

# An Improved Pitch Contour Formulation for Malay Language Storytelling Text-to-Speech (TTS)

Izzad Ramli, Nursuriati Jamil, Noraini Seman

Faculty of Computer and Mathematical Sciences, Universiti  
Teknologi MARA, Shah Alam, Malaysia  
zadzed89@gmail.com, {liza,aini}@tmsk.uitm.edu.my

Norizah Ardi

Academy of Language Studies, Universiti Teknologi  
MARA, Shah Alam, Malaysia  
norizah@salam.uitm.edu.my

**Abstract**— In this paper, an improved pitch contour formulation is introduced by modifying the existing pitch contour sinusoidal function. The aim is to convert neutral speech into storytelling speech in Malay Language. Our speech datasets (neutral and storytelling speech) were recorded by a male and a female professional speaker. They contain 116 speech sentences, 1,164 words, and 2,732 syllables. For storytelling speech, 124 prominent syllables are detected using Prosogram tool. These prominent syllables are further categorized into six clusters of pitch contour. Distance measurements using one minus Pearson correlation is done to assess the similarity of the proposed pitch contour formulae to the original storytelling pitch contour. The proposed pitch contour sinusoidal function is also compared with the existing pitch contour function used by previous work. The results showed that the proposed pitch contour formulation performed better than the existing pitch contour formulae.

**Keywords**—Storytelling; pitch contour; prosodic parameters; prosody modification; Text-to-Speech (TTS)

## I. INTRODUCTION

Intonation is one of the crucial element in the development of storytelling text-to-speech. Basically, varying intonation can be done with manipulating prosodic features such as duration, intensity and pitch [1]. Duration and intensity are easily manipulated in a straight forward manner. However, pitch modification depends on the pitch contour with respect to the original storytelling [2]. The reason being is neutral speech's pitch contour is generally well-behaved and fairly restricted in pitch range and variation [3] compared to storytelling speech. A simple increase or decrease of the mean pitch is inadequate to produce synthesized storytelling speech [1].

Manipulation of pitch contour had been done in several ways in speech synthesis even in singing synthesis. In [4] [5], pitch contour template was designed for personalized singing synthesis. The pitch contour was derived from the actual pitch contour making it more natural than modified pitch contour. However, the template designed is dependent to the speaker. Another way is predicting the pitch contour using Classification And Regression Tree (CART) [6] which needs statistical analysis to develop the model. However, the CART model is also speaker-dependent. The pitch contour modification method provides flexibility to modify pitch contour across speakers and circumstances providing better reliability. The related works reported in the literature for

pitch contour modification at a different level are documented in [7], [8], [9], [10], [11].

A study of pitch contour analysis done by [3] identified six clusters of pitch contour of prominent syllables commonly found in storytelling speech corpus. The pitch contours were clustered using hierarchical clustering technique of foot-based pitch contour. All pitch contours showed variations in terms of up-down movement, location of pitch peaks (middle or last), and shape of the pitch contour. Pitch contours are generally formulated using sinusoid function as done by [12] and [13]. In [2], a default pitch contour was designed using a constantly rising and falling of pitch. The peak of the pitch contour is formulated only at the middle of the contour as illustrated in Fig. 1. This up-down pitch contour are constantly rising and falling with the same degree [2],[12],[13]. Therefore, it is unable to match the original storytelling contour which has a different degree of rising and falling factor as shown in Fig. 2. Even by controlling the parameters of the contour shape are difficult to produce the desired contour.

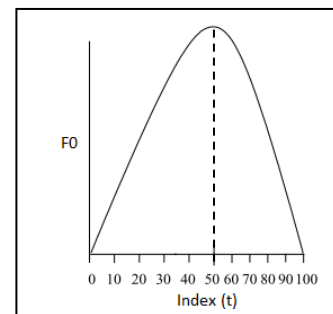


Fig. 1. Default pitch contour

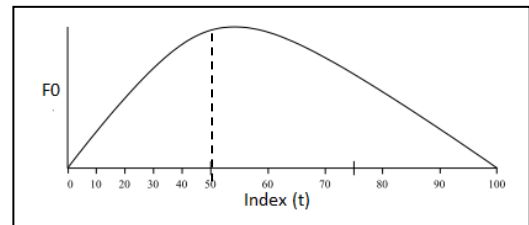


Fig. 2. Illustrated of original pitch contour

In this paper, we analysed the pitch contour of prominent syllables using our storytelling speech datasets based on six

clusters of pitch contour as proposed by [3]. Then, we proposed a new formulation of the six pitch contours based on syllables because our Malay storytelling TTS system used syllable-based unit selection approach. In addition, Malay language is an alphabetic language with salient syllabic structures [14].

