Rule-Based Storytelling Text-to-Speech (TTS) Synthesis

Izzad Ramli[1], Noraini Seman[1], Norizah Ardi[2] and Nursuriati Jamil[1]

[1]*Digital Image, Audio and Speech Technology, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*
[2]*Academy of Languages Studies, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

Abstract. In recent years, various real life applications such as talking books, gadgets and humanoid robots have drawn the attention to pursue research in the area of expressive speech synthesis. Speech synthesis is widely used in various applications. However, there is a growing need for an expressive speech synthesis especially for communication and robotic. In this paper, global and local rule are developed to convert neutral to storytelling style speech for the Malay language. In order to generate rules, modification of prosodic parameters such as pitch, intensity, duration, tempo and pauses are considered. Modification of prosodic parameters is examined by performing prosodic analysis on a story collected from an experienced female and male storyteller. The global and local rule is applied in sentence level and synthesized using HNM. Subjective tests are conducted to evaluate the synthesized storytelling speech quality of both rules based on naturalness, intelligibility, and similarity to the original storytelling speech. The results showed that global rule give a better result than local rule

1 Introduction

Speech synthesis is the process of converting written text to spoken audio, also known as text-to-speech (TTS). Expressive speech is a form of speech synthesis aimed to improve the naturalness of neutral speech output of standard TTS [1]. Storytelling speaking style is a type of expressive speech that is important towards the development of digital storytelling [2] and a humanoid robot [3]. Earlier work of storytelling is done by incorporating basic emotions to instil naturalness [4]. For instance, Silva et al. [5] synthesized emotions such as happy, sad, surprise and fear to create a virtual storyteller resembling human storyteller. On the other hand, developed an emotion transformer to alter the neutral prosody into desired emotive prosody as well as transform the voice quality of the synthetic speech. The emotive TTS is further used in the humanoid audio-visual avatar.

Rule-based prosody modifications have been a popular approach to incorporate emotions in storytelling speaking style. This approach was undertaken by many expressive TTS in Dutch [6], English [7], Catalan [8], Spanish [9], Indian [10], German [11] and Korean [12] languages. In [13], global rules are used to modify prosodic parameters of pitch, intensity, overall speech tempo and pause duration to convert neutral speech into storytelling. In a more recent work, local rules are employed at phrase-level instead of sentence-level to synthesize storytelling speaking style in Bengali, Hindi, and Telugu. Roekhaut et al. [14] also proposed local rules for varying prosodic parameters to synthesize storytelling style. The rules are applied to modify the prosody based on syllable position (initial, middle, end) in an utterance. Six-syllable categories are identified based on their positions and whether they are prominent/non-prominent syllables. Then, style conversion is applied to each of the syllable categories in a sentence.

In this paper, local rules and global rules of prosody modifications are further investigated to determine their performance in synthesizing storytelling in the Malay language. This paper is structured as follows. In Section II, an in-depth analysis of storytelling speaking style during neutral speech is conducted. Section III presents the process of deriving the global and local rule-sets for converting neutral speech to storytelling style speech. Evaluation of the synthesized storytelling speech using subjective tests is shown in Section IV. The conclusion and summary of the results are described in Section V.

2 Speech data acquisition and analysis

In this section, the data acquisition and prosodic parameters of neutral and storytelling speech are explained and analyzed, respectively.

2.1. Speech data collection

Two types of speech that is neutral speech and storytelling speech data are recorded from one male and one female experienced storytellers. Neutral speech is recorded from both storytellers as they read the scripts with no intonation. On the other hand, storytelling

speaking style is recorded using storytelling intonations of the storytellers. In storytelling literature, storytelling speech may be conveyed in narrative, descriptive or dialogue mode [15]. Narrative storytelling is mainly used to inform the listener about the actions that are taking place and the characters affecting the story. Descriptive storytelling, on the other hand, describes a character or events to the listener so that they feel the character or the storyline depicted. Meanwhile, dialogue storytelling is when the storyteller typically modifies his/her voice into a more exaggerated register of expressions, where full-blown emotions may be manifested. In this work, storytelling is done in a narrative mode. Figure 1 shows the recording sessions of the female and male storytellers.

