Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

THE FIRST MALAY LANGUAGE STORYTELLING TEXT-TO-SPEECH (TTS) CORPUS FOR HUMANOID ROBOT STORYTELLERS

I. Ramli[1], N. Jamil[1,*], N. Seman[1] and N. Ardi[2]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[2]Academy of Language Studies, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Published online: 05 October 2017

ABSTRACT

This paper describes the process undertaken and criteria considered in acquiring a storytelling speech corpus of Malay language towards the development of humanoid storyteller. The speech corpus contains 464 speech sentences, 4,656 words and 9,584 syllables. Three children's short stories were recorded by 3 female storytellers, 1 male professional speaker, 2 female speakers and 2 male speakers. The equipment specifications, recording procedures and speech annotations are described in detail in accordance to baseline work. The stories were recorded in two speaking styles that are neutral and storytelling speaking style. The first Malay language storytelling corpus is not only necessary for the development of a storytelling text-to-speech (TTS) synthesis. It is also detrimental for natural language processing and speech recognition of Malay language, an under-resourced language.

Keywords: storytelling speech corpus; humanoid storyteller; storytelling TTS; Malay language.

Author Correspondence, e-mail: liza@tmsk.uitm.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i4s.20

1. INTRODUCTION

A humanoid robot capable of storytelling has increasingly being used to interact with children and assist them in teaching and learning of languages. The humanoid played and interacted with the child acting as a social character, while telling them stories and introducing new vocabulary words [1]. The famous Honda's Asimo robot was also used in a storytelling experiment conducted in Carnegie Mellon University to test the effectiveness of human-like gaze during storytelling [2]. In a more recent work, expressive text-to-speech synthesis was utilized in a humanoid robot Nao to automatically narrate a short story to children [3] and Nao robot platform is constructed in [4] to allow direct transmission of speech and gestures produced by a human operator. Storytelling using Nao robot has also shown to be capable in improving language learning to improve children's oral language, learning new vocabulary words and increased the amount of diversity of the language [5]. An automated storytelling humanoid robot requires a storytelling text-to-speech synthesis to be embedded into the robot system. Therefore, linguistic module containing the language's resources such as speaking style speech corpus need to be constructed. In many languages such as English, Japanese and Mandarin, storytelling speaking style speech corpus is already established. Unfortunately, Malay language is still categorized as an under-resourced language [6] due to its limited presence on the web and lack of electronic resources for speech and language processing such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data and pronunciation dictionaries [7]. Although some initial efforts towards developing the language resources, there is little or no available storytelling speaking style resources for Malay language. Current text-to-speech synthesis (TTS) system in Malay language is only able to produce neutral speech sound such as news reading speech [8]. In order to modify standard text-to speech synthesis to storytelling speech synthesis, speaking style analysis should be done on storytelling speech. Similar to spontaneous speech, the study of storytelling speaking style requires large amount of data to understand the linguistic nature of the speaking style [9]. Other than speaking style, the speech quality is also an important factor to be considered [10]. Hence, to ascertain high quality recording speech, recording equipment, recording setup, storyteller chosen should be well thought out.

In this research, we will describe the process conducted in collecting the storytelling speech corpus for the Malay language. This paper is structured as follows. Section II describes a review of storytelling speech corpus of other languages. In section III, demography of the storytellers is presented followed by elaboration of the selected stories in section IV. Section V and VI presented the equipments needed and the recording conditions. Speech pre-processing is described in section VII and section VIII discussed the annotation of the recorded storytelling speeches. The succinct storytelling analysis is elaborated in section IX. Conclusion of this research is described in section X.

2. RELATED WORK

An important component of a TTS is the speech database [11], stored in syllable level, word level or sentence level format. Similarly, a storytelling speaking style TTS also requires a storytelling speech database collected from professional actors or experienced storytellers. The storytelling speech database is crucial for prosody analysis of human storytellers to generate rules and models of the storytelling speaking style.

We began our study by reviewing related work on storytelling speech corpus in different languages such as Bengali [10], Hindi [12], Dutch [13], English [3, 14], French [15], Slovak [16] and Spanish [17]. Summary of the review is presented in Table 1 and briefly described accordingly. Storytelling is closely associated to children and in all work listed in the table used children stories and fairy tales in their corpus. However, depending on the language resources, the quantity of recorded stories varies from one story to 25 stories. The person chosen as storyteller comprised professional speaker, artist or storyteller; the number of storytellers ranged from 1 to 3 persons for each speech corpus. Table 1 is further discussed in next sections as the baseline of the selection of stories, storytellers and the recording specifications of our first Malay Language storytelling speech corpus.

3. STORYTELLER SELECTION

The storyteller is the person who reads the story's scripts in a storytelling manner during the recording session, be it in narrative, descriptive or dialogue discourse mode. In [3] hired a professional speaker to read 12 tales for audio recording and the storytelling speech is later annotated into phonemes, syllables, pitch, rhythm and voice quality. Since, the storytelling

TTS is to be incorporated into a humanoid robot, they further videotaped 6 actors to act the emotional gestures for a more realistic storytelling. However, gesture is not within the scope of our current work. Other researchers that also employed professional speakers as their storytellers are [15], [14]. On the contrary, [10, 12-13] engaged professional artists and actors for their collections of storytelling audio speeches.

Table 1. Summary of related work

References	Language	Storyteller	Corpus Size
[14]	English	1 semi-professional female speaker	2 children stories, 128 sentences
[3]	English	1 professional speaker	89 short stories, 12 selected tales
[10]	Bengali, Telugu	1 professional radio artist	125 children stories, total 3 hours
[13]	Dutch	3 male Dutch actors	5 fairy tales
[12]	Hindi	1 male, 1 female professional artist	25 children stories
[16]	Slovak	1 male speaker	10 children stories
[17]	Spanish	1 Spanish storyteller	1 story

In [13], three Dutch actors recorded 5 stories. While, in [12] hired one female and one male artist as their storytellers of 25 stories. In [10] used one professional radio artist to record 125 Bengali and Telugu stories. Five different male and female speakers are further engaged by [10] to record the same stories in a neutral style like news reading speech. Only one work cited using a storyteller [17] to record a story for their speech corpus and another work cited using a male speaker [16].

