



How to cite this article:

Ramli I., Jamil, N., & Seman, N. (2021). An iterated two-steps sinusoidal pitch contour formulation for expressive speech synthesis. *Journal of Information and Communication Technology*, 20(4), 489-510. <https://doi.org/10.32890/jict2021.20.4.2>

An Iterated Two-Step Sinusoidal Pitch Contour Formulation for Expressive Speech Synthesis

Izzad Ramli, Nursuriati Jamil & Noraini Seman
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Malaysia

izzad, aini @tmsk.uitm.edu.my
lizajamil@computer.org

Received: 18/11/2020 Revised: 14/1/2021 Accepted: 8/3/2021 Published: 27/9/2021

ABSTRACT

Intonation generation in expressive speech such as storytelling is essential to produce high quality Malay language expressive speech synthesizer. Intonation generation, for instance explicit control, has shown good performance in terms of intelligibility with reasonably natural speech; thus, it was selected in this research. This approach modifies the prosodic features, such as pitch contour, intensity, and duration, to generate the intonation. However, modification of pitch contour remains a problem because the desired pitch contour is not achieved. This paper formulated an improved pitch contour algorithm to develop a modified pitch contour resembling the natural pitch contour. In this work, the syllable pitch contours of nine storytellers were extracted from their storytelling speeches to create an expressive speech syllable dataset called STORY_DATA. All the shapes of pitch contours from STORY_DATA were analyzed and clustered into the standard six main pitch contour clusters for storytelling. The clustering

was performed using one minus the Pearson product moment correlation. Then, an improved iterative two-step sinusoidal pitch contour formulation was introduced to modify the pitch contours of a neutral speech into an expressive pitch contour of natural speeches. Overall, the improved pitch contour formulation was able to achieve 93 percent high correlated matches, indicating the high resemblance as compared to the previous pitch contour formulation at 15 percent. Therefore, the improved formula can be used in a text-to-speech (TTS) synthesizer to produce a more natural expressive speech. The paper also discovered unique expressive pitch contours in the Malay language that need further investigations in the future.

Keywords: Pitch contour formulation, prosody modification, speech synthesis, storytelling.

INTRODUCTION

Expressive speech synthesis has gained interest in the last decade. As generally known, expressive text-to-speech can be widely used in a variety of applications such as diagnosis and therapy for communication disorders like dyslexia or autism (Plaisant et al., 2000). It is also crucial for the growth of digital speech (Lunce, 2007) and humanoid robots (Gelin et al., 2010). There are several studies in expressive speech synthesis, especially in emotional speech (Chang et al., 2014; Um et al., 2020; Yadav & Rao, 2015). Previous research focused on the spectral characteristics of the various emotions. In the Malay language, there are some works on emotions to be added into storytelling speech synthesis for a more natural storytelling (Jamil et al., 2017; Md Saad et al., 2018). Other than emotion, the research of speaking style synthesis was also done by several scholars (Kato et al., 2020; Raúl & Frances, 2017; Verma et al., 2015). Expressive speech synthesis approaches are therefore introduced to generate natural human-like synthesized speech with emotion and speaking style. As stated by Schröder (2009), the methods of expressive speech synthesis are categorized as playback, implicit, and explicit control. In implicit control, Hidden Markov Model is commonly useful to change the expressivity developed by the statistical models using the trained expressive speech database. Meanwhile, playback approaches concatenate large expressive speech units, such as syllables or

phonemes, using unit selection method. Both approaches need a large dataset. In contrast, explicit control, which is commonly used in storytelling, does not need a large speech corpus (Verma et al., 2015). Therefore, this study focuses on using the explicit control approach to produce expressive speech.

RELATED STUDIES

In the study by Montaña et al. (2013), the explicit control approach was shown to be highly flexible in synthesizing many different emotions and speaking styles of the speakers. This approach was also proven to be capable of producing natural synthesized speech with good intelligibility (Verma et al., 2015). In explicit control, the neutral speech is modified using the prosodic properties by referring to the analysis of the prosodic information in expressive speech. It manipulates prosodic characteristics such as duration, intensity, and pitch, and can produce various intonations (Roekhaut et al., 2010). Sometimes, pause is also considered for modification (Hamzah & Jamil, 2019). Duration and intensity are the most easily adjusted. However, the modification of the pitch must depend on the pitch contour extracted from the original expressive pitch contour (Theune et al., 2006). Pitch contour is incredibly significant for varying the speaking style of a speech (Gu & Jiang, 2015). The pitch contour of neutral speech is normally well-behaved and is reasonably small in pitch spectrum and variance (Klabbers & Santen, 2004) as compared to an expressive speech. To generate synthesized articulated voice, a straightforward modification to the mean of the pitch is insufficient (Roekhaut et al., 2010). Therefore, there is a need for an algorithm to manipulate the shape of the pitch contour.