The paper is structured as follows. Section II describes the recordings of our storytelling speech datasets. Section III details the pre-processing step that involves labelling the recordings at the sentence, word and syllable levels. In Section IV, we describe the analysis of the pitch contour at prominent syllables. Section V elaborates the improved pitch contour formulation and Section VI presents the evaluation of the proposed formulation. Results and discussion are presented in Section VII followed by conclusion and future work in Section VIII.

## II. STORYTELLING SPEECH DATASET

The speech dataset is recorded by one male and one female professional speakers. A 58-year old male professional speaker and a professional female radio caster is employed as our storyteller. Three narrative children short stories from a classic Malaysia’s collections of short stories entitled ‘200 kisah teladan haiwan’ (200 animal folklores) are selected for recording. The number of sentences, words and syllables are depicted in Table I.

The speech recorded is in two speaking styles (i.e. neutral and storytelling). The recorded neutral speech is free of all possible emphasis or stress such as news reading. Therefore, the storyteller must maintain their vocal qualities in term of intelligibility, timbre, diction and pronunciation. Recordings are made in an isolated room in Digital Image, Audio and Speech Technology Group (DIAST) laboratory, Universiti Teknologi MARA. The quiet room is equipped with a centralized air conditioner 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 dataset consists of 12 (2 storyteller x 2 speaking styles x 3 stories) audio .WAV files down-sampled at 16 kHz with a 16 bits sample size. A total of 116 sentences (29 sentences x 2 speaking styles x 2 storytellers), 1,164 words (98 words x 2 speaking styles x 2 storytellers), and 2,732 syllables (683 syllables x 2 speaking styles x 2 storytellers) are collected from the speakers.

TABLE I. STORYTELLING SPEECH DATASET

Story	No. of sentences	No. of words	No. of syllables
<i>Si angsa yang bertelur emas</i>	12	113	276
<i>Anjing dengan bayang-bayang</i>	9	80	175
<i>Semut dan merpati</i>	8	98	232
<i>Total</i>	29	291	683

## III. PRE-PROCESSING

The storytelling dataset is annotated manually using speech analysis tool known as Praat [14] at the sentence, word, and syllable level as shown in Fig. 3. The annotation produced 12 transcriptions of text grid files. The syllables are labelled based on the Malay language syllable structure [15]. The silence areas at word-and syllable- levels are not annotated and left as blanks. Pitch information is extracted every 5 ms, using STRAIGHT tool.

## IV. PITCH CONTOUR ANALYSIS

In general, the pitch contour of a phrase for the neutral speech and storytelling speech extracted by Praat is showed in Fig. 4. The observation shows that neutral pitch contour is well-behaved and likely flat and storytelling shows up and down pitch in the pitch contour.