In our work, three female storytellers, one male professional speaker, two female and two male speakers are hired for the recording of children folktales. The female storytellers are kindergarten school 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 college degree students who are eloquent speakers and have 3 to 5 years experiences giving public speeches. Our selected storytellers are all native Malay

speakers and speak Malay language as their first language. None have any speech-related problems. A total of 8 storytellers are hired comprising 5 females and 3 males with age ranging from 25 to 58 years old. The wide range of experiences and age allow us to later analyze the prosody fluctuations and changes to determine whether age, experience and even gender influence our proposed storytelling TTS rules and models.

4. COLLECTION OF STORIES

Storytelling which is a subtheme of fiction literature is based on discourse modes, typically containing narrative, descriptive and dialogue [11] modes. Narrative storytelling is mainly used to inform the listener about the actions that are taking place and the characters affecting the story. Meanwhile, descriptive mode described a character or event in precise details to the listener so that they can get a clear picture as depicted. Lastly, dialogue storytelling is when the storyteller typically modifies his/her voice into a character producing exaggerated register of expressions and full-blown emotions may be manifested. In [17] introduced narrative situations which are analyzed at sentence level in his storytelling speaking style analysis of Spanish translation of “Harry Potter and the Philosopher’s Stone”. In most storytelling speaking style, children stories and folk tales are the favourite types of stories to be narrated. In [12], the children stories are taken from Internet collections and story books such as Panchatantra and Akbar Birbal.

In [16] verified their conversion method through the analysis of 65 sentences of the Slovak story “Witch’s Garden” containing three expressive–storytelling voices (“Teller”, “Prince”, and “Witch”). Meanwhile, in [14] stated the recordings used in their study contain two childrens stories by Beatrix Potter.

Our work selected three narrative children short stories from a classic Malaysia’s collections of short stories entitled titled 200 *Kisah Teladan Haiwan* [11]. These stories are readable within five [12, 15] to ten minutes [14] of speech. The number of sentences, words and syllables are depicted in Table 2.

Table 2. Total sentences, words and syllable

Story	No. of Sentences	No. of Words	No. of Syllables
Story 1	12	113	276
Story 2	9	80	175
Story 3	8	98	148

Total	29	291	599

The selected 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 fulfils the formal Malay language with simple words easily understood by the children. For ease of recording session, the story's script is displayed using Microsoft Power Point slides in separate slides for each story. Each sentence is placed in separate lines and is numbered as shown in Fig. 1. This format reduced the possibility of reading mistakes by the storyteller such as sentence repetition.

5. RECORDING EQUIPMENT

In this section, the recording equipment used for the recording session of the storytelling speech corpus is presented. Two laptops, one head-mounted microphone and a digital camera are utilized for the storytelling recording sessions. The audio acquisition device is a Keenion KDM-E308 head-mounted microphone as shown in Fig. 2(a). The frequency response is 18-20000 Hz for the headphone and 20-16000 Hz for the microphone, respectively. The headphone is Omni-directional headphone with sound sensitivity of -48dBV.

Si Angsa yang Bertelur Emas (Story 1)

1. *Suatu masa dahulu, tinggal seorang petani yang memelihara seekor angsa.*
2. *Pada suatu hari, ketika itu dia ingin mengambil telur angsanya.*
3. *Si petani mendapati telur itu kelihatan aneh.*
4. *Warnanya kuning keemasan dan berat!*
5. *Dia menyangka jirannya cuba bergurau lalu bercadang untuk membuang telur itu.*
6. *Namun selepas berfikir, dia membawa telur itu pulang ke rumah untuk diperiksa.*
7. *Si petani berasa terkejut apabila mendapati itu adalah telur emas!*
8. *Si petani sungguh gembira.*
9. *Hari demi hari selepas itu, si angsa terus bertelur emas.*
10. *Si petani mula menjadi tamak.*
11. *Si petani mengambil pisau dan menyembelih angsa bertuahnya.*
12. *Apabila mendapati tiada sebiji pun telur emas di dalam perut angsa itu, si petani mula menyedari kesilapannya dan berasa sangat menyesal.*

Fig.1. Example story script displayed to storyteller

Fig.2. Recording devices

We also captured a video recording of the storytelling session using a Canon E0S 700D Digital SLR camera (see Fig. 2(b)). The aim of the video recording session is for visual observations of the storytelling session for housekeeping purposes. The camera can capture 18-megapixel photos and full-HD (High-Definition) video acquisition. It contains an external 8 GB memory card which is adequate for a full recording session of one short story. One laptop is used to display the story's script while the head-mounted microphone is attached to another laptop for the recording purposes. Table 3 listed the specifications of both laptops. A free, open source digital audio editor and recording software, Audacity is used to record the audio speeches.

Table 3. Specification of the laptops used for recording

Specifications		
Brand	HP Pavilion m4-1002tx	Dell Inspiron 14 (1464)
Window	Windows 8	Window 7
RAM	8 GB	4 GB
System Type	64 bit operating system	64 bit operating system
Microprocessor	2.2 GHz Intel Core i7-3632QM	2.1GHz Intel Core i3-330M
Display	14" diagonal HD BrightView LED-backlit (1366 x 768)	14" diagonal HD LED display (1366 x 768)

6. RECORDING OF THE 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 air conditioner with one door entrance. 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 air conditioner with one door entrance. Recording equipments are arranged as illustrated in the recording setup in Fig. 3.