Based on the current literature, there are several algorithms to modify the syllable's pitch contour for speech synthesis. The pitch contours were generally formulated using sinusoidal function as done by Sarkar et al. (2014) and Verma et al. (2015). However, the formula was designed for a default pitch contour that had constant rising and falling of pitch contours (Verma et al., 2015). Another problem was that the produced pitch contour would have a peak that is located at the center of the contour. Most of the time, the existing formulation was unable to produce a modified pitch contour that matched with the

desired pitch contour even by controlling the formula's parameters. Therefore, in Ramli et al. (2016), they introduced a two-step sinusoidal pitch contour formulation to tackle that problem and improved the similarity rate. However, the modification of the pitch contour was designed only to modify the prominent syllables. Furthermore, non-prominent syllables were not analyzed and modified into natural syllables. Therefore, the novelty of this paper is the production of an improved two-step formulation that can modify both the prominent and non-prominent syllables. This paper is structured as follows. The elaboration about the recording materials and the description of the pre-processing stage involved in this research are presented in the next section. Then the pitch contour analysis and the improved formulation of the pitch contour are described. This is followed by the results and discussion. The conclusion and future research are presented in the final section of the paper.

METHODOLOGY

In this section, a detailed elaboration on the formulation of the new speech contour algorithm is presented. The methodology entails the data collection, speech pre-processing, pitch contour extraction, pitch contour normalization, distance calculation, pitch contour clustering, and pitch contour formulation.

Data Collection

Storytelling is a type of formal (scripted) expressive speech (Joaquim, 1992). In almost all studies of storytelling expressive speech synthesis, the speech corpora were acquired from spoken speeches of professional speakers and actors recorded in a studio (Aparicio et al., 2016). In this research, a total of nine storytellers were recruited for the recording, consisting of six women and three men aged 25 to 58 years old. An adequate variation of speaking style was necessary to obtain a good pool of storytelling speech styles. Moreover, this effort was to increase the digital resources of speech for the Malay language.

Malaysia has many classic collections of short folktales. In 2014, Parker (2014) collected 200 short folktales and created a book entitled “200 Kisah Teladan Haiwan”. Three most popular folklores from the book were chosen as the scripts for the recording. All the nine storytellers were able to record the storytelling varying in length from five to ten minutes per storyteller. Recordings were made in an isolated and quiet room in a laboratory equipped with a centralized air conditioner. As shown, Table 1 gives the detailed information about the syllables, words, and sentences for the three stories.

Table 1

Information about the Stories

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

In the recording session, a headphone with wired omni-directional microphone (Keenion KDM-E308) was utilized as an acquisition device. The frequency response for the microphone was 20–16000 Hz with sound sensitivity of the microphone at -48 BV. The stereo signal was captured at 16-bit resolution and 44.1 kHz. This setup was the standard recording setting to deliver high quality output with adequate size.

Background noise in the isolated room was analyzed at 18 dB coming from continuous buzzing of the air conditioning system. The speech file was saved based on stories and storytellers in wave files (.WAV). The duration of a storyteller to finish reading and recording all the stories was about less than one hour. Finally, the total of recorded files was 27 (9 storytellers x 3 stories) audios. The recorded files comprising 261 sentences (29 sentences x 9 storytellers), 2,619 words (291 words x 9 storytellers), and 6,147 syllables (683 syllables x 9 storytellers) were collected from the speakers.

Speech Pre-Processing

Disruptions such as background noise and buzzing noise from the microphone could not be prevented during recording. It was important to attenuate the existing artefacts in the speech file to minimize any future issues caused by them.

Prior to further progress and analysis, the need for pre-processing was to preserve and generate a high-quality speech signal. Therefore, spectral analysis, which included sampling, quantization, framing and windowing the audio, and filtering the noise, took place in the pre-processing stage. After that, the labeling and segmentation were carried out to prepare the speech signals for feature extraction.

Sampling and quantization. Speech signal has audio frequency elements between 20 Hz to 20 kHz of the electromagnetic spectrum (Hargus, 2005). The normal sampling frequency is 44.1 kHz in the stereo channel to capture the speech (Birkholz, 2013; Sebastian, 2014). It is because the frequency could capture up to 20 kHz of audio frequency as the highest frequency component to preserve the audio quality. Chowdhury (2006) also supported that 20 kHz sampling frequency was sufficient to digitize the signals to hold all the details of the voice. By considering the impacts of losing the naturalness of the speech, quality sampling frequency of 44.1 kHz was the best option. Quantization or bit resolution was the next critical parameter in the digitization process. Greater numbers of quantization levels could retain the speech originality. It was thus a trade-off between the number of bits and the representation of information. Therefore, the recommended bit resolution for speech processing was 16-bit (Sarkar et al., 2014).

