Prosody Analysis of Malay Language Storytelling Corpus towards Storytelling Speech Synthesis

Izzad Ramli[1],Nursuriati Jamil[1],and Noraini Seman[1]

[1]Digital Image, Audio and Technology Group, Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, 40450 Shah Alam, Malaysia
zadzed89@gmail.com
{liza,aini}@tmsk.uitm.edu.my

Abstract. In this paper, the prosody of the storytelling speech corpus is analyzed. The main objective of the analysis is to develop prosody rules to convert neutral speech to storytelling speech. The speech corpus (neutral and storytelling speech) contains 464 speech sentences, 4,656 words, and 9,584 syllables. It was recorded by three female storytellers, one male professional speaker, two female speakers and two male speakers. The prosodic features considered for analysis are tempo, pause (sentence and phrase-level), duration, intensity, and pitch. Further analysis of the word categories exist in storytelling speech is such as verb, adverb, adjective, noun, conjunction and amplifier are also constructed. The global result showed that mean prosodic of storytelling is higher than neutral speech, especially intensity and pitch. Investigation on the word categories showed that words categorized as adverb, adjective, amplifier and conjunctions are considered prominent and noun and verb only in 20% small amount. The position of the word (initial, middle, last) for word categories in a phrase also proved has versatility.

Keywords: Neutral speech, storytelling speech, prosodic parameters, expressive speech, prosody rule-set, Text-To-Speech (TTS).

1. Introduction

Speech synthesis is widely used in various applications. However, it produces neutral speech sound like news reading speech [1].Therefore, there is a growing need for an expressive speech synthesis to vary the speaking style speech especially for digital communication and humanoid robotic [2]. In order to convert neutral speech to storytelling speech, speech prosody must be manipulated. The efforts of varying prosody can be controlled using rule-based methods [3][4] or data-driven methods [5]. Data-driven methods are preferred these years. However, it required a large amount of training data, very costly and difficult to record so much speech with the same quality

applied. The rule-based method is also not an easy task and needs a proper prosody analysis to understand the linguistic nature and describe the prosody characteristics comprehensively by rules [6]. In literature, the development of storytelling speech synthesis is done using the rule-based method with various prosody analysis and rules [3][7][4].

In general, the prosody analysis in storytelling is based on tempo, pause, duration, intensity and pitch [3]. Theune (2006) in [3] analyzed these prosodic globally and develop the global rules of storytelling application. The prosody is also analyzed in Hindi storytelling at phrase and sentence level [7]. On the other hand, the analysis was also done locally in syllable level on adjective and adverb words [3]. The extra emphasis in adjective and adverb indicates a prominent syllable. This prominent syllable has an extra-long duration, a higher pitch means and rising pitch movement from their local environment [8]. Based on Roekhaut et al.(2010) in [8], the prominent syllable is categorized at initial accents (first syllable of word or phrase) and final accents (last syllable of a word of phrase). The prominent syllable for final accent is usually detected at noun, adjective, verb or adverbs and initial at several word categories.

The position at initial, middle and final words of a phrase are also analyzed [4]. This is because the storytellers produced a unique intonation at initial, middle and final words of a phrase [4]. As an example, the word at the final phrase has an increased duration compared to word located at initial and middle phrase [4]. These word location played important roles for manipulating the types of the speaking style [8]. The prosody based on word location (initial, middle, final) are varied by [7] to develop storytelling speech with various emotion.

In this work, we analyzed the prosody of storytelling corpus in the Malay language. The criteria that are considered are tempo, pauses (phrase and sentence level), the last syllable in the adjective, adverb, noun and verb, an initial syllable in potential word categories and word location (initial, middle, last) in a phrase. Our contribution is comparison value of the accented syllable in word categories with a different position.

This paper is structured as follows. In Section 2, the speech corpus is presented. The global prosody analysis of storytelling is shown in Section 3. Then, local prosody analysis (in syllable level) is described in Section 4. The summary of the results is described in Section5.

2. Storytelling Corpus

In this section, the storytelling corpus used for analysis is discussed. It explains the selection of text corpus, quantitative description of text corpus, storyteller description, the condition of audio recording and audio labeling.

2.1. Text Corpus

Three narrative children short stories from a classic Malaysia's collections of short stories entitled '*200 kisah teladan haiwan'*(200 animal folklores) [11] are selected for recording. The number of sentences, words and syllables are depicted in Table 1.