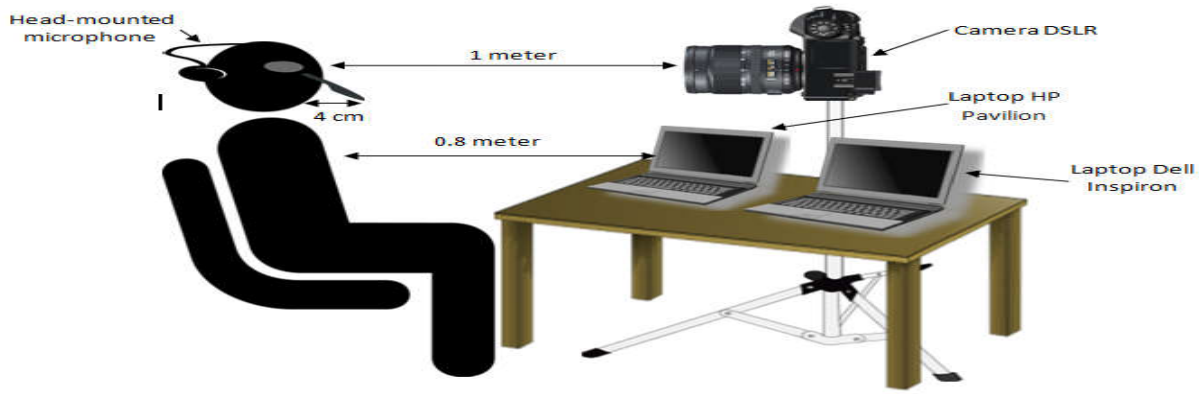

Fig.3. The recording setup

The HP Pavillion laptop is placed in front of the storyteller displaying the story's scripts. On its right side, the Dell Inspiron laptop is used to run Audacity software and recorded the storytelling thorough the connected mounted headphone. The microphone is placed on the speaker's head, approximately 4 cm from the speaker's mouth. It recorded a stereo signal at 44.1 kHz and 16-bit resolution. The recorded audio is stored in wave file (.WAV) with background noise level measured as 18 dB.

The DSLR Camera is placed behind the desk about one meter away from the storyteller. The built-in camera's microphone recorded storytelling audio at stereo signal of 48 kHz and 16-bit resolution, while visual recording is captured at 25 frames per second with a resolution of 1920x1088. The video recordings are stored in video format .MOV file. The main purpose of video recording using camera DSLR is to provide an alternative recording session.

Fig.4. Recording session for eight storytellers

Each storyteller was given ample time to practice and get familiarize with the story's content before entering the recording room for recording session. This was done in order to make the recordings as natural and fluent as possible. The recording session may began at any time when the storyteller was comfortably seated on the chair and wearing a head-mounted microphone as shown in Fig. 4. During the recording session, the storytellers may repeat the

recording until they were fully satisfied. No specific instructions on how to read the stories were given and the storytellers were free to move their body or arms to act out the narrations. When disfluencies occurred, the stories were re-recorded until it was fluent.

The storytellers need to record the three stories in two versions that are neutral speech and storytelling speaking style. In [12-13], the storytellers were requested to deliver the same story in neutral speech style like news reader speech and storytelling speaking style. The objective is to compare neutral speech with storytelling speech and analyze the differences between these two speaking styles [18]. In our work, the recording started with recording of the neutral speech where storytellers were asked to record a story with minimal intonation without emphasis. They need to maintain the constancy of pitch and intensity at the time of recording. To achieve all these as well as to obtain good voice quality, the storyteller must maintain their vocal qualities in term of intelligibility, timbre, diction and pronunciation [26]. Once completed, they were given time to rest and when they were ready; the recording session proceeded with the same story in storytelling speaking style. The storyteller needs to narrate the story script with their storytelling style without influence by another storyteller. Two different recorded files (i.e. neutral and storytelling speech) are collected from one story. The same processes continued for the other two stories. A complete recording session for one storyteller took approximately one hour.

7. SPEECH PRE-PROCESSING

A total of 48 (8 storyteller x 2 speaking styles x 3 stories) audio .WAV files and 48 video .MOV files were produced at the end of the recording sessions. The video files are compressed to normal quality with video size 1080 × 720 to reduce the size and saved as .AVI files. The audio file was pre-processed to maintain and produce good quality of speech signal before further processing and analysis. Pre-processing started with spectral analysis such as spectrogram analysis, sampling, framing, windowing and filtering.

7.1. Spectrogram Analysis

Spectrogram is used since speech signal information is best expressed in both time- and frequency-based domains.