Framing. The speech signal is a non-stationary time variable signal response to difference in frequency and spectral components over time. Ikkunointi (2016) mentioned that human speech is constructed from the phoneme dictionary, where most of the phoneme properties have remained invariant for a limited period of time (~5–100 ms). Consequently, the framing technique was utilized to convert the non-stationary speech signal into stationary signals (Hamzah, 2016). As known, framing techniques make the adjustment of the frame length. Since smooth spectral features need to be achieved, frame length must be in a high number. Paliwal et al. (2010) recommended to use a

frame length between 20 ms and 40 ms and a frame shift of 10 ms. In this study, a 20 ms frame length was selected.

Windowing. Windowing is a technique to reduce the discontinuity of the signal at the start and end of each frame. The irregularity called “leakage” is triggered at the beginning and end of the frame by abrupt shifts in the frame. The Hanning window provided a finer and more consistent signal (Podder et al., 2014), thus it was picked in this study. Due to the recording equipment, analogue to digital (A/D) conversion and ambient noise, the recording speech often included nonlinear disturbances that disrupted the quality of the speech. Noise filtering by using noise reduction technique was also carried out to eliminate interfering signals and to minimize noise in the environment that was analyzed at 18 dB.

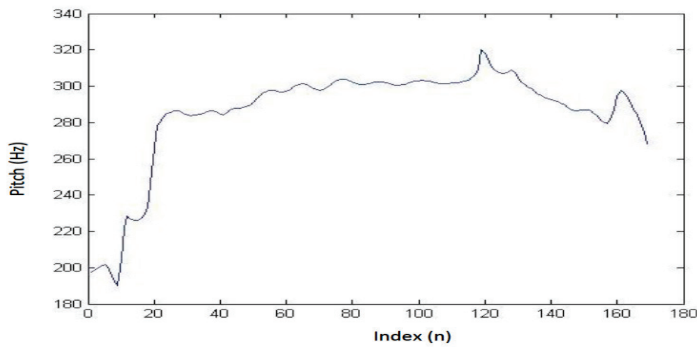
Speech labeling and segmentation. Annotation of the storytelling speech dataset was done manually by using a phonetic analysis tool called PRAAT (Boersma et al., 2015) as a final step in the pre-processing stage. The sentence, word, and syllable were annotated based on their levels once the speech file was imported to the PRAAT environment. The silence areas were not annotated and left as blanks. 27 transcriptions were created and stored in a text grid format (.textgrid) after annotations of all 27 audio files (9 storytellers x 3 stories) were completed. To produce syllable speech datasets for storytelling, all the syllable segments were extracted into a collection of syllables and the dataset was named STORY_DATA.

Pitch Contour Extraction

Before analyzing the pitch contours, the pitch features needed to be extracted in Hertz (Hz) for all syllables in STORY_DATA. Therefore, the sequence of pitch over time was extracted using STRAIGHT as illustrated in Figure 1.

Figure 1

Extracted Pitch Contour of One Syllable

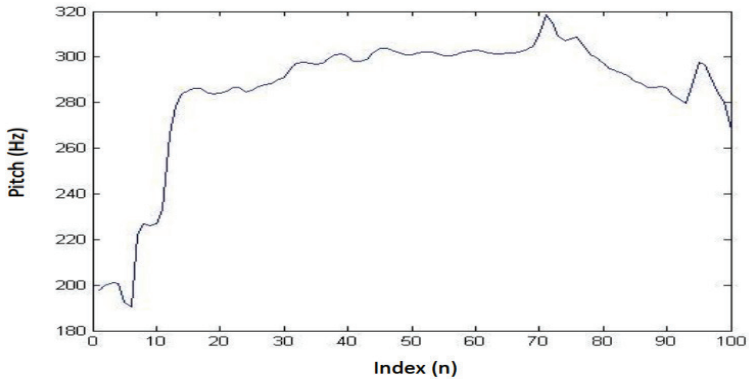


Pitch Contour Normalization

The storytelling speech dataset contained 6,147 syllables, thus 6,147 pitch contours were extracted and stored in a file. To analyze the shape of pitch contour in the dataset, each pitch contour needed to be normalized into 100 data points. 100 data points as a number of sampling was adequate to show an up-down pitch movement and differentiate the shape of the pitch contour. Therefore, a simple linear interpolation operation was used to sample the pitch contours as shown in Figure 2.