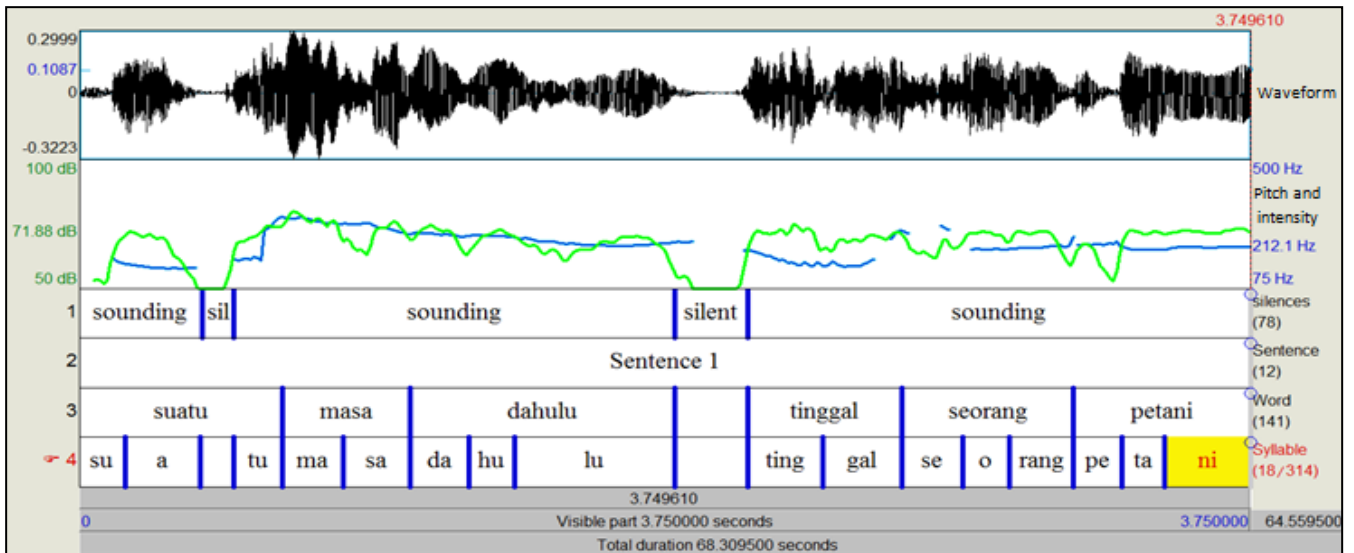


Fig. 3. Speech labeling and segmentation

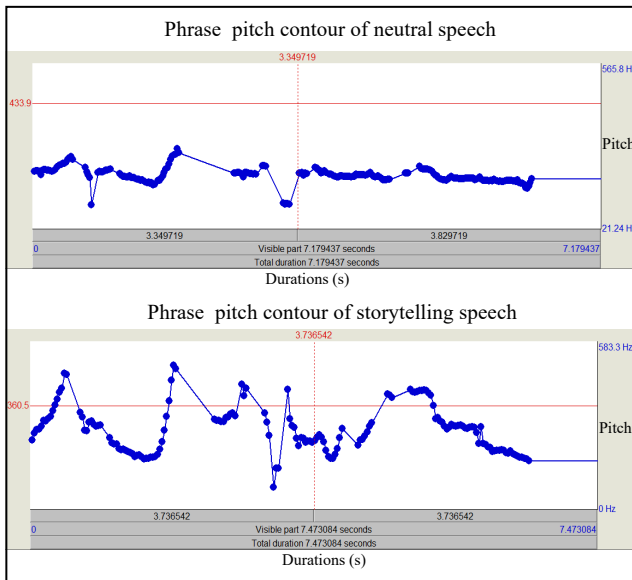


Fig. 4. Prominent syllable detection

After that, prominent syllables of the storytelling datasets are detected using Prosogram tool based on typical acoustic parameters of prosody (i.e. pitch, intensity and duration). Two main processes involved are 1) units segmentation and pitch stylisation, and 2) computation of acoustic parameters for detection.

The first process, the stylized pitch contour is computed and draw based on a tonal perception model for each syllable. In the syllable, vocalic nucleus segments which found as voiced segment with sufficient intensity (difference thresholds relative to the local peak) is stylized its pitch. The syllable can be more than one stylized pitch. These stylized pitches can be flat or with a melodic slope.

In the next process, two acoustic prosodic parameters are computed for each nucleus (i.e. duration and maximum pitch). The parameters also computed to the adjacent syllables (the preceding two and the following one with weight) to obtain the local value of these parameters. On the basis of these parameters, the discrimination analysis was conducted on to determine the prominent or non-prominent syllable. The prominent syllable was labeled (with label P) and non-prominent syllable (no label).