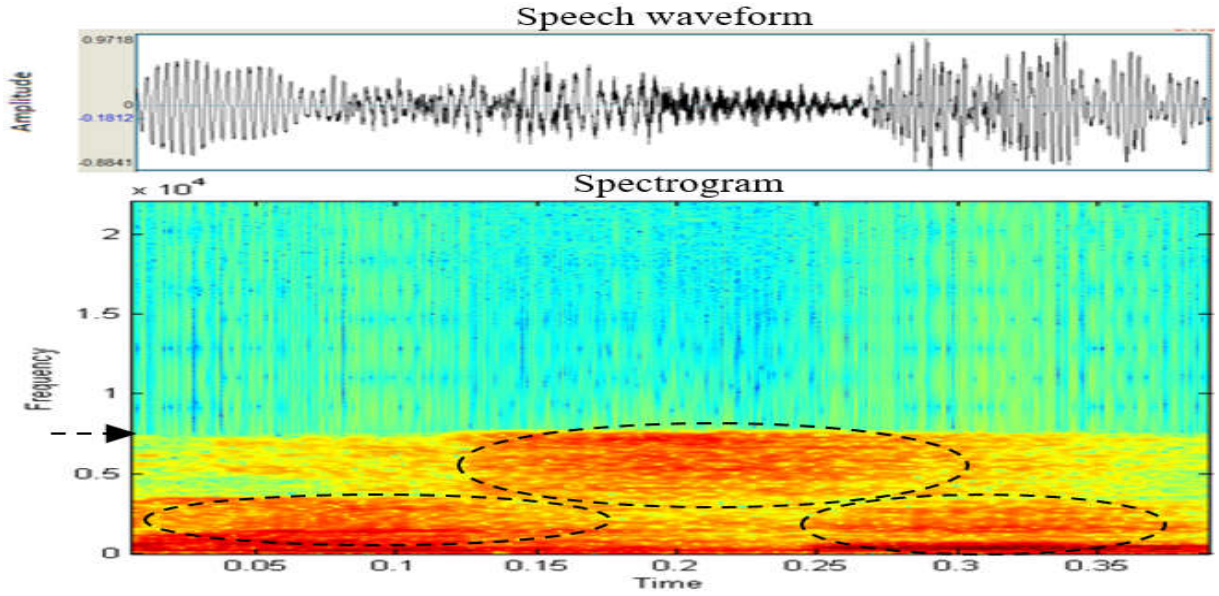

Fig.5. Spectrogram analysis result from male storyteller

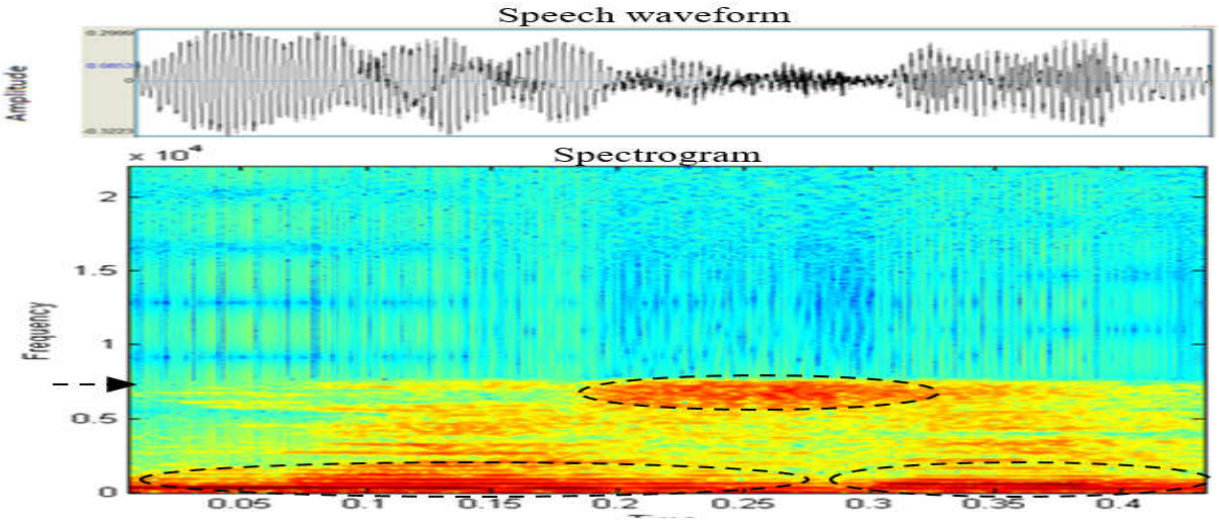

Fig.6. Spectrogram analysis result from female storyteller

In Fig. 5 and Fig. 6, a speech waveform of a male and female for the word "masa" recorded at sampling rate of 44.1 kHz and its corresponding spectrogram are demonstrated. Based on the figures, a few important observations can be highlighted as follows.

- It shows the concentrated signal energy of the voiced speech (as shown in circle) and proves that male professional speaker has larger energy area (i.e. speak louder) than female storyteller.
- The frequency information of recorded speech for male professional and female storyteller mainly ranging from 0 up to 8000 Hz as shown by the arrow.

7.2. Speech Sampling

The speech signal has frequency components in the audio frequency range (20Hz to 20 kHz) of the electromagnetic spectrum. The standard sampling frequency to speech recorded is

chosen to be 44.1 kHz in the stereo channel. This is because 20 kHz is the maximum frequency component and allowing some guard band. The 20 kHz sampling frequency is the good enough to digitize the signals to keep all the speech information within the range [19]. Lesser sampling rate will be a loss of naturalness of sound quality. However, the human ear is most sensitive to frequency components between 500Hz to 4000 Hz [20].

Thus, in order to capture significant speech information of the recorded speech, sampling frequency of our speech data are analyzed. An example of a recorded speech signal for the word "masa" from the male professional speaker with a different sampling frequency (44.1 kHz, 32 kHz, 16 kHz and 8 kHz) is shown in Fig. 7.

The observation showed that there are no significant frequency components in the spectrum beyond 8 kHz (in the dashed circle) for sampling frequency of 44.1 kHz. It also demonstrates that 44.1 kHz sampling rate is too high to capture information present in the speech signal. At 32 kHz sampling rate, some frequency component is not filtered (in the dashed circle), thus deemed as unsuitable. Since speech information is up to about 8 kHz, 16 kHz sampling rate is the optimal sampling for our speech data.

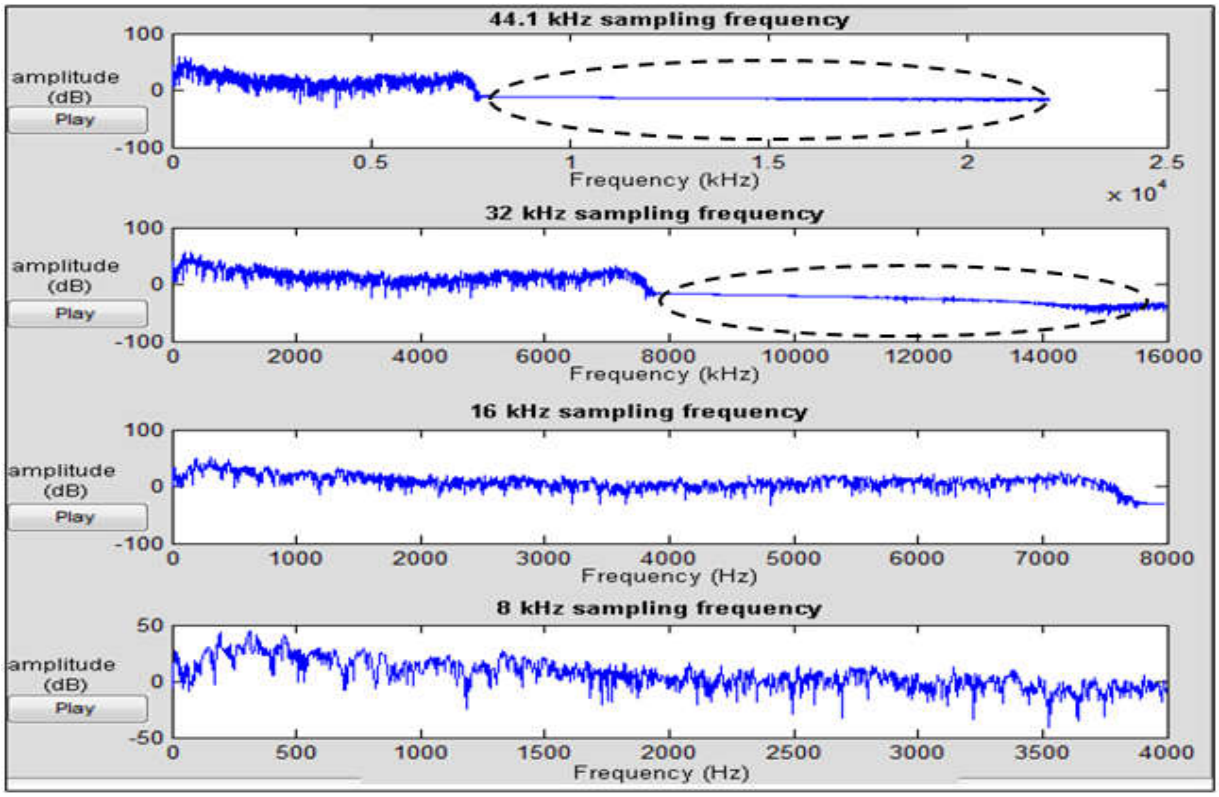

Fig.7. Recorded speech in different sampling frequency

7.3. Framing

The speech signal is a non-stationary time variant signal because of the changes in frequency and spectral component over time. The human speech signal is built from the dictionary of