Figure 2

Normalized Pitch Contour of One Syllable



Distance Calculation

The method of normalization allowed the researchers to effectively quantify distances between pitch contours. These distances were used to cluster pitch contours of the same form that appeared identical; therefore, variations in pitch height or pitch range in the distance were not included in the calculation. One minus the Pearson product moment correlation (Klabbers & Santen, 2004) was employed to calculate the distance value (*d*) between two pitch contours as stated in Equation 1.

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

where,

- D* Distance value
- F_{0i}* First pitch contour
- \bar{F}_{0i} Mean of the first pitch contour
- F_{0j}* Second pitch contour

E_{0j}	Mean of the second pitch contour
N	Length of the pitch contour
sdF_{0i}	Standard deviation of first pitch contour
sdF_{0j}	Standard deviation of second pitch contour

Distance value (D) produced is from the range of 0.0 to 1.0. The meaning and description of each interval for distance values (Keith, 2005) are shown in Table 2.

Table 2

Description of the Range for Distance Value (D)

Distance value (D)	Descriptions
1.0 – 0.9	Very high correlation
0.9 – 0.7	High correlation
0.7 – 0.5	Moderate correlation
0.5 – 0.3	Low correlation
0.3 – 0.0	Little or no correlation

The distance value from 0.7 to 1.0 revealed that the pitch contour measured was significantly related with the targeted pitch contour. Therefore, the distance value of 0.7 was then used in this study as a threshold value to determine the resemblance of the pitch contour because the range of 0.7 to 1.0 was considered an acceptable threshold as stated in Keith (2005). A pitch contour was assigned as resembling another pitch contour if the distance value (D) between both pitch contours was more than 0.7.

A study by Klabbers and Santen (2004) identified six pitch contours that are commonly found in expressive speech. Therefore, this study adopted the six pitch contours to benchmark the pitch contours of all the syllables in the dataset. The pseudocode of assigning a pitch contour to a cluster using Algorithm 1.

Algorithm 1: Pitch Contour Clustering

```
For  $j = 1$  to  $N$ 
   $D = 0$ 
  For  $i = 1$  to  $K$ 
    Step 1: Calculate  $D$  (distance value) between pitch contour  $j$  with
    cluster  $i$ .
    Step 2: If  $D$  more than 0.7 and greater than  $D$  of pitch contour  $j$ .
      Step 2.1: Update  $D$  of pitch contour  $j$ .
      Step 2.2: Assign pitch contour  $j$  is in cluster  $i$ .
  end
end
 $D$  = Distance value
 $j$  = Pitch contour
 $I$  = Pitch contour cluster
 $N$  = Total number of pitch contours
 $K$  = Number of pitch contour cluster
```

The value of D for each pitch contour was initialized to 0. Every pitch contour j was compared with pitch contour cluster i , by calculating the distance value. If the value of D was more than 0.7 and greater than D of pitch contour j , the value of D was updated, and the pitch contour j was assigned in cluster i . The step is repeated until all the pitch contours were assigned to a cluster using the value D . Out of 6,147 pitch contour syllables, 80 percent of the pitch contours were clustered into their corresponding clusters. On the other hand, 20 percent were unassigned to any clusters. Upon closer inspection of the pitch contours, it was found that 10 percent of the unassigned pitch contours did not have the up-down movement (i.e., intonation) commonly present in an expressive speech. Therefore, these pitch contours could be disregarded as the synthesized speech could be reproduced directly. Other than that, it was also discovered that 10 percent unique pitch contours with high variations of vibrato existed in the pitch contours. These pitch contours may be unique to the Malay language. However, it is not within the scope of this paper to address these unique pitch contours. Table 3 shows the six clusters for 4,880 syllable pitch contours. Column 1 is the cluster number, while column 2 demonstrates the shape of the pitch contour. The x-axis is the normalized time at 100 data points and y-axis represents pitch

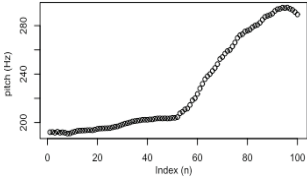
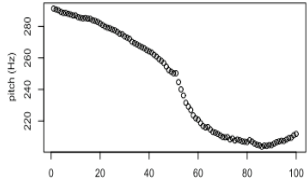
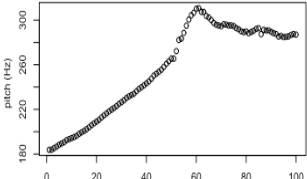
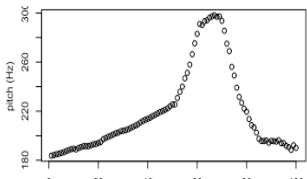
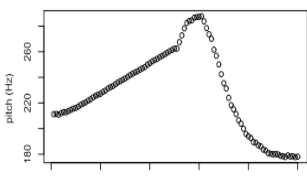
(Hz). The third column shows the total number of syllables assigned to the respective clusters, and the last column indicates the percentage of syllables in the cluster.

