefore it is easy to point out that the speaker is sad. Our data showed that in this open test, most participants from Middle East tend to perceive anger as "neutral", while participants from Western countries tend to presume sadness as "anger". This discovery is interesting to us as in some cases; even samples that produce clear angry expression (that can be easily detected by Malay listeners) are deemed as neutral. The low recognition rates clearly show that the non-native participants may have different perceptions from native participants. The findings based on data also proved Silzer's statement that "expression and perception of emotions vary from one culture to another"

Overall, the recognition rates by native listeners show higher figures compared to previous research work([20]; [9];[21];[22].. 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). This is possibly because due to the effort in localizing the emotion for better perception.

8 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. This work has also been partially funded by the Spanish Ministry of Education and Science under contract DPI2007-66846- C02-02 (ROBONAUTA).

9 REFERENCES

[1] Nass, C., Isbister, K., & Lee, E. , *Truth Is Beauty: Researching Embodied Conversational Agents*, in *Embodied Conversational Agents*, J. Cassell, et al., Editors. 2000, MIT Press: Cambridge, MA. p. 374-402.

[2] Maldonado, H. and B. Hayes-Roth, *Toward Cross-Cultural Believability in Character Design*, in *Agent Culture: Human-Agent Interaction in a Multicultural World* S. Payr and R. Trapple, Editors. 2004, Lawrence Erlbaum Associates: Mahwah, NJ. p. 143-175.

[3] Physorg.com (2007) *Research Finds that Culture is key to interpreting facial expressions*. DOI: Article 96297525

[4] Krenn, B., et al., *Life-Like Agents for the Internet: A Cross-Cultural Case Study*, in *Agent Culture: Human-Agent Interaction in a Multicultural World*, S. Payr and R. Trappl, Editors. 2004, Lawrence Erlbaum Associates: Mahwah,NJ.

[5] Hayes-Roth, B. and P. Doyle, *Animate Characters.* Autonomous Agents and multi-agent systems, 1998. **1**(2): p. 195-230.

[6] Reeves, B., & Nass, C, *The Media Equation: How People Treat Computers, Televisions, and New Media Like Real People and Places.* 1996, NY: Cambridge University Press.

[7] Silzer, P.J. *Miffed, upset, angry or furious? Translating emotion words*. in *ATA 42nd Annual Conference*. 2001. Lost Angeles, CA.

[8] Brave, S. and C. Nass, *Emotion in Human-Computer Interaction*, in *The Human-Computer Interaction Handbook: FundamentALS, Evolving Technologies and Emerging Applications*, J.A. Jacko and A. Sears, Editors. 2003, Laurence Elbaum Associates (LEA): Mahwah, NJ. p. 81-93.

[9] Nass, C., et al., *The effects of emotion of voice in synthesized and recorded speech*. 2000: Standford, CA.

[10] Wazir-Jahan, K., ed. *Emotions of Culture: A Malay Perspective*. ed. W.J. Karim. 1990, Oxford University Press: NY.

[11] Mohamad, M., *The Malay Dilemma*. 1981, Kuala Lumpur: Federal Publications.

[12] Razak, A.A., M.I.Z. Abidin, and R. Komiya. *Emotion pitch variation analysis in malay and english voice samples*. in *The 9th Asia-Pacific Conference on Communications 2003*. 2003.

[13] El-Imam, Y.A. and Z.M. Don, *Text-to-speech conversion of standard Malay.* International Journal of Speech Technology, 2000. **3**(2): p. 129-146.

[14] Syaheerah, L.L., et al. *Template-driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment*. in *4th International Conference of Information Technology in Asia (CITA 05)*. 2005a. Kuching, Sarawak.

[15] Syaheerah, L.L., et al. *Adding Emotions to Malay Synthesized Speech Using Diphone-based templates*. in *7th International Conference on Information and Web-based Applications & Services (iiWAS 05)*. 2005. Kuala Lumpur, Malaysia: University Malaya.

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

[17] Wu, C.-H. and J.-H. Chen. *Template-driven generation of prosodic information for chinese concatenate synthesis*. in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1999. Phoenix, Arizona.

[18] Murray, I.R. and J.L. Arnott. *Synthesizing emotions in speech: is it time to get excited?* in *4th International Conference on Spoken Language Processing 1996*. 1996.

[19] Boersma, P. and D. Weenink, *Praat*. 2005: Amsterdam, NL.

[20] Bulut, M., S. Narayanan, and A.K. Syrdal. *Expressive speech synthesis using a concatenative synthesizer*. in *ICSLP*. 2002. Denver, CO.

[21] Murray, I.R. and J.L. Arnott, *Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion.* Journal Acoustical Society of America, 1993. **93**(2): p. 1097-1108.

[22] Murray, I.R., J.L. Arnott, and E.A. Rohwer, *Emotional Stress in synthetic speech: progress and future directions.* Speech Communication, 1996. **20**(1-2): p. 85-91.

APPENDIX A: A Visualization of the Phonemes Matching Process