ISCA Archive
http://www.isca-speech.org/archive

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

EM-HTS: Real-Time HMM-Based Malay Emotional Speech Synthesis

Mumtaz B. Mustafa[1], *Raja N. Ainon*[1], *Roziati Zainuddin*[1]

[1] Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

mumshaka4@perdana.um.edu.my, ainon@um.edu.my, roziati@um.edu.my

Abstract

This research aims at developing a real-time HMM-based Malay emotional speech synthesis (EM-HTS) that has the ability to synthesize any form of text input in four different expression which are neutral, anger, sadness and happiness. The quality of the emotional speech synthesis was improved by using Neutral to Angry, Sad, and Happy (NASH) duration generator, which uses context-dependent duration generation method to improve the duration information to the label files of target emotions for training purpose. We conducted three forms of evaluations to determine the accuracy, intelligibility and naturalness of the speech generated by EM-HTS. All the three tests show that the adopted method (NASH) gives a better reproduction of prosody compared to conventional method using the same training speech data.

Index Terms: HMM-based emotional speech synthesis, context-dependent duration conversion

1. Introduction

Recent researches have indicated that corpus-based speech synthesizer such as unit-selection and HMM-based speech synthesis can generate high quality natural human speech including emotional speech using an appropriate size of training data [1], [2] and [3].This was made possible by referring to the F0 and spectral model derived during the training process using real human speech [4]. HMM-based speech synthesis has been developed for many languages including Japanese, English and Thai. At this moment, these systems were mostly meant for research purposes rather than real life applications.

This research aims at development of an HMM-based Malay emotional speech synthesis (EM-HTS) that enables users to input any text and evaluate the quality of the synthesized expression as either neutral, anger, sadness and happiness. The quality of the emotional speech being synthesized in this research is improved by applying appropriate duration information during training. For this purpose we have developed an emotional duration generator (NASH) that applies a Malay linguistic context-dependent decision tree. Duration changes are made to the training label files (also known as utterance file). Figure 1 shows the overall working mechanism of HMM-based Malay emotional speech synthesis.

In our previous work on Malay speech synthesis, we have developed a rule-based prosody conversion method for Malay diphone text–to-speech synthesis system [5]. Neutral synthesized speech was re-synthesized to emotional speech by applying prosodic factors of the target emotions. We only considered F0 and duration factors to re-synthesize emotional speech. The limitation of our previous work is that we were unable to take into consideration other speech factors such as intensity and spectral features because such factors were not available for us to manipulate for synthesizing emotional speech. On the other hand, HMM-based speech synthesis makes use of more speech factors such as spectral, F0, duration and loudness to synthesize speech. HMM-based speech synthesis can generate better emotional speech compared to diphone-concatenative speech synthesis.

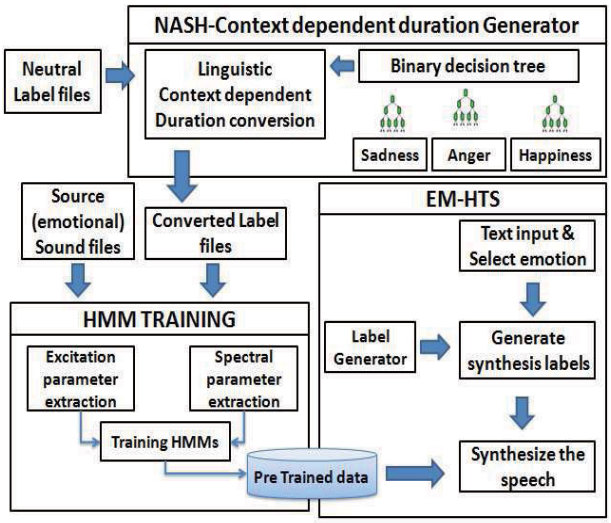

Figure 1: *The process of real-time synthesis of Malay emotional speech using EM-HTS.*

2. Data and Evaluation

The quality of HMM-based speech synthesis depends on the quality of the input data which comprises of recorded human speech and label files that contain phonetic and prosodic information of the training data particularly the duration length of each phone of the respective speech utterance. Malay is an under-resourced language with no readily available emotional speech database and a front end natural language processor to generate the label files.

2.1. Malay Label generation using Festival

Malay is spoken by more than 300 million people in several countries including Malaysia, Indonesia, Brunei, Singapore and Southern Thailand. This research focuses on Standard Malay (SM), which is used in formal contexts such as education and the mass media [6]. The writing system of SM uses the standard Roman alphabets with 21 consonant letters and 5 vowel letters. Although the alphabets are similar to English, the sound system of Malay differs from that of English [7].