Cluster 1 contained a generally rising pitch with 9 percent occurrence in the total syllable of 6,147. This contour typically preceded the contours of Cluster 2 and 6. It was because the syllables in this cluster were located at the initial of a word with a rising pitch at the end and preceded with a contour that had a high pitch at the beginning of the pitch contour, such as Clusters 2 and 6. As an example, the syllable /di/ for the word [dia]. After the /di/ syllable, the pitch contour of syllable /a/ was either from Cluster 2 or 6. Cluster 2 comprised 8 percent of the contours from the total pitch contours. They depicted a downward trend from the top to the bottom. However, there was a slight increment in pitch at the end of the contour. In this paper, pitch contours of Cluster 3 allocated the largest amount with 27 percent of the pitch contours. It showed a slight incline to the center of the contour, followed by a plateau to the end. These contours indicated a steady rise and were mostly accompanied by the contours of Cluster 2. An example is the word [semut]. The syllable /se/ was a pitch contour from Cluster 3, followed by the syllable /mut/ from Cluster 2.

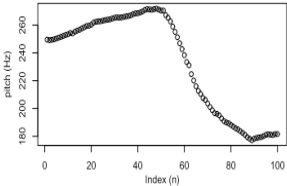
Cluster 4 comprised 5 percent of the overall syllable pitch contours, and it was the lowest occurrence of pitch contours. Clusters 4 and 5 were quite similar; nevertheless, the difference was that the location of the peak in Cluster 4 was late and higher as compared to Cluster 5. Cluster 5 was one of the larger clusters after Cluster 3, containing 19 percent of the total pitch contours. The shape of pitch contour in Cluster 5 was a default up-down movement that began with lower pitch, rising gradually until it reached the center of the contour and dropped very rapidly to the bottom. Cluster 6 contained 12 percent of the pitch contours. It operated at a higher pitch than Cluster 5 and the rising accent curve could be seen until the top of the plateau and reduced to the edge.

Table 3

Syllable Pitch Contours of STORY_DATA Grouped into Six Clusters

Cluster	Pitch Contour	Total Syllable	Percentage
Cluster 1		581	9%
Cluster 2		465	8%
Cluster 3		1,686	27%
Cluster 4		297	5%
Cluster 5		1,138	19%

(continued)

Cluster	Pitch Contour	Total Syllable	Percentage
Cluster 6		713	12%
	Others	1,267	20%
Total		6,147	100%

Pitch Contour Formulation

The aim of the pitch contour formulation is to synthesize a neutral speech pitch contour to a desired natural pitch contour. In Ramli et al. (2016), they formulated a two-step sinusoidal pitch contour formulation that modified a pitch contour to produce a natural synthesized pitch contour. However, as stated earlier, the formulation in Ramli et al. (2016) only worked for prominent syllables. This paper introduced an iteration of the two-step sinusoidal pitch contour formulation for both prominent and non-prominent syllables. The two-step formula modified a pitch contour using two steps by splitting the pitch contour of the syllable into two parts. The two steps were important because some of the targeted pitch contours had varying degrees of increasing and decreasing for both parts of a pitch contour. The advantage of the iterated two-step formulation was that it could manipulate a dynamic pitch contour as opposed to the formula by Verma et al. (2015). The previous formula of Ramli et al. (2016) is shown in Equation 2.

$$m'(t) = s(t) \times \left(1 + \alpha \times \sin \left(\frac{t - t_1}{t_2 - t_1} \right) \times \beta \times \pi \right) \quad (2)$$

where,

$m'(t)$ = Modified pitch contour

$s(t)$ = Neutral pitch contour

α = Desired maximum pitch shift

β = Constant for determining the contour shape

The previous formula in Equation 2 modified a syllable’s pitch contour by adjusting the value for α and β . The parameter α refers to the desired maximum pitch shift, while the parameter β determines the contour shape such as constantly increasing or falling. However, only a default pitch contour that was constantly rising or falling worked with Equation 2. This was because Equation 2 had difficulties in emulating pitch contour clusters with inconsistent contour shape at the first and second halves. As can be seen in the pitch contour in Cluster 5, the pitch contour started with a constant rise in the first half and then was consistently smooth in the second half of the pitch contour. However, the rigid variable value introduced from a previous work (Ramli et al., 2016) was only applied for pitch contour of prominent syllables. Therefore, an iteration of the two-step pitch contour formulation was proposed to find the suitable value of variables for modifying the pitch contour to further resemble the desired pitch contour. The main function of the iterated two-step sinusoidal pitch contour formulation was to modify the neutral pitch contour in two steps.