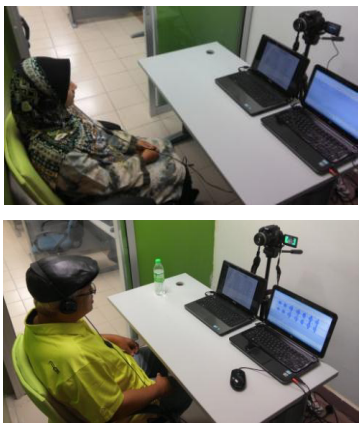

Figure 1. Recording sessions of the narrative storytelling.

In this preliminary investigation, one children's Malay folklore entitled "*Si angsa yang bertelur emas*" is selected to be narrated by the storytellers. The story is consists of 12 sentences, 113 words, and 276 syllables. The recording session is done in a quiet room, and the storyteller is seated in a comfortable chair approximately 2 feet away from the video camera. The recorded speech is stored in .wav format at a sampling rate of 41Khz with 16-bit resolution. The storytellers are given ample time to practice before recording session and they may repeat the recording until they are satisfied. All the neutral and storytelling speaking speech data are further transcribed at the sentence, word, and syllable levels using Praat [16] tool.

2.2. Prosodic parameter analysis

Analysis of prosodic parameters is important to examine which aspects of the prosody are pertinent to each other. This knowledge is used to transform neutral speech into storytelling speaking style. A total of 48 sentences, 452 words and 1,104 syllables are transcribed from the neutral and storytelling speech data of both storytellers. Five prosody parameters are further extracted from both speech data; they are pitch, intensity, duration, tempo and pauses. Table 1 tabulated the average pitch, intensity and duration of all syllables from the neutral and storytelling speech of both storytellers.

Table 1. Average prosodic parameters of both storytellers.

		Duration	**Intensity**	**Pitch**
Neutral speech	Female	0.24	61.23	171.74
	Male	0.24	76.94	144.21
Storytelling speech	Female	0.2	68.99	238.61
	Male	0.19	77.59	150.6

The analysis shows that average intensity and pitch for storytelling speech at 68.99 and 238.61 are larger than a neutral speech at 61.23 and 171.74, respectively for the female storyteller. The average intensity and pitch for the male storyteller also show a slight increase in storytelling speech. Duration, on the other hand, shows a lower average for storytelling speech at 0.2 compared to neutral speech at 0.24. Hence, this indicates that the speech tempo for storytelling speech is faster than neutral speech. The same is also true for the male storyteller.

The pause feature is analysed at phrase and sentence level. In our work, a phrase is defined as a collection of words and determined by the symbol comma (,) that exist in a sentence. Table 2 shows the average pause at phrase and sentence level. It shows that neutral speech has a longer average pause for phrase and sentence level compared to storytelling speech for both storytellers. At the sentence level, both storytellers tend to pause much longer than at phrase level before continuing to the next sentence.

Table 2. Average pause at phrase and sentence levels.

		Phrase Level	**Sentence level**
Neutral speech	Male	0.46	0.87
	Female	0.40	0.82
Storytelling speech	Male	0.42	0.60
	Female	0.25	0.67

Further analysis also showed that several words, typically adjectives and adverbs contain accented syllables that are stressed during pronunciations. Accented syllables tend to have increased pitch, duration, and intensity. Table 3 presents an example of an adjective word '*berat*' containing accented syllables /be/ and /rat/. It can be seen that tremendous increase of 320% for the duration, 12 % for intensity and 109% of pitch occurred during storytelling for accented syllable /be/. Similar phenomena is also shown for syllable /rat/. However, the rate of escalation is different depending on the position of the accented syllable in the word. Syllable /be/ that is located at the beginning of the word has a large increase for the duration. Meanwhile, syllable /rat/ that is positioned at the end of the word has the largest increase in pitch. Upon more inspections of other accented syllables, we discovered that the prosody parameters of storytelling speaking style varied according to the syllable positions: initial, middle or end.