Table 1. Total sentences, words and syllable in each story

Story	No. of sentences	No. of words	No. of syllables
Si angsa yang berteluremas	12	113	276
Anjingdenganbayang-bayang	9	80	175
Semutdanmerpati	8	98	148
Total	29	291	599

The script of three stories made up a total of 29 sentences, 291 words, and 598 syllables. The scripts do not contain any dialogue and description as our scope in the narrative discourse mode. The language used in the stories fulfills the formal Malay language, with simple words easily understood by the children.

2.2. Audio Corpus Recording

The corpus is recorded by three female storytellers, one male professional speaker, two female and two male speakers. The female storytellers are school kindergarten teachers who have the proper training and experience in delivering storytelling. Their ages range from 30 to 45 years old. A 58 year old professional speaker who has more than 30 years delivering lectures and public speeches is also employed as our storyteller. The four speakers are degree college students who are eloquent speakers and have 3 to 5 years experiences giving public speeches. The corpus exists in two speaking style (neutral and storytelling). Recordings are made in an isolated room in Digital Image, Audio and Speech Technology Group (DIAST) laboratory. The quiet room is equipped with a centralized airconditioner with one door entrance. Background noise of the audio storytelling data was analyzed at 18 dB due to the constant humming of the centralized air-conditioning system. In the end, the speech corpus consists of 48 (8 storyteller x 2 speaking styles x 3 stories) audio.WAV files and down-sampled at 16 kHz with a 16 bits sample size.

2.3. Corpus Labeling

The corpus was annotated using speech analysis tool known as Praat[10] at the sentence, word, and syllable level. There are 48 transcriptions file in textgrid file (.textgrid) that was annotations from 48 audio files. At the first placed, the speech and non-speech region are automatically labeled as speech and silence respectively using Praat. The label was used as guidance for manually labeling end point of the sentence-, word-, and syllable-level. The syllables are labeled based on Malay language syllable structure [17]. The empty labels at word-and syllable- levels are the silence areas which are not annotated and left as blanks.

3. Global Prosody Analysis of Storytelling Speech

In this section, a general prosody analysis in global was analyzed. We have analyzed neutral speech and storytelling speech with respect to different prosody parameters. For this purpose, we have considered 464 neutral sentences and 464 storytelling sentences from eight storytellers. Neutral speech and storytelling speech have been compared to storytelling in global level which is respect to the tempo, pause (phrase and sentence), average syllable duration, average syllable intensity and average syllable pitch as shown in Table 2.

Table 2. Prosodic comparison between neutral and storytelling

Prosodic parameters	Neutral	Storytelling
Mean tempo	4.2	4.48
Mean pause (sentence level)	0.81	0.77
Mean pause (phrase level)	0.29	0.37
Mean syllable duration	0.22	0.2
Mean intensity (dB)	66.18	67.89
Mean pitch (Hz)	191.83	210.04

Tempo is also known as the speaking rate of a person and is calculated based on syllable per second (SPS).Previous research showed that the tempo of storytelling is faster than neutral speech [3]. However, in this research, the observation showed that storytelling speech tempo is faster than neutral speech. Our speech data showed that six out of eight storytellers have higher tempo than neutral speech. This phenomenon occurs because while recording the neutral speech, the storyteller always put their attention on each word pronunciations in an utterance which is time-consuming. The similar phenomenon also occurred in [7].

The pause feature is analyzed at phrase and sentence level in second (s). In our work, a phrase is define as a collection of words and determined by the symbol comma (,) that exist in a sentence. Based on Table 2, the neutral speech has a longer average pause at sentence level compared to storytelling speech. The total average pause at sentence level for neutral speech is longer than storytelling at 0.81s and 0.77s, respectively. However, at phrase level, pause for storytelling is longer than neutral

speech with the total average is 0.37s and 0.29s, respectively. It showed that, at phrase level, storytelling speech has much longer pause before continuing to the next phrase rather than neutral speech.

The syllable duration determines the tempo of the overall speech. The analysis of the duration syllable is to determine the average syllable duration for a certain style and storytellers. The average duration syllable for neutral is longer than storytelling speech. It proved by the total average of the syllable duration of the neutral is 0.22s, and storytelling speech is 0.20s. The average of the syllable duration of storytelling can further be considered for developing a rule for storytelling speech synthesis.

The intensity of the prosody is a measure of loudness in the utterance (Bulut & Narayanan, 2008). It is calculated in unit Decibels (dB). The analysis of the mean intensity of neutral and storytelling speech is 66.18 dB and 67.89 dB. We discover that five storyteller has higher intensity or speech energy compared to neutral speech. It means that storyteller tends to speak louder when delivering a tale as compared to his/her normal reading style. The analysis intensity based on gender also signifies that male will speak louder than female storyteller for both neutral and storytelling speech.