The pitch contour was formulated using Equation 3 in the first step. The parameters β and d define the form of the contour, which was continuously growing, rising, and then dropping; constantly falling or falling, and then rising. Meanwhile, both variables α and c are the desired maximum pitch shift. The second step was formulated using Equation 4. Finally, these two altered pitch contours were merged using Equation 5 to create a full pitch contour of a syllable. The value of the parameters α and β were used in the first step, while the parameters c and d were utilized in the second step.

$$\text{Step 1:} \quad p'(t) = s(t) \times \left(1 + \alpha \times \sin \left(\frac{t-t_1}{t_2-t_1} \right) \times 2\beta \times \pi \right) \quad (3)$$

$$\text{Step 2:} \quad f'(t) = p'(t) \times \left(1 + c \times \sin \left(\frac{t-t_1}{t_2-t_1} \right) \times c \times \pi \right) \quad (4)$$

$$\text{Overall:} \quad m'(t) = \begin{cases} p'(t) \rightarrow \text{if } (t < |t_2 / 2| \\ f'(t) \rightarrow \text{if } (t \geq |t_2 / 2| \end{cases} \quad (5)$$

where,

$m'(t)$ = Modified pitch contour

$s(t)$ = Neutral pitch contour

$p'(t)$ = Modified pitch contour at first half

$f'(t)$ = Modified pitch contour at second half

α = Desired maximum pitch shift for first half

β = Constant determining the contour shape for first half

c = Desired maximum pitch shift for second half

d = Constant determining the contour shape for second half

t = Current duration

t_1 = First duration

t_2 = Last duration

In the proposed method, the variables α , β , c , and d were iterated to find the most suitable variables for modifying the pitch contour toward achieving the desired pitch contour. The iteration began with the variable value α and c set to -0.5, and β and d were initialized to 0.1. The pseudocode of the iteration is as demonstrated in Algorithm 2.

Algorithm 2: The Pseudocode

$D = 0$

For $\alpha = -0.5$ to 0.5

 For $\beta = 0.1$ to 1.0

 For $c = -0.5$ to 0.5

 For $d = 0.1$ to 1.0

 Step 1: Modify neutral pitch contour using Equation (5) with the value α , β , c , and d .

 Step 2: The modified pitch contour are compared to original storytelling pitch contour using one minus the Pearson product moment correlation.

 Step 3: If D is greater than 0.7 and greater than D (current value).

 Step 3.1: Update D for modified pitch contour.

 Step 3.2: Update that modified pitch contour is similar to original storytelling pitch contour.

 end

 end

 end

end

The iteration continued until the last value of α was 0.5. The increment for each iteration was 0.1. Every modified pitch contour from the iteration was compared with the original storytelling pitch contour by calculating the distance value (D). The iteration could also be stopped when the distance value (D) of the modified and targeted pitch contours achieved a high distance value (i.e., greater than 0.9) to indicate a high resemblance of the desired pitch contour.

RESULTS AND DISCUSSION

The new proposed iterated two-step pitch contour formulation and Equation 2 were tested to evaluate the performance of converting a neutral pitch contour into a natural storytelling pitch contour. A speech test dataset containing 1,366 natural pitch contours and 1,366 neutral pitch contours was acquired from the same one man and woman storytellers. The man and woman storytellers were chosen from the storyteller pool as defined in Section 2. A neutral speech is described as a monotonous speech and has a smooth intonation (Lutfi, 2007). Therefore, the recording and speech pre-processing of the neutral speeches were also done following the same procedures as stated in Section 2. The neutral pitch contours from the test dataset were modified using the proposed iterated two-step pitch contour formulation to produce the pitch contours altered by the proposed Equation 5. On the other hand, the same neutral pitch contours were modified using Equation 2. Then, the modified pitch contour from both equations and the corresponding natural storytelling pitch contours were compared. The difference between the two pitch contours was determined using Equation 1 that consisted one minus the Pearson product moment correlation. As mentioned, this equation only measured the difference in pitch height or range. A distance value (D) with a value of 1 implied a greater resemblance. The performance for both equations is shown in Table 6.