Table 3. Prosody analysis for accented syllables.

Word	Syllable	Prosody	Neutral speech	Storytelling speech
'berat'	/be/	Duration	0.15	0.63 (320%)
		Intensity	77.37	86.70 (12%)
		Pitch	110.38	231.11 (109%)
	/rat/	Duration	0.32	0.47 (46%)
		Intensity	75.74	77.61 (2.5%)
		Pitch	116.18	219.96 (89%)

3 Development of prosodic rules for storytelling speaking style

This section discusses the process of converting neutral speech to storytelling style speech as shown in Figure 2.

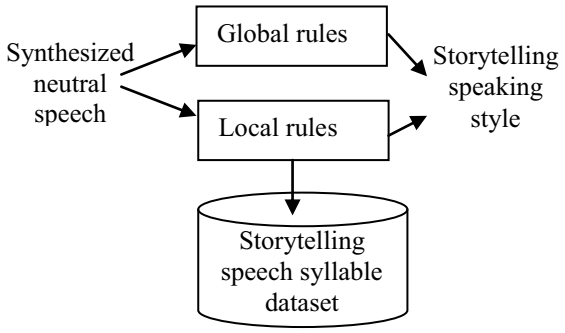

Figure 2. Rule-based storytelling speech synthesis.

The synthesized neutral speech is generated using a Malay language TTS [17]. As the aim of this paper is to compare the performances of local and global rules for synthesizing storytelling speaking styles, two set of rules are constructed to produce the storytelling speaking style speech. Global rules are set to modify prosodic parameters of accented syllables only, leaving other syllables unchanged. Global rules also alter overall speech tempo and pause duration. On the other hand, local rules modify both accented and unaccented syllables of the neutral speech to produce the storytelling speech. As mentioned earlier, prosody parameters of the accented syllables are modified based on the positions of the syllables in the word (i.e. initial accent, middle accent, final accent) [14]. The local rules also altered unaccented syllables locally based on the average prosody of the targeted speech. To construct the local and global rules of modification factors, prosody parameters of the storytelling speech syllables are referred and analyzed.

Modification factors for the global and local rules are determined by pitch contour, duration, intensity patterns, tempo and pause of both neutral and storytelling speeches. The factors are calculated by dividing the average desired (i.e. storytelling) prosodic values with the average neutral prosodic values [13]. For example, global rule modification factor of accented syllables for pitch in female storyteller is calculated as follows:

$$\frac{\text{Average pitch of accented syllables for storytelling}}{\text{Average pitch of all syllables in neutral speech}}$$

$$\frac{256.92\text{Hz}}{171.74\text{Hz}} = 1.5$$

Modification factors for duration and intensity of accented syllables for the female storyteller are calculated similarly and tabulated in Table 4. Note that pitch and intensity parameters of unaccented syllables remain unchanged (i.e. NC = no change) for unaccented syllables. Tempo (syllable per second) or speaking rate in the global rule is applied to unaccented syllable only while duration is used for accented syllables. Modification factor of tempo is calculated as such:

$$\frac{\text{Total syllables in storytelling speech}}{\text{Total duration without pause in storytelling speech}}$$

$$\frac{276}{66.3} = 4.1$$

Modification factors for pauses are average pauses computed at phrase and sentence level of the storytelling speeches. The factor at phrase level for the female storyteller is 0.25 and at the sentence level is 0.67 as depicted in Table 4.

Table 4. Global rules modification factors of the female storyteller.