The spelling system of Malay is rather straightforward whereby it follows the pronunciation of a word. The features of vowels can be classified based on the position of tongue during pronunciation which are front (/i/, /e/, /a/), centre (/ə/, /a/) and back (/u/, /o/) as well as the height of the tongue as high, mid-high and low. Malay language also has three diphthongs which are /ai/, /au/ and /oi/. Consonant has different kind of manner of articulation such as plosive or stop, affricatives, nasal and fricatives.

For the purposes of generating label files for Malay language, we have used Festival TTS system [8] that has been modified to generate Malay training label files. This was possible because Malay uses similar graphemes to English. There are two common approaches in deriving phonemic representation of words which are the rule-based and database. For Malay, rule-based method of grapheme-to-phoneme may seem to be the best way to generate phoneme representation. However, Malay language has many borrowed words that do not follow the common grapheme-to-phoneme rules set for Malay native words. For example, in SM, the vowel 'a' is usually pronounced as schwa in-stem final open syllable position with the exception to proper nouns and borrowed words. Similarly, the letter 'e' has several corresponding sound. In most instances it is pronounced as schwa, and in other instances as the mid- front vowel /ie/.

In view of many exceptions to general phoneme representation of Malay words, we decided to build a limited grapheme-to-phoneme database for Festival using 33 phonemes including pauses. The limited domain dictionary comprising of 2,763 words taken from 500 phonetically balanced sentences. We then use the database to generate the neutral sounding label files of Malay using Festival TTS system. Table 1 shows the phoneme classification of Malay according to IPA and its comparison to English.

Table 1. Lists of phonemes for Malay and English

	Vowel	Diphthong	Voiced-consonant	Unvoiced-consonant
Malay Phonemes	i, e, a, ə, u , o	au, ai, oi	b, d, g, j, l, m, n, ng, ny, r, w, y, v, z	p, t, k, c, f, h, kh, s, sy
Malay phonemes (used)	iy, ey, aa, er, uw, ow	aw, ay, oy	b, d, g, jh, l, m, n, ng, ny, r, w, y, v, zh	p, th, k, ch, f, hh, kh, s, sh
English Phonemes (available)	aa, ae, ah, ax, eh, er, ey, ih, iy, ow, uh, uw	aw, ay, oy, el, em, en, ey, axr, er	b, d, dh, g, jh, l, m, n, ng, r, w, y, v, z, zh	p, t, th, k, ch, f, hh, kh, s, sh

2.2. Building Malay Speech Corpus and evaluation

With the absence of readily available speech data for Malay, we developed 500 phonetically balanced Malay sentences which we have classified into short and long sentences (short referred to sentences with 6 or less words, while long refers to 7 or more words). We choose 500 sentences because most of the previous research for other languages uses this amount of sentences (503 sentences) for synthesizing good quality emotional speech [9, 10]. These 500 sentences were expressed by 2 speakers (1 professional male and 1 professional female) in 4 different expressions which are neutral, happiness, anger, and sadness. To evaluate the quality of the Malay Emotional Speaker-Dependent Speech (MESDS) corpus, 1,500 utterances were randomly selected comprising of equal short and long utterances (250 sentences x 2 speakers x 3 emotions) and were divided into fifteen sets of files comprising 100 snippets each.

A total of thirty evaluators were involved in the perceptual test that represents balanced gender, age and profession. Each evaluator listened to 100 voice snippets and determined the type of emotions being expressed from a choice of four which are anger, sadness, happiness or others.

2.3. Evaluation results

From the evaluation test, sadness has the highest rate of recognition among the three emotions with recognition rate of 90.18%. This is followed by anger at 86.17% and happiness at 84.44%. Sadness and anger were confused with each other during the evaluation indicating that there was some similarity between these two utterances particularly on the F0 contour. Anger was also confused with happiness mainly because of similarity to the speed of the expression.