phonemes, and most of the phonemes properties remain invariant for a short period of time (~5-100 ms) [21]. Thus, the non-stationary speech signal need to be transformed as stationary using framing method [22]. Framing is the process of blocking the speech signal into frames of N samples. The adjacent frames are separated and shifted for M samples to overlap with the previous frames. The shift of the M samples determines the smoothness of the spectral features. If the M samples are small, the spectral features will be smooth. Without overlapping between the adjacent frame, the correlation between frames and adjacent frames will contain noisy component [20]. The illustration of the framing process is shown in Fig. 8.

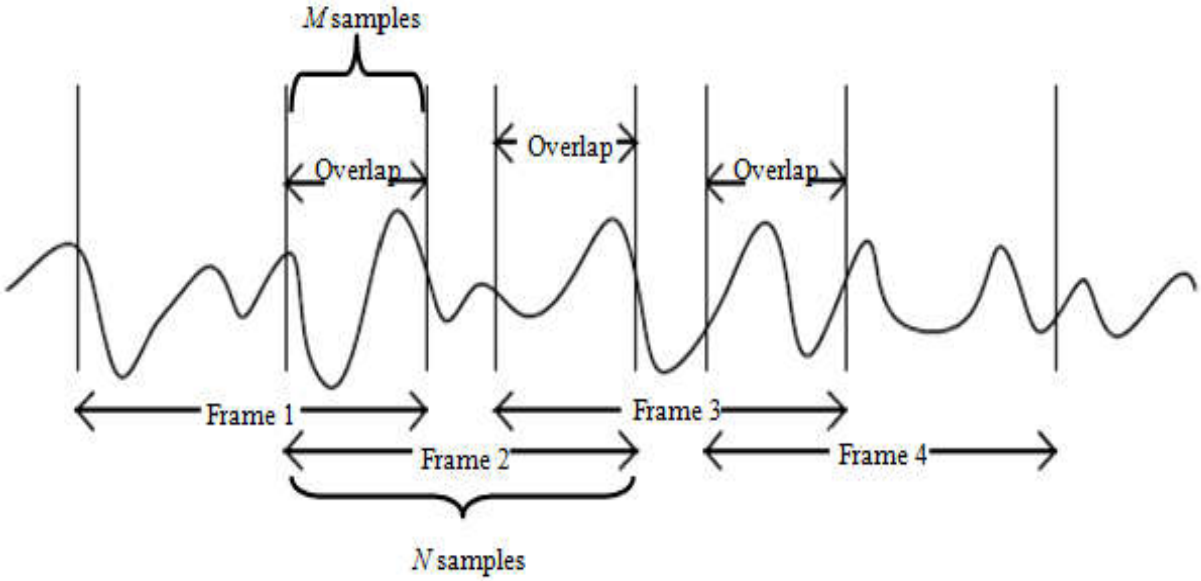

Fig.8. Illustration of framing process

The number of N sample for each frame can range from 160 to 640 samples for the duration of 10ms to 40ms at 16 kHz sampling rate. The general equation for frame blocking is shown in equation (1). The symbol S represents the length l_{th} frame of the speech signal and L represents the entire speech signal.

$$x_l(\mathrm{N}) = \overline{\mathrm{S}}\,(M_l + \mathrm{N}) \tag{1}$$

where x_l = Frame of speech, N = 0,1....., N-1 sample and l = 0,1....., L-1 frames.

The speech data is tested over various frame lengths (10ms, 20ms, 30ms and 40ms) with 10ms (160 samples) frame shift. An example of different framing for speech utterance “suatu masa dahulu” are shown in Fig. 9.

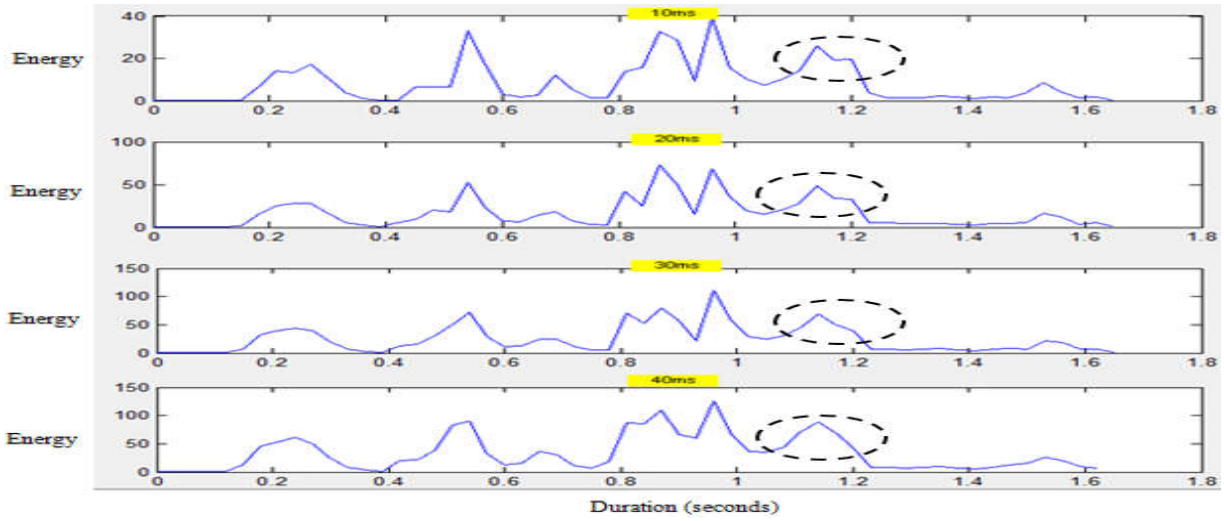

Fig.9. Short time energy over various frame length

The effects of the frame length choice are demonstrated. A long frame would result in very little changes of the measurement in time, whereas a short frame resulted with a non-smooth speech signal (as shown in dashed circle). As the frame length increases, short-time energy becomes smoother, as expected. Since 10 ms frame length produce too many details, the preferred length is between 20 ms to 40 ms (20 ms is chosen for our speech data) for frame length and 10 ms for frame shift [23].

