Acoust. Sci. & Tech. **42**, 6 (2021)

PAPER

SUST TTS Corpus: A phonetically-balanced corpus for Bangla text-to-speech synthesis

Arif Ahmad*, Md. Reza Selim, Md. Zafar Iqbal and M. Shahidur Rahman

Department of Computer Science and Engineering, Shahjalal Univeristy of Science and Technology, Sylhet-3114, Bangladesh

(*Received 30 March 2021, Accepted for publication 29 July 2021*)

Abstract: This paper presents the *Shahjalal University of Science and Technology Text-To-Speech Corpus (**SUST TTS Corpus**)*, a phonetically balanced speech corpus for Bangla speech synthesis. Due to the advancement of deep learning techniques, modern speech processing researches such as speech recognition and speech synthesis are being conducted in various deep learning methods. Any state-of-the-art neural TTS system needs a large dataset to be trained efficiently. The lack of such datasets for under-resourced languages like Bangla is a major obstacle for developing TTS systems in those languages. To mitigate this problem and accelerate speech synthesis research in Bangla, we have developed a large-scale, phonetically-balanced speech corpus containing more than 30 hours of speech. Our corpus includes 17,357 utterances spoken by a professional voice talent in a sound-proof audio laboratory. We ensure that the corpus contains all possible Bangla phonetic units in sufficient amounts, making it a *phonetically-balanced* speech corpus. We describe the process of creating the corpus in this paper. We also train a neural Bangla TTS system with our corpus and obtain a synthetic voice which is comparable to the state-of-the-art TTS systems.

Keywords: Speech synthesis, Bangla TTS, Phonetically balanced corpus, Merlin TTS

1. INTRODUCTION

Speech technology researches have been improved rapidly in recent years. These improvements are made possible because of the introduction of various deep learning techniques [1]. In a world of smartphones and intelligent devices, text-to-speech (TTS) or speech synthesis plays a crucial role. Not surprisingly, state-of-the-art TTS researches are being conducted with the help of deep learning tools as well. The quality of a synthesized speech generated by a neural text-to-speech system depends greatly on the quality and the size of the training data. Obtaining a free, high-quality, and large-scale TTS corpus for some resource-rich languages [2,3] is easier. But the same is not true for the under-resourced languages.

Bangla (also known as Bengali), is an Indo-Aryan language spoken by the inhabitants of Bangladesh and some portion of India. Although Bangla is the native language of more than 250 million people, resources for computational research in this language are inadequate. The lack of publicly available high-quality speech corpus in Bangla is one of the major hindrances to adopt any existing state-of-the-art neural TTS systems. Preparing a carefully curated speech corpus for speech synthesis is a challenging task. Only a few resource-rich languages (e.g. English) has a publicly available large TTS corpus. For Bangla TTS, there is no such large corpus available publicly. Although some attempts were made to prepare speech corpus for Bangla, such as [4] and [5], those are not large enough to train a modern TTS system. Nowadays a typical neural TTS system requires around 25 hours of speech data [2] to be able to generate natural-sounding speech.

In this connection, we have prepared a *clean* TTS corpus called the *Shahjalal University of Science and Technology Text-To-Speech Corpus (**SUST TTS Corpus**)*. It contains 30 hours of speech recording of over 17,000 sentences. It is a *phonetically-balanced* corpus, meaning that it contains all possible Bangla phonetic units in sufficient amounts. We have evaluated our corpus by building a *Bangla Statistical Parametric Speech Synthesizer* and obtained a satisfactory result compared to the existing Bangla TTS systems.

The subsequent sections of this report are organized as follows. We briefly discuss and summarize the previous works done in creating Bangla speech corpus in Sect. 2. Section 3 contains the description of our corpus prepara-

*e-mail: arif_ahmad-cse@sust.edu
[doi:10.1250/ast.42.326]

tion along with some statistics of the corpus. In Sects. 4 and 5, we give the evaluation of our corpus on a neural TTS system. Finally, Sect. 6 ends the report by indicating some applications and future scope of this work.

2. RELATED WORKS

Most of the text-to-speech systems in modern days are being developed using various deep learning techniques. TTS models based on Recurrent Neural Networks (RNNs) such as Tacotron [6], Tacotron 2 [7], Deep Voice [8], Deep Voice 2 [9] have shown impressive performance. But using RNNs has a higher computational cost and a slower inference time. This has led the researchers to build fully convolutional models such as DCTTS [10] and Deep Voice 3 [11]. All these neural TTS systems are data-hungry, they need a sufficiently large dataset to be trained effectively. The public implementations of the modern TTS models are usually trained on the LJSpeech dataset [2]. It is an English TTS corpus containing around 24.6 hours of speech recording of a single speaker. Some of the TTS models have also been trained on the datasets of other languages of comparable size. JSUT [12] (Japanese), Ruslan [13] (Russian), KSS dataset [14] (Korean), etc. are some of the examples of such datasets.

For developing a Bangla speech synthesizer with those modern TTS models, we need a large TTS corpus as well. There were some attempts in the recent past to develop such datasets. Mumit *et al.* released a Bangla annotated speech corpus [15] of around 13 hours for phonetic research. This can be used as a good starting point but is not quite enough for a deep learning TTS model. Google released a multi-speaker TTS corpus [5] of around 3 hours in 2016. *IndicSpeech* [16] is another attempt at creating TTS corpus for Indian languages, which contains a Bangla speech corpus of around 22 hours, but this dataset is not publicly released yet. In this work, we present a *large-scale*, *phonetically balanced* Bangla TTS corpus of 30 hours, which will enable us to further the Bangla TTS research.