Syllable type	Prosodic parameter	Modification factor
Accented	Pitch	1.5
	Duration	0.9
	Intensity	1.13
Unaccented	Pitch	NC
	Intensity	NC
	Tempo	4.1
NA	Pause (phrase level)	0.25
NA	Pause (sentence level)	0.67

NC – no change; NA – not applicable

Table 5. Local rules modification factors of female storyteller.

Syllable type	Prosodic parameter	Modification factor		
		Initial	Middle	End
Accented	Pitch	2.63	1.72	2.23
	Duration	1.16	NC	1.5
	Intensity	1.1	1.2	1.08
Unaccented	Pitch	1.3		
	Intensity	1.11		
	Tempo	4		
NA	Pause (phrase level)	0.25		
NA	Pause (sentence lev.)	0.67		

NC – no change; NA – not applicable

In Table 5, modification factors of local rules are presented. Factors for local rules are calculated similarly

as global rules. However, local rules of accented syllables are segregated based on the syllable's position in the word (i.e. initial, middle, end). Another set of global and local rules are calculated for the male storyteller and are listed in Table 6 and 7, respectively.

Table 6. Global rules modification factors of male storyteller.

Syllable type	Prosodic parameter	Modification factor
Accented	Pitch	1.47
	Duration	1.14
	Intensity	1.01
Unaccented	Pitch	NC
	Intensity	NC
	Tempo	5.4
NA	Pause (phrase level)	0.42
NA	Pause (sentence level)	0.6

NC – no change; NA – not applicable

Table 7. Local rules modification factors of male storyteller.

Syllable type	Prosodic parameter	Modification factor		
		Initial	Middle	End
Accented	Pitch	2.5	1.38	1.4
	Duration	0.7	1.1	1.5
	Intensity	0.8	1.03	1.06
Unaccented	Pitch	NC		
	Intensity	0.98		
	Tempo	5.4		
NA	Pause (phrase level)	0.42		
NA	Pause (sentence lev.)	0.6		

NC – no change; NA – not applicable

3.1. Pitch

Pitch varies significantly in storytelling speech compared to neutral speech. Analysis of pitch contour in accented syllables of storytelling speeches shows that a falling or raising contour patterns. Therefore, to formulate this rise-fall pattern of pitch contour, non-linear functions are explored. In this paper, the pitch contour patterns are formulated using sine functions [18]. The formulation multiplies all pitch values within a relevant time domain $[t_1, t_2]$ for a given modification factor which is governed by the variable α . The calculation is shown in equation (1).

$$s' = s(t)*(1+(\alpha -1) \times \sin \{[(t - t_1)/(t_3- t_2)] \times 0.5 \times \pi\} \quad (1)$$

where

s Original pitch values at time t
s' Manipulated pitch values at time t
α Pitch modification factor

The pitch modification factor for global rules represented by variable α is substituted with 1.5 (see Table 4) and 1.47 (see Table 6) for female and male, respectively. Meanwhile, pitch modification factors for local rules (see Table 5 and 7) are applied based on the positions of the syllables.

3.2 Duration

In order to efficiently impose the dynamics of storytelling speech prosody on neutral speech, duration should be modified to appropriate levels. For the global rule, two duration constant modification factor are designed. These duration modification factors are uniformly applied to accented and unaccented syllable respectively. In local rule, each position (initial, middle and final syllable) of accented syllable are scaling with different constant modification factors. Unaccented syllable is scale locally proportional to the tempo. The entire modification factor for duration parameters in both rules are showed in rows 4 in Table 4 and Table 5.

3.3 Intensity

Based on the analysis, the average intensity of a storyteller turns out to be higher as compared with neutral speech. Intensity for a speech signal may vary between speakers. Also, intensity may vary due to the recording device and distance to the speaker. In this work, the same recording device applied, and distance recording devices to the speaker are unchangeable. The rules derived from converting intensity of neutral into storytelling style speech are given below.

$$y'(t) = y(t) \times \alpha \quad (2)$$

where:

y Original intensity values at time t
y' Manipulated intensity values at time t
α Intensity modification factor

The intensity modifications factor for female and male as showed at rows 5 in Table 4 and Table 5 respectively. For local rule intensity modification factor at initial, middle and final of the syllable is given in the last three columns with a different value in both Tables.