7.4. Windowing

Windowing is to minimize the signal discontinuities at the beginning and ending of each frame. The "leakage" aberrations from sudden changes in the frame at the start and end frame. Three types of windowing that are Hamming, Rectangular and Hanning windows were tested using 20 ms window length as shown in Fig. 10.

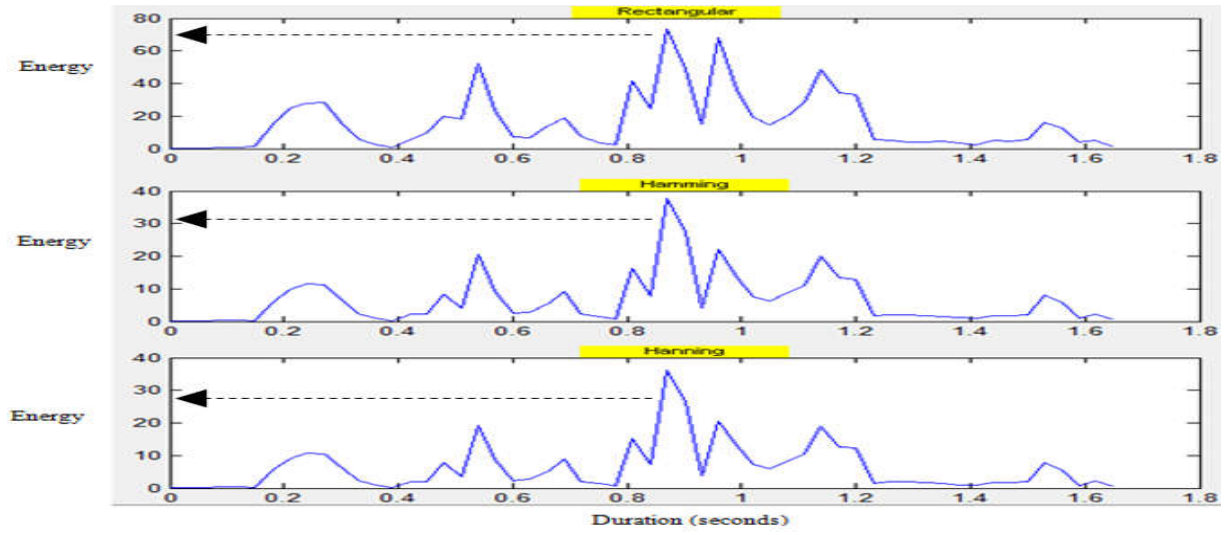

Fig.10. Speech signal with various windowing techniques

It shows that the Rectangular windowing produced larger energy (as shown by the arrow), compared to Hamming and Hanning window. Even though all windows are suitable to be used, Hanning window was chosen and applied to our speech signal because it more smooth and accurate [24].

7.5. Filtering

A recorded speech contained non-linear distortions due to recording devices, A/D conversion, and environment noise disturbing the quality of the speech. In order to suppress interfering signals and reduce environment noise, filtering is done. The goal of filtering is to remove unwanted components or features from a speech signal. Environment noise was analyzed at 18 dB and removed using noise reduction technique. The noise reduction technique can reduce constant background sounds such as hum, whistle, whine, buzz and "hiss" such as tape hiss, fan noise or FM/webcast carrier noise. Noise reduction applied high pass filter which sets the noise floor in each of the frequency bands and used this as the threshold. The speech waveform before and after filtering is shown in Fig. 11 and Fig. 12.

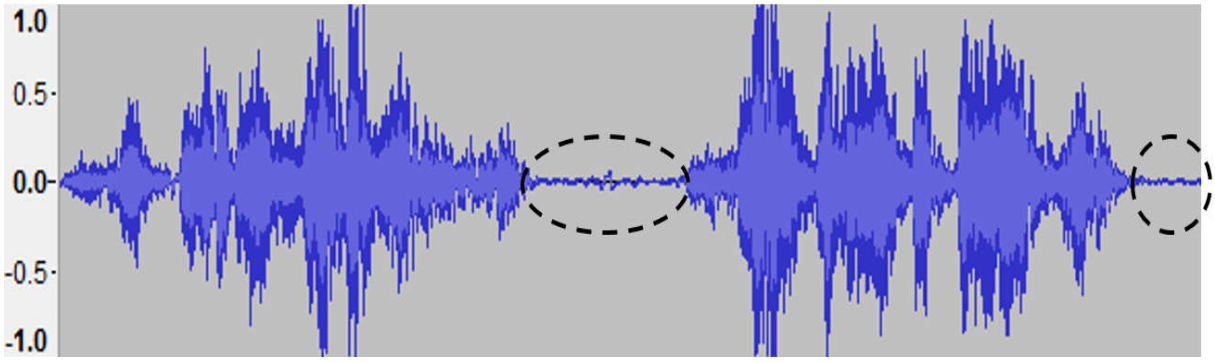

Fig.11. Speech signal before filtering

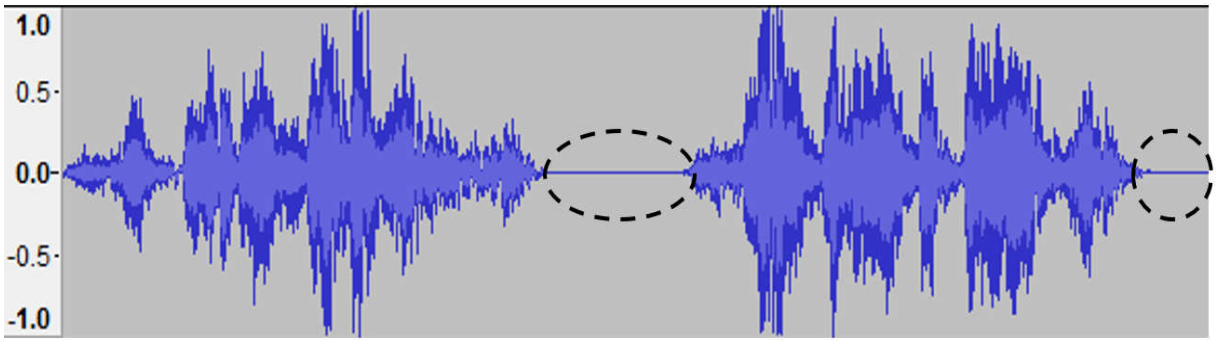

Fig.12. Speech signal after filtering

8. SPEECH ANNOTATIONS

The storytelling audio corpus was annotated manually using speech analysis tool known as Praat [19]. Each audio file is imported to Praat [25] and annotated at sentence, word and

syllable level. However, annotation also can be at a certain level only depending on the research attention. As example, in [12] only labelled their storytelling corpus of three Indian languages: Bengali, Hindi and Telugu at syllable level. After annotations of all 48 audio files are done, a total of 48 transcriptions are produced and stored in textgrid file (.textgrid).