124 syllables out of 683 detected as prominent syllables. In general, prominent syllable will have an extra-long duration (i.e. longer than the local duration), a higher pitch (i.e. higher than the local pitch) or pitch movement (increasing or decreasing stylized pitch contour) [1]. An example of a prominent syllable is demonstrated in Fig. 5.

The syllable *rat* is detected as prominent and labeled as (P) because it has pitch movement drastically. The syllable *be* is not prominent syllable because it does not have extreme pitch change, longer duration (i.e. duration syllable *be* is 0.14 seconds which local is 0.32 seconds) or higher pitch (pitch syllable *be* is 263.09 Hz which local is 302.91 Hz).

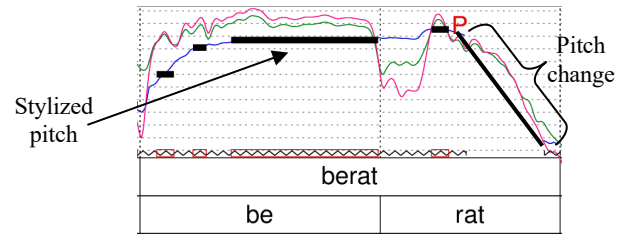


Fig. 5. Prominent syllable detection

Then, the pitch contours are clustered into six clusters and shown in Table II. The y-axis is pitch (Hz) and the x-axis is the normalized time set to 100 data points. Cluster 1 is the largest cluster, containing 33% of the pitch contours (41 out of a total 124 pitch contours). The contours represented in this cluster exhibit the default up-down movement. It starts at a lower pitch and increasing steadily until it reaches the middle of contour and falling rather quickly to the end. In our dataset, Cluster 1 frequently occurs at the end of a phrase at 52% (21 contours out of 40) or end of the sentence at 48% (20 contours out of 40).

Cluster 2 contains 26% of the pitch contours (33 contours out of 124). They exhibit a gradual descent from the peak to the bottom. However, the last contour shows a slight increase of the contour. Interestingly, 76% (25 out of a total 33) occurrence of Cluster 2 is at the end of a sentence and 24% (25 out of a total 33) occurrences are at phrase-final position. It shows that this contour exists more in syllables positioned at the end of a sentence than in syllable at the end of a phrase.

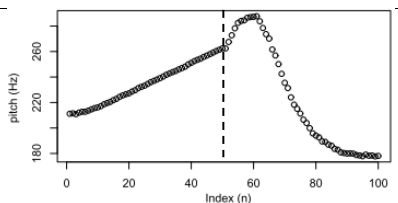
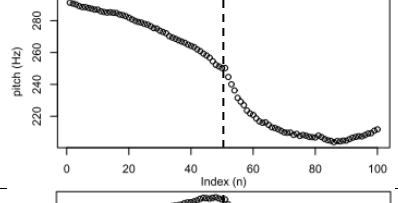
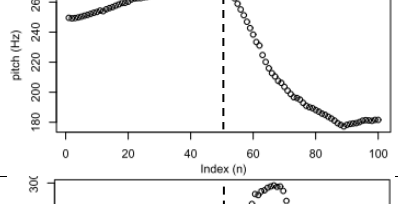
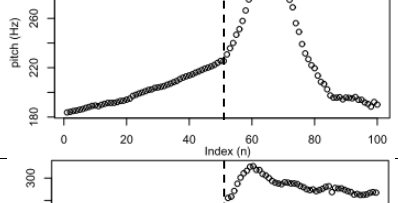
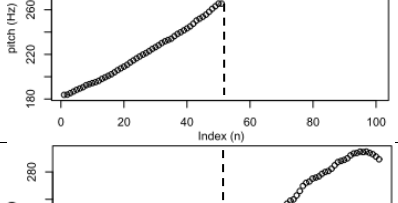
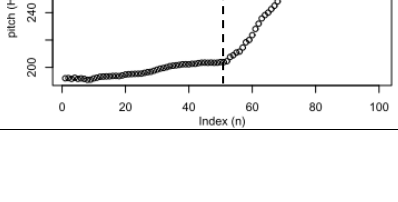
Cluster 3 contains 26% of the pitch contours (32 out of 124). It starts at a higher pitch with a comparison to Cluster 1. The growing accent curve can be observed until the end of the plateau and reducing until the end. This type of contour is observed in phrase-final position in 53% of the cases.