3.4 Pause

The pause is analyzed from sentences of storytelling and neutral style speech at phrase and sentence level. In this work, the stories recorded by the storyteller have less duration of pauses as compared to neutral speech. Female storyteller average pause length of neutral at phase level is 0.4 s and storyteller 0.25 s. At sentence-level, average pause length of neutral is 0.82 s and storyteller 0.67 s. On the other hand, the male storyteller has the same pause at phase level at 0.4 s. However, average pause length of neutral is 0.87 s at sentence level which is longer than storyteller with an average pause at 0.6. The pause duration (phrase and sentence level) applied to both storytellers as showed last two columns in Table 4 and Table 5.

4 Evaluations and discussions

A subjective test is carried out to evaluate the quality of the synthesized speech. The criteria are considered (naturalness, intelligibility, similarity) compared to the original storytelling speech. In this test, subjects have to listen and evaluate the synthesized speech sentences. 5 native Malay-speaking subjects in the age group of 21-51 participated in the listening test.

For listening test, a complete text story is divided into sentences. The sentences consist of 5 to 21 words. The sentences are converted to synthesize storytelling speech using global and local rule for female and male storyteller. It synthesized using prosodic modification methods HNM. A total of 96 synthesized storytelling speech sentences are collected, which is 48 sentences of each storyteller. Here, the subjects were presented with both original storytelling speech and synthesized storytelling speech. The subjects are given time to judge the quality of the synthesized storytelling speech by comparing the original expressive speech. The evaluation has been done in a quiet place and using proper headphone to prevent external sound. Before evaluation, the subjects are explained about the criteria of naturalness, intelligibility, and similarity. They can replay the synthesized speech for better evaluation. The higher quality synthesized speech was given a mean opinion score of five and poor quality synthesized speech were given of one. The results for the prosodic rule and modification methods for female and male are given in Table 3 and Table 4 respectively.

Table 8**.** Evaluation result for synthesized female storyteller

Prosodic rule	**Natural-ness**	**Intelligibi-lity**	**Similarity**	**Mean (μ)**
Global	2.8	3.0	3.0	2.9
Local	2.4	2.95	2.85	2.73

Table 9**.** Evaluation result for synthesized male storyteller

Prosodic rule	**Natural-ness**	**Intelligibi-lity**	**Similarity**	**Mean (μ)**
Global	2.98	3.8	3.05	3.27
Local	2.8	3.55	2.96	3.1

Based on Table 3 and Table 4, global rules outperform compared to local rule for both storyteller. Mean result for global rule of female storyteller is 2.9 is better than local rule at 2.73 out of 5. While, mean results of male storyteller is 3.27 and 3.1 for global and local rule. Hence, the global rule is used for modification yields a better result than the local rule in both storytellers.

The global rules do not manipulate the entire syllable but only at accented syllable to produce much better quality of speech. It is because modification of syllable sometimes is over exaggerated and quality degenerates that happen in the local rule which modified the entire syllable to the target speech. The values of modification factor in local rule derived based on average are the cause of this matter. These modification factors are not compatible to apply in the entire syllable context. Therefore, in selecting the modification factor, the result should be considered. We can modify the prosody, however, at a certain value it meets a limitation. Better technique needs to explore to find the perfect modification factor that can match with every syllable context in converting neutral to storytelling style speech.