Fig. 13 illustrated the transcription of neutral and storytelling speech file of a female speaker. In the figure, the first row consists of the speech waveform. The second row illustrated the pitch (i.e. blue coloured line) and intensity (i.e. green coloured line) patterns of the speech waveform. For annotation or labelling, the first tier (marked as 1) shows the automated sound/silent labelling of the phrase.

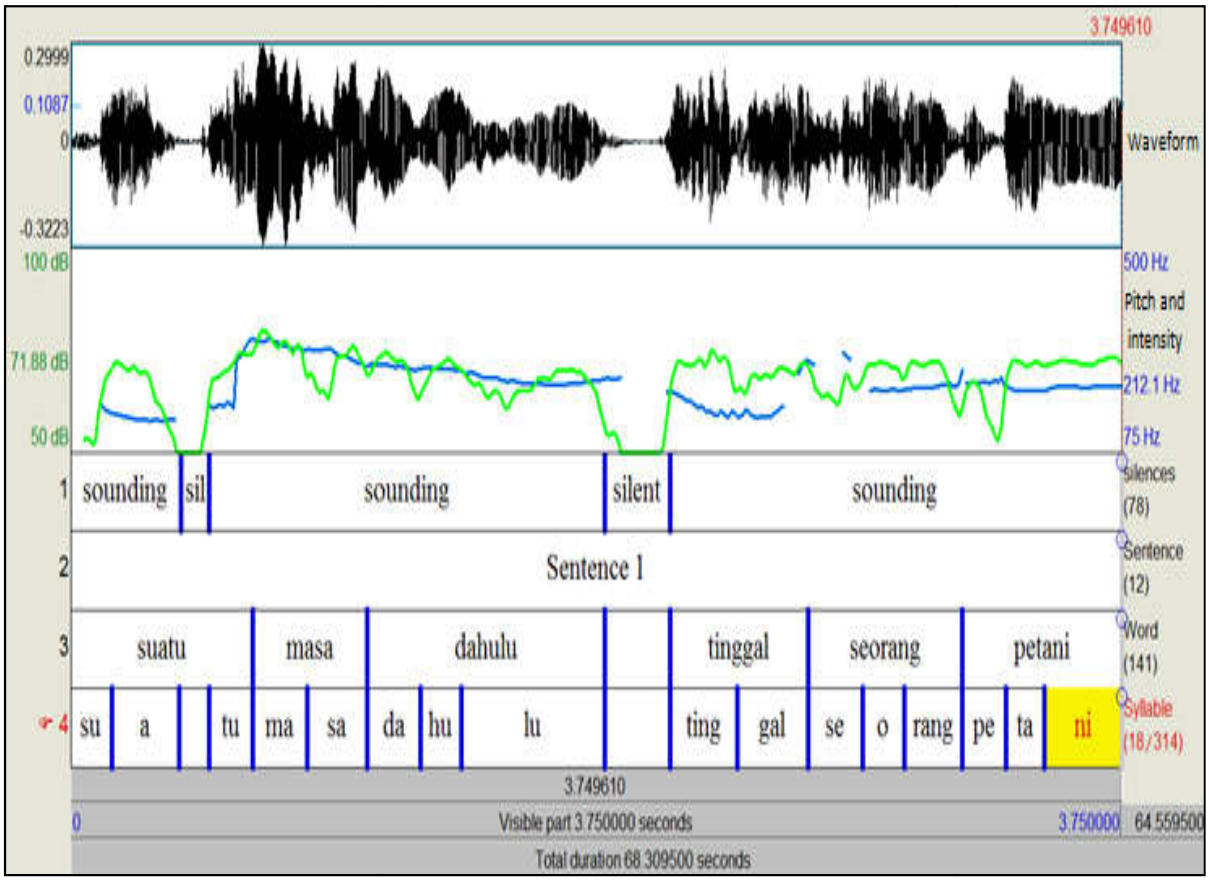

Fig.13. Storyteller speech (Phrase: "suatu masa dahulu, tinggal seorang petani")

Table 4. Tempo analysis of a story and storyteller

Story	MPS1	FSp1	FSp2	FSp3	MSp1	MSp1	FSp1	FSp2	Mean
Story 1	5.22	4.94	4.39	5.43	5.59	5.00	5.39	4.45	5.05
Story 2	5.14	4.93	4.3	5.65	6.14	5.34	5.37	4.36	5.15
Story 3	3.29	2.97	2.73	3.79	3.70	3.08	3.39	2.96	3.23
Mean	4.55	4.28	3.81	4.96	5.14	4.47	4.72	3.92	4.55

MPS1-Male professional speaker; FSt1-First female storyteller; FSt2-Second female storyteller; FSt3-Third female storyteller; MSp1-First male speaker; MSp2-Second male speaker; FSp1-First female speaker; FSp2-Second female speaker

Table 5. Syllable analysis for word type (adjective, adverb and intensifier)

Word	Type	Duration	Intensity	Pitch
Keemasan	Adjective	7	4	7
Menyesal	Adjective	6	5	5
Malang	Adjective	6	5	4
Tersasar	Adverb	6	6	4
Segera	Adverb	4	6	7
Sungguh	Intensifier	4	6	7
Sangat	Intensifier	7	6	8

The second, third and fourth tier (marked as 2, 3 and 4) shows manual labelling at sentence-, word- and syllable-level respectively. The syllables are labelled 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.

9. STORYTELLING ANALYSIS AND DISCUSSION

The storytelling speech was analyzed for each storyteller. Table 4 shows the tempo (s) for the story and storytellers. The tempo was calculated in syllable per second (SPS). The results showed that the average tempos for stories (i.e. story 1, story 2 and story 3) are 5.05, 5.15 and 3.23.