Cluster 4 contains 14% of the pitch contours (17 out of 124). The pitch contours in this cluster are similar to those in Cluster 1. The difference is that the peak occurs later and is higher than the pitch contour of Cluster 1. The peak location is late at the pitch contour. This cluster occurred more at the end of a sentence at 82% (14 out of 17).

In this research, pitch contour of Cluster 5 comprises only 1% of the pitch contours (1 out of 124). It displays a gradual incline until the middle of contour and followed by a plateau until the end. These contours exhibit the continuation rise and mostly followed by contours from Cluster 2.

Cluster 6 contains a generally rising pitch. Our 124 pitch contours of prominent syllables do not match with this type of cluster. However, we believe it occurs in other non-prominent syllables because this contour generally precedes contours from Cluster 2 and 3 [3].

TABLE II. SIX PITCH CONTOUR CLUSTER

Cluster	Pitch contour	Total syllable
Cluster 1		41
Cluster 2		33
Cluster 3		32
Cluster 4		17
Cluster 5		1
Cluster 6		0

## V. A MODIFIED SINUSOIDAL PITCH CONTOUR FORMULATION

The sinusoidal formula as in Eq. (1) is used to formulate storytelling pitch contour in [12] and [13]. The formula modifies a syllable pitch contour by manipulating variable  $a$  and  $\beta$ . However, the manipulation is difficult to match perfectly for certain clusters with different contour shape at the first half contour and second half contour.

$$m'(t) = s(t) * (1 + a \times \sin(\frac{t-t_1}{t_2-t_1}) \times \beta \times pi) \quad (1)$$

- $m'(t)$  Modified pitch contour
- $s'(t)$  Neutral pitch contour
- $a$  Desired maximum shift
- $\beta$  Constant determined contour shape

Identify applicable sponsor/s here. If no sponsors, delete this text box (sponsors).

In order to gain similarity modified contour to the desired pitch contour cluster, we proposed a modified formulation of the pitch contour by separating a pitch contour of a syllable into two sections (i.e. first half and second half). The first half of pitch contour is formulated using Eq. (2) and controlled by parameters  $a$  and  $b$ . While the second half of the pitch used Eq. (3), with  $\alpha$  and  $\beta$  as the control parameters. The parameters  $b$  and  $\beta$  determined the contour shape which is constantly increasing, rising and then falling, constantly falling or falling and then rising. Lastly, these two modified contours are combined to produce a complete pitch contour using Eq. (4).

$$p'(t) = s(t) * (1 + a \times \sin(\frac{t-t_1}{t_2-t_1}) \times 2b \times pi) \quad (2)$$

$$f'(t) = p'(t) * (1 + \alpha \times \sin(\frac{t-|t_2/2|}{t_2-|t_2/2|}) \times \beta \times pi) \quad (3)$$

$$m'(t) = \begin{cases} p'(t) & \text{if } (t < |t_2/2|) \\ f'(t) & \text{if } (t \geq |t_2/2|) \end{cases} \quad (4)$$

- $m'(t)$  Modified pitch contour
- $s'(t)$  Neutral pitch contour
- $p'(t)$  Modified pitch contour at first half
- $f'(t)$  Modified pitch value at second half
- $a$  Desired maximum pitch shift for first half
- $b$  Constant determines contour shape for first half
- $\alpha$  Desired maximum pitch shift for second half
- $\beta$  Constant determines contour shape for second half

The variables (i.e.  $a$ ,  $b$ ,  $\alpha$ ,  $\beta$ ) and their corresponding values used for modulating the pitch contour from Cluster 1 to 6 are shown in Table III. It also shows the neutral and modified pitch contour separated based on the six clusters. As an example, in order to create a pitch contour resemble to pitch contour of cluster 1 as shown in Table II, the neutral pitch contour is converted to target pitch contour (i.e. pitch contour of cluster 1) using Eq. (4) with the variables value of  $a$ ,  $b$ ,  $\alpha$ ,  $\beta$  is  $a, 0.1, -\alpha, 0.5$  respectively. The value of  $a$  (desired maximum shift) is determined based on the analysis result and calculated using Eq. (5).