Table 6

Result of using Equation 2 and Proposed Equation 5

Distance value (<i>D</i>)	Equation 2	Proposed Equation 5
1.0 – 0.9	1%	13%
0.9 – 0.7	14%	80%
0.7 – 0.5	85%	7%
0.5 – 0.3	0%	0%
0.3 – 0.0	0%	0%
Total	100%	100%

Based on Table 6, 93 percent of the modified neutral pitch contours using the proposed iterated two-step pitch contour formulation achieved the distance value (*D*) of above 0.7, indicating a high correlation matching of the synthesized neutral pitch contour with the natural pitch contour. 13 percent of the neutral pitch contours had a very high correlation distance value (*D*), while 80 percent scored a high correlation value of *D*. However, a small percentage of 7 percent in the proposed Equation 5 was classified as moderate correlation. As compared to modification by using Equation 2, only 15 percent achieved a high correlation but most of the results for distance value (*D*) were between 0.7 and 0.5 at 85 percent. None of the modified pitch contours had low correlation values of *D*, which was below 0.5 for both equations. Based on the results, it proved that the proposed iterated two-step pitch contour formulation was able to convert the pitch contour of the neutral speech to resemble an expressive speech. Furthermore, it is believed that 7 percent of the moderate correlation for the distance value (*D*) came from the unique syllable pitch contours that could not be clustered into the six clusters of the expressive speech. These pitch contours, which comprised 10 percent of the total syllables that have been presented in the pitch contour clustering section, needs to be further studied.

CONCLUSION AND FUTURE WORK

This research demonstrated the success of an iterated two-step pitch contour formulation to modify the pitch contours of neutral speeches to resemble their corresponding natural speeches. Even though only storytelling speaking style was conducted in this paper for the analysis until the synthesis process, the method of modifying the pitch contour can be used for other speaking styles, such as commentary, news reading, conversation, poetry reading, and debate. Instead of the tempo and intensity of the speech, the shape of pitch contour could also lead the style of the speeches. Furthermore, this paper showed that 80 percent of the pitch contour of expressive speeches in the Malay language can be generalized into the standard six pitch contour clusters. However, the Malay language has some unique expressive speech pitch contours that need further investigation, especially toward contributing to the development of Austronesian language. In the future, the researchers plan to establish a pitch prediction model to predict the pitch contour of the synthesized storytelling speech for the text-to-speech storytelling application and toward the completion of the Malay language intonation model of expressive speech, which can synthesize storytelling speech in the Malay language.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia for the facilities and financial support.

REFERENCES

- Aparicio, R. M., Salle, L., Ramon, U., Tècnica, E., & Electrònica, E. (2016). *Prosodic and voice quality cross-language analysis of storytelling expressive categories oriented to text-to-speech synthesis*. (Doctoral dissertation, Universitat Ramon Llull).
- Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4), 1–17. <https://doi.org/10.1371/journal.pone.0060603>

- Boersma, P., David, W., & Heuven, V. (2015). *Praat doing phonetics by computer* (v. 5.3.39). <http://www.praat.org/>
- Chang, R. R., Yu, X. Q., Yuan, Y. Y., & Wan, W. G. (2014). Emotional analysis and synthesis of human voice based on STRAIGHT. *Applied Mechanics and Materials*, 536–537, 105–110. <https://doi.org/10.4028/www.scientific.net/AMM.536-537.105>
- Chowdhury, S. (2006). *Concatenative text-to-speech synthesis: A study on standard colloquial Bengali*. (Doctoral dissertation, Indian Statistical Institute).
- Gelin, R., D'Alessandro, C., & Le, Q. (2010, November). Towards a storytelling humanoid robot. In *AAAI Fall Symposium Series on Dialog with Robots* (pp. 137–138).
- Gu, H., & Jiang, K. (2015, July). A pitch-contour generation method combining ANN, global variance, and real-contour selection. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 396–402). <https://doi.org/10.1109/ICMLC.2015.7340954>
- Hamzah, R. (2016). *Discriminative classification model of filled pause and elongation for Malay language spontaneous speech*. (Doctoral dissertation, Universiti Teknologi MARA).
- Hamzah, R., & Jamil, N. (2019). Investigation of speech disfluencies classification on different threshold selection techniques using energy feature. *Malaysian Journal of Computing*, 4(1), 178–192.
- Hargus, S. (2005). Athabaskan phonetics and phonology. *Language and Linguistics Compass*, 4(10), 1019–1040.
- Ikkunointi. (2016). *Windowing*. http://www.cs.tut.fi/kurssit/SGN-4010/ikkunointi_en.pdf
- Jamil, N., Apandi, F., & Hamzah, R. (2017, July). Influences of age in emotion recognition of spontaneous speech: A case of an under-resourced language. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1–6). IEEE.
- Joaquim, L. (1992, July). Speaking styles in speech research. In *Workshop on Integrating Speech and Natural Language* (pp. 15–17).
- Kato, S., Yasuda, Y., Wang, X., Cooper, E., Takaki, S., & Yamagishi, J. (2020). Modeling of Rakugo speech and its limitations: Toward speech synthesis that entertains audiences. *IEEE Access*, 8, 138149–138161.