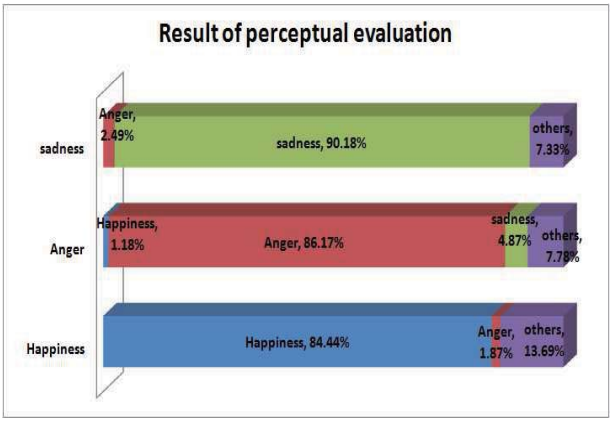

Figure 2: *Results of perceptual evaluation.*

Mixed results were obtained when we evaluated the emotion expression according to gender. Male expressed happiness and anger were recognized higher than female speakers. From the perceptual test, shorter Malay utterances has higher recognition rate than longer sentence indicating that shorter utterance portray better emotion than longer utterances. Table 2 shows the result of evaluation according to gender, and utterance length.

Table 2. *Recognition of emotion by categories.*

CATEGORIES		EMOTIONAL IDENTIFICATION RATE (%)		
		Happiness	Anger	Sadness
Gender	Male	81.22	85.92	84.93
	Female	80.08	81.42	92.47
Utterance	Long	77.23	80.78	85.00
	Short	84.07	86.56	92.40

3. HMMs training

EM-HTS uses pre-trained MESD speech data. Each type of expression was trained separately using HMM-based speech synthesis [11] for male and female voices using experimental conditions shown in table 2. We used 480 speech data from the 500 recorded utterances for training and the remaining 20 for evaluation. For training the neutral speech data, we used the standard label files generated by festival synthesis as discussed in part 2.

For training emotional speech, the label files generated by festival cannot be used because the phone duration information does not match the phone duration of the recorded emotion speech. In view of this, we have modified the neutral label files to contain emotional duration information using NASH duration generator. We adopted the linguistic contextual factor of Malay speech to determine the appropriate duration factor to be applied. We have developed a separate

decision tree for each emotion. In building the decision tree, we have applied the following contextual factors:

- number of words in a sentence
- word position in a sentence (WPOS 1 – 7)
- position of syllable in word (SY1: starting syllable, SY2: ending syllable, SY3: other syllable)
- type of phoneme
- position of phoneme in syllable

Table 2. *Experimental conditions for preparing MESD pre-trained data.*

Experimental condition	Neutral speech	Emotional speech with unmodified labels	Emotional speech with modified labels
Number of training data	480	480	480
Label files	unmodified	unmodified	NASH modified
Sampling rate	16 kHz		
Windowed	25ms Blackman with a 1ms shift		
HMMs	5-state left-to-right HMM		
Other features	25 mel-cepstral coefficient, zeroth coefficient, logarithm of F0, delta and delta-delta coefficients		

Emotional dependent phone duration was generated by applying conversion factors to the phone duration of neutral label files. We formulated the conversion factors by comparing the phone durational length of neutral recorded speech with the emotional speech. For this research, we have applied phoneme-level linguistic features that are common to all emotional utterances. The types of features that we have used are consistent with the HMM-based speech synthesis linguistic features. Table 3 lists the linguistic context of Malay language that we have used for NASH.

Table 3. *List of Malay linguistic contextual factors.*

Context	*phoneme*
vowel	a , e , i, o , u, ə
nasal	m , n ,ny, ng
plosive	p, b, t, d, k, g
fricative	f, v, s, sy, h
diphthong	ai ,oi, au

We applied position of speech tagging mechanism similar to word position identification in [12] to identify the position of phoneme in the source neutral sentence for NASH to apply relevant duration factors. For any form of text input, the first three words are tagged as 1, 2 or 3, the last three words are tagged as 4, 5 or 6 and any other words are tagged as 7. Syllable in a word is tagged as sy1 for beginning syllable, sy2 for ending syllable and sy3 for other syllables.

Within each syllable, NASH identifies the phone classes and its position to apply the most appropriate duration conversion factors. During our analysis, we found that vowels have the most significant duration movement compared to consonants. The duration of vowels also changes with their position in the syllable. In Malay, the vowel can take position at the beginning (V), in the middle (CVC) or at the end (CV) of a syllable. Based on this, we have formulated different duration factor for each types of vowel at its different syllabic position. Figure 4 shows the decision tree applied in NASH to determine the duration conversion factor for vowels.