The analysis of mean pitch between neutral and storytelling speech showed increasing of the pitch from neutral to storytelling from 191.83 Hz to 210.04 Hz. It is because five storytellers have higher average pitch as compared to their neutral speech. It indicated that storyteller will increase their pitch in storytelling speaking style. The analysis on the gender showed that female has a high frequency rather that male storyteller and is able to manipulate their pitch with ease.

4. Local Prosody Analysis

The literature mentioned that word categories such as noun, verb, adjectives, and adverbs emphasized the last syllable of a particular word during pronunciations. In this research, we also analyzed accented syllables within conjunction and amplifier (*kata penguat*) and found that they existed in both word categories. However, accented syllable of the amplifieris located at the initial syllable of a word. The total words for all three stories based on word categories are shown in Figure 1.

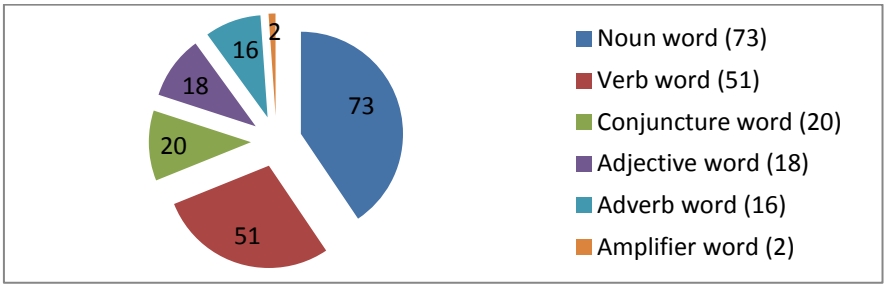

Fig.1. Tabulated of selected word categories in story

The analysis of each word categories is done with comparing words in neutral with the storytelling speech. The prosody parameters compared are duration, intensity and pitch. Figure 2 shows the percentage number of words that storytellers tend to increase their duration, intensity, and pitch for word categories.

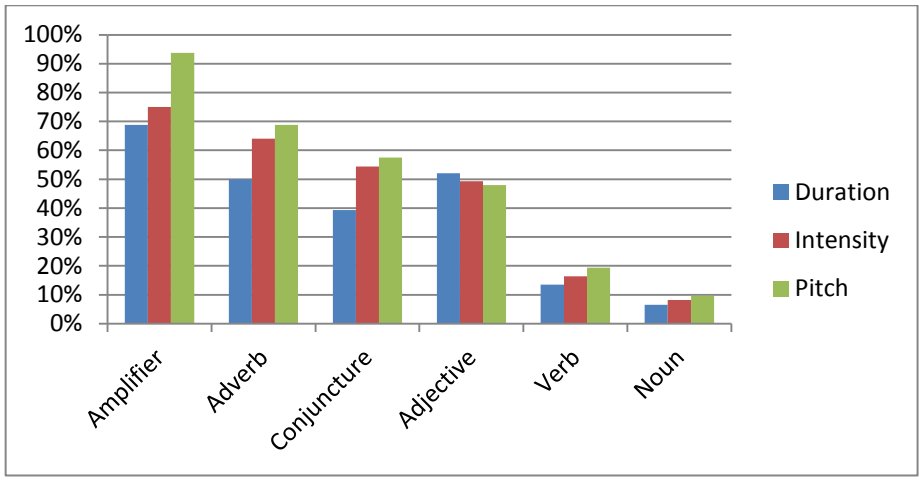

Fig.2. Percentage storyteller increase prosody with word categories

Based on Figure 2, it shows that less than 20% of verb and noun words have the minimal increase in duration, intensity and pitch at the last syllable. It indicates that only a small percentage of the verb and nonwords contain accented syllable. On the other hand, amplifier words have the highest accented syllable at more than 65% compared to other word categories. The overall observation concludes that only certain words in each word category have accented syllables and these words are used as the basis of the rules.

The words are further examined based on their positions (i.e. initial, middle and last) in a phrase or sentence. Table 3 shows the percentage increased prosody of accented syllable in the initial, middle, last position of a word. The comparison is done between neutral speech and storytelling speech. For an adjective, the accented syllable at the last word has higher pitch increased that is 40% as compared to the middle word at 18%. As an example, the last word *berat* in the sentence *telur itu sangat berat* has accented syllable of the syllable *rat*. The syllable *rat* in storytelling speech has an increased pitch by a factor of 1.4 times from syllable *rat* in the neutral speech. However, the duration of the accented syllable in the middle of a word is longer than accented syllable of the last word.