- Keith, G. (2005). *Correlation coefficients*. <https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm>
- Klabbers, E., & Santen, J. (2004, June). Clustering of foot-based pitch contours in expressive speech. In *Fifth ISCA Workshop on Speech Synthesis* (pp. 73–78).
- Lunce, B. C. (2007). Digital storytelling as an educational tool. *Indiana Libraries*, 30(1), 77–80.
- Lutfi, S. L. (2007). *Adding emotions to synthesized Malay speech using diphone-based templates*. (Master's thesis, University of Malaya).
- Md Saad, M., Jamil, N., & Hamzah, R. (2018). Evaluation of support vector machine and decision tree for emotion recognition of Malay folklores. *Bulletin of Electrical Engineering and Informatics*, 7(3), 479–486. <https://doi.org/10.11591/eei.v7i3.1279>
- Montaño, R., Alías, F., & Ferrer, J. (2013, September). Prosodic analysis of storytelling discourse modes and narrative situations oriented to text-to-speech synthesis. In *8th ISCA Speech Synthesis Workshop* (pp. 171–176).
- Paliwal, K. K., James, L., & Kamil, W. (2010, December). Preference for 20-40 ms window duration in speech analysis.. In *2010 4th International Conference on Signal Processing and Communication Systems (ICSPCS)* (pp. 1–4). IEEE.
- Parker, V. (2014). *200 kisah teladan haiwan*. Edukid Publication.
- Plaisant, C., Druin, A., Lathan, C., Dakhane, K., Edwards, K., Vice, J. M., & Montemayor, J. (2000, November). A storytelling robot for pediatric rehabilitation. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies - Assets '00* (pp. 50–55). <https://doi.org/10.1145/354324.354338>
- Podder, P., Khan, Zaman, T., & Haque Khan, M. (2014). Comparative performance analysis of hamming, hanning and blackman window. *International Journal of Computer Applications*, 96(18), 1–7.
- Ramli, I., Jamil, N., Seman, N., & Ardi, N. (2016, November). An improved pitch contour formulation for Malay language storytelling text-to-speech (TTS). In *IEEE Industrial Electronics and Applications Conference (IEACon)* (pp. 250–255). IEEE.
- Raúl, M., & Francesc, A. (2017). The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages. *Speech Communication*, 88, 1–16.

- Roekhaut, S., Goldman, J., & Simon, A. C. (2010, May). A model for varying speaking style in TTS systems. In *Fifth International Conference on Speech Prosody* (pp. 4–7).
- Sarkar, P., Haque, A., Dutta, A. K., Gurunath Reddy, M., Harikrishna, D. M., Dhara, P., Verma, R., Narendra, N. P., Sunil Kr, S. B., Yadav, J., & Rao, K. S. (2014, August). Designing prosody rule-set for converting neutral TTS speech to storytelling style speech for Indian languages: Bengali, Hindi and Telugu. In *2014 7th International Conference on Contemporary Computing (IC3)* (pp. 473–477). <https://doi.org/10.1109/IC3.2014.6897219>
- Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. *Affective Information Processing*, 2, 111–126.
- Sebastian, D. (2014). *Unit selection text to speech system for Polish*. [Unpublished thesis] Faculty of Mechanical Engineering and Robotics, AGH University of Science and Technology.
- Theune, M., Meijs, K., Heylen, D., & Ordelman, R. (2006). Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 1099–1108.
- Um, S.-Y., Oh, S., Byun, K., Jang, I., Ahn, C., & Kang H.-G. (2020, May). Emotional speech synthesis with rich and granularized control. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7254–7258). IEEE.
- Verma, R., Sarkar, P., & Rao, K. S. (2015, January). Conversion of neutral speech to storytelling style speech. In *8th International Conference on Advances in Pattern Recognition* (pp. 1–6). <https://doi.org/10.1109/ICAPR.2015.7050705>