5 Conclusions

In this paper, the process of converting neutral speech from standard text-to-speech (TTS) system to storytelling style speech is carried out by using global and local rule. Five prosodic parameters, namely, pitch, duration, intensity, tempo and pause are analyzed for modification. The modification factors are derived by analyzing the difference between prosodic parameters of neutral and storyteller speech sentences. The subjective test results for naturalness, intelligibility, and the similar value indicated that the global rule was effective in imposing the story style on neutral speech and maintain the quality of speech. However, the used of average to determine the modification factor to suit with all syllable context is not the best idea. Further research needs to explore to improve the synthesized speech in the global and local rule. We believe that local rule can give a better result if the modification factors are applied in the correct circumstances for every syllable. The rules may be extended to develop others speaking styles such as news reading, debate, and conversation.

References

[1] D. Govind, Prasanna, and Mahadeva, “Expressive speech synthesis: a review,” *International Journal of Speech Technology*, vol. 16, 237–260, Oct. (2012)

[2] B. C. Lunce, “Digital Storytelling as an Educational Tool,” *Inidiana Libriaries*, 30, 77–80, (2007)

[3] R. Gelin, C. D’Alessandro, and Q. Le, “Towards a storytelling humanoid robot,” *AAAI Fall Symposium Series on Dialog with Robots*, 137–138, (2010)

[4] J. Adell, A. Bonafonte, and D. Escudero, “Analysis of Prosodic Features Towards Modelling of Emotional and Pragmatic Attributes of Speech,” in *Procesamiento del lenguaje natural*, 277–283, (2005)

[5] A. Silva, M. Vala, and A. Paiva, “The Storyteller : Building a Synthetic Character That Tells Stories,” in *In Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communication Acts to Achive Contextual Embodied Agents, at Autonomous Agents Conference*, 53–58, (2001)

[6] S. J. L. Mozziconacci, “Speech variability and emotion: production and perception,” Technical University Eindhoven, 1998

[7] J. Y. Zhang, A. W. Black, and R. Sproat, “Identifying Speakers in Children ’ s Stories for Speech Synthesis,” 2041–2044, (2003)

[8] I. Iriondo, F. Alías, J. Melenchón, and M. Llorca,

"Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis," *Affective Dialogue Systems*, 197–208, (2004)

[9] J. M. P. M Montero, J M Gutierrez-Arriola, S Palazuelos, E Enriquez, S Aguilera, "Emotional speech synthesis from speech database." (1998)

[10] P. Sarkar, A. Haque, A. K. Dutta, G. R. M, D. M. Harikrishna, P. Dhara, R. Verma, N. P. Narendra, S. K. S. B, J. Yadav, and K. S. Rao, "Designing Prosody Rule-set for Converting Neutral TTS Speech to storytelling style speech for Indian Languages : Bengali , Hindi and Telugu," 0–4, (2014)

[11] M. Schröder, "Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions," *Affective Dialogue Systems*, 3068, 209–220, (2004)

[12] H.-J. Lee, "Fairy Tale Storytelling System: Using Both Prosody and Text for Emotional Speech Synthesis," in *In Convergence and Hybrid Information Technology*, Springer, 317–324, (2012)

[13] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for story telling applications," *IEEE Transactions on Audio, Speech, and Language Processing*, **14(4)**, 1099–1108, (2006)

[14] S. Roekhaut, J. Goldman, A. C. Simon, U. D. M. Umons, U. De, D. De Linguistique, U. De Genève, and I. Langage, "A Model for Varying Speaking Style in TTS systems," 4–7, (2010)

[15] R. Montaño, F. Alías, and J. Ferrer, "Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis," *8th ISCA Workshop on Speech Synthesis*, 171–176, (2013)

[16] P. Boersma and W. David, "Praat, doing phonetics by computer," (2015)

[17] I. Ramli, N. Jamil, N. Seman, and N. Ardi, "An Improved Syllabification for a Better Malay Language Text-to-Speech Synthesis (TTS)," in *International Symposium On robotics and intelligent sensors*, 417–424, (2015)

[18] R. Verma, "Conversion of Neutral Speech to Storytelling Style Speech," (2015)