For adverb word category, there is no significance difference (less than 10%) of intensity and pitch between initial, middle and last position. However, duration of accented syllable in adverb longer than initial and middle position. As for amplifier word category, there is no changes initial and last word because of the non-existence.

Table 3. Comparison of accented syllable based on position

Position	Duration	Adjective Intensity	Pitch	Duration	Adverb Intensity	Pitch
Initial word	NC	NC	NC	+52%	+4%	+21%
Middle word	+54%	+6%	+18%	+40%	+6%	+29%
Last word	+26%	+6%	+40%	+62%	+9%	+24%
		Amplifier			Conjunction	
Initial word	NC	NC	NC	+49%	+6%	+21%
Middle word	+99%	+9%	+31%	+35%	+6%	+20%
Last word	NC	NC	NC	NC	NC	NC
		Noun			Verb	
Initial word	+36%	+6%	+22%	+103%	+3%	+15%
Middle word	+39%	+7%	+26%	+36%	+7%	+28%
Last word	+45%	+6%	+25%	+31%	+6%	+28%

Analysis on the amplifier in the middle position for our speech data is described. It is interesting to note that duration of the amplifier is increased by 99% for the middle word, which is the highest increment compared to others. A 40% increase in pitch of the last word and 54% increase of duration in the middle word of adjectives are observed. No changes are noted for the initial words. Conjunction words have a slightly higher increase in duration at the initial word compared to middle word. No analysis on the conjunction at the last word and we described it as no change for this research.

Even though nouns and verbs have less than 20% accented syllables, analysis revealed that noun at the last word showed that an increased duration of 45% that is higher than noun located at the initial and middle word. Intensity and pitch do not show much difference (less than 10% different). Verb words, however, shows an increase of 103% duration at the initial word. Nevertheless, it has the lowest intensity increased compared to all the other word categories.

5. Summary and Future Work

In this paper, we have analyzed the prosody of the storytelling as compared to the neutral speech. The corpus has been presented, and the analysis of the corpus has been described. The result discussed the difference in the global prosody of the storytelling and the neutral speech. The analysis in local prosody showed that only selected word in word categories has accented syllable. The words were determined and considered during development of the storytelling rule. Our future work is to develop the prosody rule of the storytelling based on the data analysis especially on percentage increased prosody in word categories at different word position which is our contribution to this research.

References

1. Khaw, Y.J., Tan, T., Sciences, C.: Preparation of MaDiTS Corpus for Malay Dialect Translation and Speech Synthesis System. In: Speech, language and Audio in Multimedia Workshop (SLAM 2014). pp. 53–57 (2014).
2. Gelin, R., D'Alessandro, C., Le, Q.: Towards a storytelling humanoid robot. AAAI Fall Symposium Series on Dialog with Robots. 137–138 (2010).
3. Theune, M., Meijs, K., Heylen, D., Ordelman, R.: Generating expressive speech for story telling applications. IEEE Transactions on Audio, Speech, and Language Processing. 1099–1108 (2006).
4. Sarkar, P., Haque, A., Dutta, A.K., M, G.R., Harikrishna, D.M., Dhara, P., Verma, R., Narendra, N.P., B, S.K.S., Yadav, J., Rao, K.S.: Designing Prosody Rule-set for Converting Neutral TTS Speech to storytelling style speech for Indian Languages : Bengali , Hindi and Telugu. 0–4 (2014).
5. Mustafa, M.B., Don, Z.M., Ainon, R.N., Zainuddin, R., Knowles, G.: Developing an HMM-based speech synthesis system for Malay: a comparison of iterative and isolated unit training. IEICE Transactions on Information and Systems. (2014).
6. Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous speech corpus of Japanese, (2000).
7. Verma, R.: Conversion of Neutral Speech to Storytelling Style Speech. (2015).
8. Roekhaut, S., Goldman, J., Simon, A.C., Umons, U.D.M., De, U., Linguistique, D. De, Genève, U. De, Langage, I.: A Model for Varying Speaking Style in TTS systems. 4–7 (2010).
9. Issam Rebai, Y.B.: Arabic text to speech synthesis based on neural networks for MFCC estimation. In: Computer and Information technology (WCCIT). pp. 1–5. IEEE (2013).
10. Boersma, P., David, W.: Praat, doing phonetics by computer.
11. Montaño, R., Alías, F., Ferrer, J.: Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis. 8th ISCA Workshop on Speech Synthesis. 171–176 (2013).
12. Bulut, M., Narayanan, S.: On The Robustness of Overall F0- Only Modifications to the Perception of Emotions in Speech. Journal of the Acoustical Society of America. 123, 4547–4558 (2008).