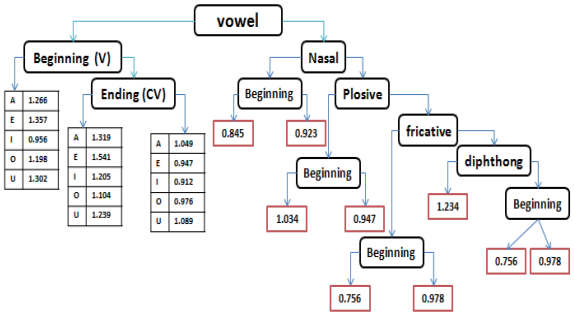

Figure 3: *Binary decision tree to determine duration conversion factors for vowels.*

To evaluate the effectiveness of NASH, we conducted two separate training for synthesizing emotional speech. The first training use 480 recorded speeches with unmodified label files (Denoted as HCS-480). The second used the same speech data using NASH modified label files (Denoted as HPCS-480).

4. System Evaluation

We conducted three forms of evaluations to determine the accuracy, intelligibility and naturalness of the speech generated by EM-HTS. To evaluate the accuracy of the synthesized speech, we compared the prosody of recorded speech and synthesized speech and established the root mean square error (RMSE). We then compare the RMSE of NASH enabled EM-HTS with conventional adaption synthesis. To investigate the naturalness of the synthesized speech, a perceptual evaluation similar to the first evaluation were conducted. Intelligibility was evaluated through an acceptance test whereby evaluators input any text using the interface and validate whether the text input was properly uttered by the system.

4.1. Accuracy testing and results

For evaluating the accuracy of NASH, we have conducted an objective evaluation test to investigate the impact of using NASH on the prosody of synthesized emotional speech. All twenty speakers dependent synthesized speech of each experimental condition for both male and female were compared with the recorded speech obtained earlier and a root means square error (RMSE) was calculated. Figure 4 shows the RMSE of synthesized female speech and figure 5 shows the RMSE for male synthesized speech.

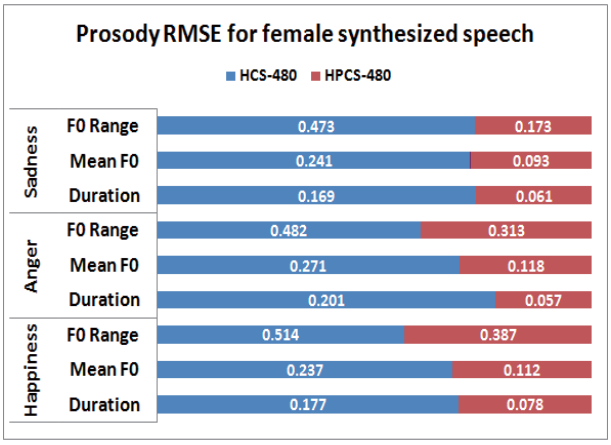

Figure 4: *Prosody RMS error for female synthesized speech.*

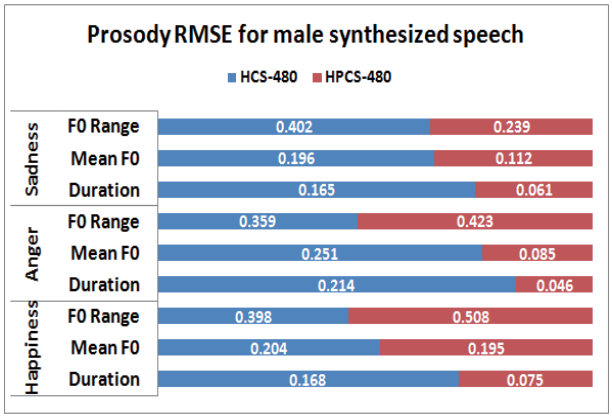

Figure 5: *Prosody RMS errors for male synthesized speech.*

The objective evaluation on synthesized speech shows among others, HMM-based conventional emotional speech synthesis (HCS-480) recorded high RMSE for F0 and duration with duration recorded a lower error comparative to F0. Although the sound generated was satisfactory, high error rate for F0 (mean F0 and F0 range) indicated poor F0 modeling by HMMs when synthesizing emotional speech. This is because HMM-based speech synthesis depends on the content of the label files to generate the appropriate F0 model. When we synthesize emotional speech using neutral label files using the 480 training data, the F0 model generated was less accurate since the duration reference used by HMMs during training was not the same with the duration of the actual speech data.