Average tempo for story 3 is much slower than the others. As we can see in Table 4, every storyteller reduces their tempo for story 3. It indicated that the tempo of the storytelling is depending on the script and content of the story.

Further analysis on syllable in storytelling speech was showed that storytellers will emphasize (increase duration, intensity and pitch) at the last syllable of the certain adjective and adverb word and at the first syllable of the intensifier word (*kata penguat*) as compared to neutral speech. Table 5 shows the total storytellers (from 8 storytellers) that emphasize the adjective, adverb and intensifier. Thus, these finding is crucial during modelling storytelling speech synthesis system.

10. CONCLUSION

This paper presented a new corpus of storytelling speech in Malay language. The efforts that were done are speech recording, pre-processing, and speech annotation. The analyzed data of storytelling prosody such as tempo, duration, intensity and pitch will be used for modeling of a storytelling speech synthesis system.

Thus, the complete storytelling synthesis system may be embedded to a humanoid storyteller that is able to speak in neutral and storytelling style. The corpus is also beneficial to speech recognition, and natural language processing. Upon request, the corpus is available to any academic institutions and publics to be used and contribute for other researches.

11. ACKNOWLEDGEMENTS

Due acknowledgment is accorded to the Research Management Center and Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA for the funding received through the BESTARI Grant (600-IRMI/DANA 5/3/BESTARI (0002/2016).

12. REFERENCES

[1] Kory J, Breazeal C. Storytelling with robots: Learning companions for preschool children's language development. In 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014, pp. 643-648

[2] Mutlu B, Forlizzi J, Hodgins J. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In 6th IEEE-RAS international conference on Humanoid Robots, 2006, pp. 518-523

[3] Gelin R, d'Alessandro C, Le Q A, Deroo O, Doukhan D, Martin J C, Pelachaud C, Rilliard A, Rosset S. Towards a storytelling humanoid robot. In 24th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, 2010, pp. 137-138

[4] Bremner P, Leonards U. Iconic gestures for robot avatars, recognition and integration with speech. Frontiers in Psychology, 2016, 7:1-14

[5] Kory J J. Storytelling with robots: Effects of robot language level on children's language learning. Master thesis, Cambridge: Massachusetts Institute of Technology, 2014

[6] Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In International Conference on Speech and Computer, 2003, pp. 8-17

[7] Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 2014, 56:85-100

[8] Khaw Y M, Tan T P. Preparation of MaDiTS corpus for Malay dialect translation and speech synthesis system. In Speech, Language and Audio in Multimedia Workshop, 2014, pp. 53-57

[9] Maekawa K, Koiso H, Furui S, Isahara H. Spontaneous speech corpus of Japanese. In 2nd International Conference on Language Resources and Evaluation, 2000, pp. 947-952

[10] Sarkar P, Haque A, Dutta A K, Reddy G, Harikrishna D M, Dhara P, Verma R, Narendra N P, Sunil K S B, 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. In 7th IEEE International Conference on Contemporary Computing, 2014, pp. 473-477

[11] Rebai I, BenAyed Y. Arabic text to speech synthesis based on neural networks for MFCC estimation. In IEEE World Congress on Computer and Information Technology, 2013, pp. 1-5

[12] Verma R, Sarkar P, Rao K S. Conversion of neutral speech to storytelling style speech. In 8th IEEE International Conference on Advances in Pattern Recognition, 2015, pp. 1-6

[13] Theune M, Meijs K, Heylen D, Ordelman R. Generating expressive speech for storytelling applications. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4):1137-1144

[14] Alm C O, Sproat R. Perceptions of emotions in expressive storytelling. In 9th European Conference on Speech Communication and Technology 2005, pp. 533-536

[15] Doukhan D, Rilliard A, Rosset S, Adda-Decker M, Alessandro C. Prosodic analysis of a corpus of tales. In 12th Annual Conference of the International Speech Communication Association, 2011, pp. 3129-3132

[16] Přibil J, Přibilová A. Application of expressive speech in TTS system with cepstral description. In A. Esposito, N. G. Bourbakis, N. Avouris, & I. Hatzilygeroudis (Eds.), Verbal and nonverbal features of human-human and human-machine interaction. Berlin: Springer, 2008, pp. 200-212

[17] Montaño R, Alías F, Ferrer J. Prosodic analysis of storytelling discourse modes and narrative situations oriented to text-to-speech synthesis. In 8th ISCA Workshop on Speech Synthesis, 2013, pp. 171-176

[18] Lee B, Hasegawa-Johnson M, Goudeseune C, Kamdar S, Borys S, Liu M, Huang T S.

AVICAR: audio-visual speech corpus in a car environment. In 8th International Conference on Spoken Language Processing, 2004, pp. 2489-2492

[19] Chowdhury S. Concatenative Text-to-speech synthesis: A study on standard colloquial bengali. Phd thesis, Kolkata: Indian Statistical Institute, 2006

[20] Seman N. Coalition of artificial intelligence (AI) algorithms for isolated spoken Malay speech recognition. Phd thesis, Selangor: Universiti Teknologi MARA, 2011

[21] Ikkunointi. Windowing. 2016

[22] Hamzah R. Discriminative classification model of filled pause and elongation for Malay language spontaneous speech. Selangor: Universiti Teknologi MARA, 2016

[23] Paliwal K K, Lyons J G, Wójcicki K K. Preference for 20-40 ms window duration in speech analysis. In 4th IEEE International Conference on Signal Processing and Communication Systems, 2010, pp. 1-4

[24] Podder P, Khan T Z, Khan M H, Rahman M M. Comparative performance analysis of hamming, Hanning and Blackman window. International Journal of Computer Applications, 2014, 96(18):1-7

[25] Boersma P, Weenink D. Praat: Doing phonetics by computer. [Computer program]. Version 6.0. 19

[26] Bahari A R, Musa A, Nuawi M Z, Rizman Z I, Saad S M. Novel statistical clustering method for accurate characterization of word pronunciation. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(4): 1172-1177

How to cite this article:
Ramli I, Jamil N, Seman N, Ardi N The first malay language storytelling text-to-speech (tts) corpus for humanoid robot storytellers. J. Fundam. Appl. Sci., 2017, *9(4S), 340-358.*