$$\text{Desired maximum shift} = \frac{\text{Desired maximum pitch increase}}{\text{Average pitch}} \quad (5)$$

As example;

$$\begin{aligned} \text{Desired maximum shift} &= \frac{40 \text{ Hz}}{200 \text{ Hz}} \\ &= 0.2 \end{aligned}$$

Based on the analysis result on the pitch, let desired maximum pitch increase is 40 Hz, it will divide with the average pitch (i.e. 200 Hz) to get the 0.2 as desired maximum shift. The value 0.2 will be used in Eq. (4) which previously representation with the variable  $a$ . Every cluster has a unique variable value of  $b$  and  $\beta$  that will determine the up and down contour and shape for the first half and the second half of the contour respectively. These variables value mainly contribute to the development of six pitch contour clusters in this research.

TABLE III. PROPOSED PITCH CONTOUR PARAMETERS FOR SIX CLUSTERS

Clusters	$a$	$b$	$\alpha$	$\beta$
Cluster 1	$a$	0.1	$-a$	0.5
Cluster 2	$a$	0.2	$-a$	0.8
Cluster 3	0.1	0.1	$a$	0.2
Cluster 4	0.1	0.2	$a$	0.9
Cluster 5	$-a$	0.1	$-a$	0.9
Cluster 6	$a$	0.5	0.1	0.1

## VI. EVALUATION

The main aim of evaluation is to assess the modified pitch contour formulation (i.e. Eq. 4) against the pitch contour formulation used in previous studies (i.e. Eq. 1). All 124 prominent syllables identified earlier are extracted from the

neutral speech and modified using pitch contour formulation of Eq. 1 and our proposed formulation of Eq. 4. The original storytelling pitch contours of all 124 prominent syllables are also extracted as benchmark dataset. These datasets are summarized in Table IV.

TABLE IV. COLLECTION OF PITCH CONTOURS

Pitch contour	Total syllables
Original pitch contour from storytelling syllable	124
Modified pitch contour using (Eq. 4) from neutral syllable	124
Modified pitch contour using (Eq. 1) from neutral syllable	124

The evaluations are done by comparing the distance between the original pitch contour with the modified pitch contours produced by Eq. (1) and our proposed Eq. (4). The distance between each pair of pitch contours is calculated using one minus the Pearson product moment correlation as in Eq. (6) [3]. This equation calculates the difference in pitch height or range by subtracting pitch contour with a mean pitch and dividing them by their standard deviation. A distance value,  $D$  that is closer to 1 indicates a higher similarity.

$$D = 1 - cor(F_{0i}, F_{0j})$$

$$= 1 - \left( \frac{1}{n-1} \sum \left( \frac{F_{0i} - \bar{F}_{0i}}{sdF_{0i}} \right) \left( \frac{F_{0j} - \bar{F}_{0j}}{sdF_{0j}} \right) \right) \quad (6)$$

$D$  Distance value  
 $F_{0i}$  Neutral pitch contour  
 $F_{0j}$  Modified pitch contour  
 $n$  Length pitch contour

## VII. RESULTS AND DISCUSSION

Table V shows the result of the distance measurements achieved by pitch contour formulation of Eq. (1) and our improved pitch contour formulation of Eq. (4). As can be seen from the table, the average distance value for Eq. (1) is 0.65 compared to Eq. (4) at 0.71. The details results also shows that our proposed pitch contour formulation performed slightly better than the previous pitch contour formulation in which Cluster 1 until Cluster 4 have higher similarity with the original storytelling pitch contour. Pitch contour of Cluster 5 attained equal similarity values for Eq. (1) and Eq. (4).