The NASH-modified 480 training data (HPCS-480) has lower RMSE compared to HCS-480 for both the F0 and duration. This is because the phone duration generated by NASH was close to the actual phone duration of the emotional speech. As a result, the synthesized phone duration of HPCS-480 reflects the phone duration of the natural emotional speech. When the label files contain accurate duration information that reflects the duration of emotional speech, HMM-based speech synthesis developed a better F0 models during training. A lower RMSE for F0 of NASH-modified synthesis was obtained compared to conventional synthesis. Figure 6 compares the RMSE for conventional emotional synthesis and NASH-modified synthesis.

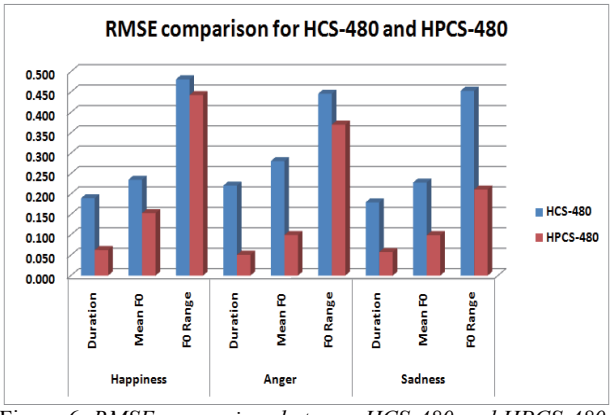

Figure 6: *RMSE comparison between HCS-480 and HPCS-480*

4.2. Naturalness testing and results

To evaluate the naturalness of the synthesized output, we have conducted a perceptual evaluation on the 20 synthesized speech in a manner similar to the first evaluation (refer to part 2.3). The same evaluators were involved in this test. Figure 7 shows the emotions recognition rate of the synthesized emotional speech trained using NASH modified label files.

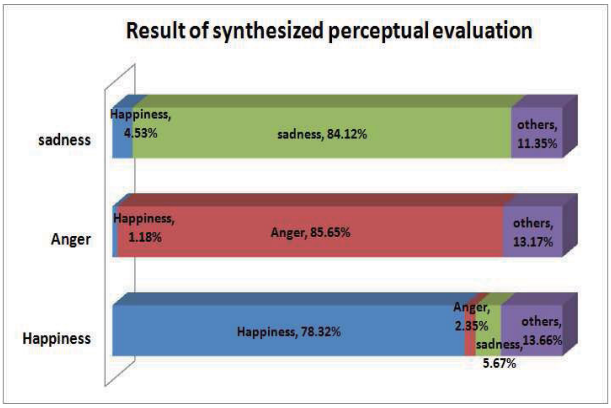

Figure 7*: Results of synthesized speech perceptual evaluation.*

Generally, the recognition rates of all three synthesized emotions were lower than the recognition rate of recorded speech. Between the three emotions anger has the highest recognition rate of 85.65% followed by sadness at 84.12% and happiness at 78.52%. From the evaluation, we can conclude that the deterioration of the recognition rate for synthesized speech is attributed to the weakness of HMM-based speech synthesis to accurately model the F0 particularly for emotional speech that has high F0 range within each phoneme like sadness and happiness. As a result the synthesized sadness emotion was no longer the highest recognized emotions. The low recognition rate of happiness and sadness was also due to the inability of the HMM-based speech synthesis to replicate other speech factors available in recorded speech such as giggle, laughter and breathiness.

Just like the earlier perceptual evaluation, male expressed happiness and anger achieved higher recognition than female speakers. Likewise, shorter Malay utterances has higher recognition rate than longer sentences. Table 4 shows the result of second perceptual evaluation classified to gender and utterance length.

Table 4: *Recognition of emotions by categories.*

CATEGORIES		EMOTIONAL IDENTIFICATION RATE (%)		
		Happiness	Anger	Sadness
Gender	Male	79.26	87.26	83.48
	Female	77.38	84.04	84.76
Utterance	Long	73.46	81.94	80.86
	Short	83.18	89.36	87.38

4.3. Intelligibility testing and results

We evaluate the intelligibility of EM-HTS by conducting an acceptance test. Thirty users were involved in this test, whereby they randomly input five text sentences of any length and then synthesized them in all the four expressions. Based on the synthesized output, the evaluators answered two questions which required the users to approve whether the system synthesize the text with proper articulation (Approval rate; 1 = yes, 0 = no). Based on the synthesized output, users rated the clarity of the synthesized output (1 = not clear, 2 = clear and 3 = very clear).

From the evaluation, the approval rate for the neutral synthesized speech was highest at 94.67% and lowest approval rate was for the happiness synthesized speech at 84.00%.

Anger approval rate is at 88.67% followed by sadness at 86.67%. In terms of clarity scale, neutral has the highest rate at 2.69 followed by anger at 2.41 and sadness at 2.37. Happiness has the lowest approval rate at 2.16. Table 5 shows the result for the intelligibility test.

Table 5: *Results of the intelligibility test.*

	Neutral	*Happiness*	*Anger*	*Sadness*
Approval rate (%)	94.67	84.00	88.67	86.67
Clarity scale	2.89	2.16	2.41	2.37

From this test, we found that the user's satisfaction to the synthesized neutral speech by HMM-based speech synthesis was generally high with approval rate of 95% and clarity scale of 2.89 (close to very clear). This shows that HMM-based speech synthesis can synthesize high quality speech for any form of text input. However, the approval rate and clarity scale for emotional speech synthesized by EM-HTS was lower compared to neutral speech. The average approval rate for all the emotional speech is at 86.44% and the average clarity scale is at 2.31 (close to clear). The lower approval rate for emotional speech when compared to neutral speech shows that synthesizing emotional speech requires more than just F0, duration, intensity and spectral but also other speech features such as giggle, laugher and breathiness.

5. Conclusions

This research allows end users to have hands-on experience using HMM-based speech synthesis and enables them to evaluate and comment on the performance of HMM-based speech synthesis based on their own choice of text input. The EM-HTS developed in this research has been tested for its accuracy of prosody generation, naturalness and intelligibility. The use of NASH duration generator has improved the accuracy of the prosody generation by EM-HTS with a much lower RMSE when compared to conventional method of emotional synthesis.

For naturalness test, we have evaluated 20 synthesized speeches in a listening test involving 30 users. Although the recognition rate of the synthesized emotional speech was lower than the recorded speech, the overall recognition of synthesized speech is about 80%. Finally, the EM-HTS has a high score of approval rating of 95% for neutral speech. This shows that the Malay emotional speech database (MESD) used in this research was phonetically-balanced and has a good coverage of the Malay Language.

6. Acknowledgement

This work is supported by a research grant from University of Malaya, Malaysia.

7. References

[1] Zen, H., Tokuda, K., Black, A.W., "Statistical parametric speech synthesis", Speech Communication, 51 (11), 1039-1064, 2009.

[2] Barra-Chicote, R., Yamagashi, J., King, S., Montero, J.M., and Macias-Guarasa, J., "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech", Speech Communication, vol. 2009.

[3] Black, A.W., "Unit-selection and emotional speech", In Proceeding of Eurospeech 2003, pp.1649-1652, 2003.

[4] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," In Proceeding of Eurospeech-99, pp. 2374-2350, 1999.

[5] Mumtaz, B., Ainon, R.N., Roziati, Z., and Zuraidah, M.D., "Integrating rule and template-based approaches for emotional Malay speech synthesis", in Proc. of Interspeech 2008, pp. 253-256, 2008.

[6] Al-Emam, Y.A., Zuraidah, M.D., "Rules and Algorithms for Phonetic Transcription of Standard Malay," IEICE Trans. Inf. And Syst., Vol.E88-D, No. 10, 2005.

[7] Teoh, B.S., The sound system of Malay revisited, Dewan Bahasa and Pustaka, Ministry of Education Malaysia, Kuala Lumpur, 1994.

[8] Festival homepage: http://www.cstr.ed.ac.uk/projects/festival/

[9] Yamagashi, J., Onishi, K., Masuko, T. and Kobayashi, T., "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis", IEICE Trans. Information and Systems, E88-D, 3, pp.502-509, 2005.

[10] Takashi, N , and Kobayashi, T., "A technique for estimating intensity of emotional expressions and speaking styles in speech based on multiple-regression HSMM", IEICE Trans. Inf. & Syst., vol. E93-D, No 1 January 2010.

[11] Tokuda, K., Zen, H., Yamagashi, J., masuko, T., Sako, S., Black, A., Nose, T., "The HMM-based speech synthesis system (HTS) Version 2.1. http://hts.sp.nitech.ac.jp/

[12] Zeynep, I., and Young, S., "Data-driven emotion conversion in spoken English", Speech Communication, 51, pp. 268-283, 2009.