Upon closer inspection of the results, we discovered several factors affecting the results of distance measurements. Several neutral pitch contours of the prominent syllables spoken by the storyteller are not flat contour which have pitch movement. This is due to unnecessary emphasis of syllable during neutral speech recording. Upon modification of these pitch contours, the desired pitch contour is not realized. Another factor is due to the vibrato produced by the vibrating glottis resonance of the vocal cord. The existence of vibrato in the pitch contour resulted as noises affecting the shape of the pitch contour. These two factors significantly affect the distance measurements.



TABLE V. EVALUATION RESULTS

Cluster No.	Equation (1) [4][5]	Improved Equation (4)
Cluster 1	0.62	0.64
Cluster 2	0.64	0.69
Cluster 3	0.59	0.65
Cluster 4	0.60	0.74
Cluster 5	0.91	0.91
Cluster 6	NA <sup>a</sup>	NA <sup>a</sup>
Average	0.65	0.71

<sup>a</sup> NA-not applicable

## VIII. CONCLUSION AND FUTURE WORK

This paper introduced a modified formulation of the pitch contour by separating a pitch contour of a prominent syllable into two sections. Six different pitch contours are produced from the proposed formula. The pitch contour analysis also showed that higher percentage of pitch contour clusters occurred either at the end of a sentence or phrase. The proposed pitch contour formulation needs further refinement to solve the factors affecting the shape of the pitch contour leading to a dissimilar distance measurement. We also need to increase the size of our storytelling speech dataset to further validate our present results. The improved pitch contour formulation will be applied to our storytelling TTS and evaluated using perceptive test. The duration and intensity information will be added to enrich the intonation of the storytelling speech. These are our considerations for future work.

## REFERENCES

- [1] S. Roekhaut, J. Goldman, A. C. Simon, U. D. M. Umons, U. De, D. De Linguistique, U. De Genève, and I. Language, "A Model for varying speaking style in TTS systems," in *Fifth International Conference on Speech Prosody*, 2010, pp. 4–7.
- [2] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for story telling applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006.
- [3] E. Klabbbers and J. Santen, "Clustering of foot-based pitch contours in expressive speech," *Fifth ISCA Workshop on Speech Synthesis*, pp. 73–78, 2004.
- [4] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4509–4512, 2012.
- [5] Z. Nurmaisara, N. Jamil, S. Salleh, and N. Rahman, "Synthesizing Asli Malay Song: Transforming Spoken Voices into Singing Voices.," in *The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications*, 2014, pp. 303–310.
- [6] C. Thomas, P. Gokul, N. Thomas, and D. P. Gopinath, "Synthesizing intonation for malayalam TTS," in *International Conference on Control, Communication and Computing India*, 2015, pp. 522–527.
- [7] R. S. Deo, "Pitch contour modelling and modification for expressive marathi speech synthesis," pp. 2455–2458, 2014.
- [8] M. Gurunath Reddy and K. S. Rao, "Neutral to happy emotion conversion by blending prosody and laughter," in *8th International Conference on Contemporary Computing, IC3*, 2015, pp. 342–347.
- [9] M. C. Anil and S. D. Shirbahadurkar, "Pitch and duration modification for expressive speech synthesis in Marathi TTS system," in *International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC*, 2015, pp. 3–6.
- [10] B. Gerazov and I. Zaron, "Analysis of extracted pitch contours across speakers for intonation modelling in TTS synthesis," in *International Symposium on Communications, Control and Signal Processing*, 2012, pp. 2–4.
- [11] A. H. Warsi, T. Basu, K. Hirose, and H. Fujisaki, "Analysis and synthesis of F0 contours of declarative, interrogative, and imperative utterances of Bangla," in *International Conference on Speech Database and Assessments*, 2012, pp. 56–61.
- [12] R. Verma, P. Sarkar, and K. S. Rao, "Conversion of Neutral Speech to Storytelling Style Speech," in *8th International Conference on Advances in Pattern Recognition*, 2015, pp. 1–6.
- [13] 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," in *Contemporary Computing (IC3), Seventh International Conference*, 2014, pp. 673–677.
- [14] L. W. Lee, H. M. Low, and A. R. Mohamed, "A comparative analysis of word structures in Malay and English children ' s stories," vol. 21, no. 1, pp. 67–84, 2013.
- [15] 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*, 2015, pp. 417–424.