3. CORPUS DESIGN

We describe the process of designing and preparing the *SUST TTS corpus* in this section. We start with a discussion of creating a phonetically balanced text corpus. Then we discuss the speaker selection and speech recording process in detail. We finalize the discussion by presenting some statistics of the corpus.

3.1. Preparing Phonetic Units

To compile a phonetically balanced text corpus, we need to ensure that all the *phonetic units* of Bangla are present in sufficient amounts. *Contextual phonetic units* are extensively applied to speech technology researches given their ability to encompass allophonic variation and coarticulation effects. We choose to keep track of *triphones* and some modified *syllables* for preparing our corpus. It is observed that triphones are capable of describing the surrounding environment of a given phone [17], and have a huge impact on the performance of the acoustic models for speech synthesis.

The general structure of Bangla phonetic units has three parts: the onset, the nucleus, and the coda. The nucleus is usually a vowel or diphthong. The onset is usually a consonant that comes before the nucleus, and the coda is the part that comes after the nucleus. Bangla language allows syllables/triphones with empty codas, and empty onsets as well. A phonetic unit of the form CV (consonant + vowel, with an empty coda) is called an open unit, while a unit that has a coda (CVC, etc.) is called a closed or checked unit [18].

The Bangla alphabet contains 11 letters representing vowels, but there are only 7 distinct vowel sounds present in the language [19]. The number of distinct consonant letters/phones is 28. We have created our triphone/syllable inventory by taking all possible combinations of vowels and consonants. We also considered the position of a phonetic unit in the word. For example, a phonetic unit /x/ can appear at the beginning (denoted as /{x/), middle (denoted as /x/), or at the end of a word (denoted as /x}/). We count all three occurrences of the unit separately. Bangla language has some phonetic units in the form CCV or CCCV. These are called *conjuncts*. In Bangla, there are 41 *conjuncts* that occur at the beginning of a word/syllable. We have counted the 18 *diphthongs* of Bangla as well. All of the above combinations would constitute around 1,400 phonetic units. But some of those units never appear in a valid text. After removing those invalid units, we get a total of **1,208** phonetic units to count. Table 1 shows the summary of the phonetic units we have prepared for measuring our corpus.

We make sure that all the phonetic units are present in our text corpus. Table 2 shows the 41 conjuncts that appear at the start of a word/syllable. The reason for listing these units is they were not discussed in any previous works of Bangla phonetic studies. So it will be a good resource for anyone who wants to explore this area of research in the future.

Table 1 Summary of prepared phonetic units.

Unit	Form	Amount
Vowel	V	7
Diphthong	VV	18
Syllable/Triphone	CV/VC/VCV	903
Conjunct	CCV/CCCV	280
Total		**1,208**

Table 2 List of 41 conjuncts that appear at the start of words/syllables.

Conjunct	Pronunciation	Example
ক্র	/kr/	ক্রমশ /krɔmɔʃ/
ক্ল	/kl/	ক্লেশ /kleʃ/
খ্র	/kʰr/	খ্রিস্টান /kʰriʃʈan/
খ্ল	/kʰl/	খ্লেবনিকফ /kʰlebnikɔpʰ/
গ্ল	/gl/	গ্লানি /glani/
গ্র	/gr/	গ্রহণ /grɔhɔn/
ঘ্র	/gʰr/	ঘ্রাণ /gʰran/
ট্র	/ʈr/	ট্রাফিক /ʈrapʰik/
ড্র	/dr/	ড্রাগন /dragɔn/
ত্র	/tr/	ত্রাস /traʃ/
থ্র	/tʰr/	থ্রোট /tʰroʈ/
দ্র	/dr/	দ্রবণ /drɔbɔn/
ধ্র	/dʰr/	ধ্রুবক /dʰrubɔk/
ন্র	/nr/	নৃপতি /nripɔti/
প্র	/pr/	প্রহরী /prɔhɔri/
প্ল	/pl/	প্লাবন /plabɔn/
ফ্র	/fr/	ফ্রেম /frem/
ফ্ল	/fl/	ফ্লোর /flɔr/
ব্র	/br/	ব্রাত্য /bratːɔ/
ব্ল	/bl/	ব্লগার /blɔgar/
ভ্র	/bʰ/	ভ্রমণ /bʰrɔmɔn/
ম্র	/mr/	ম্রিয়মান /mrijɔman/
ম্ল	/ml/	ম্লান /mlan/
শ্র	/sr/	শ্রমিক /srɔmik/
শ্ল	/ʃl/	শ্লথ /ʃlɔtʰ/
স্র	/sr/	স্রস্টা /srɔʃʈa/
স্ল	/sl/	স্লোগান /slogan/
স্ট্র	/sʈr/	স্ট্রোক /sʈrok/
স্ট	/sʈ/	স্টেশন /sʈeʃɔn/
স্প	/sp/	স্পন্দন /spɔndɔn/
স্ক্র	/skr/	স্ক্রিপ্ট /skripʈ/
স্ক্ল	/skl/	স্ক্লেরা /sklera/
স্ক	/sk/	স্কুল /skul/
স্খ	/skʰ/	স্খলন /skʰɔlɔn/
স্ত্র	/str/	স্ত্রীরোগ /strirog/
স্ত	/st/	স্তবক /stɔbɔk/
স্থ	/stʰ/	স্থাপত্য /stʰapɔtːɔ/
স্ন	/sn/	স্নায়ু /snai̯u/
স্ফ	/sf/	স্ফটিক /sfɔʈik/
স্ম	/sm/	স্মর্তব্য /smɔrtɔbːɔ/
স্প্র	/spr/	স্প্রিন্টার /sprinʈar/

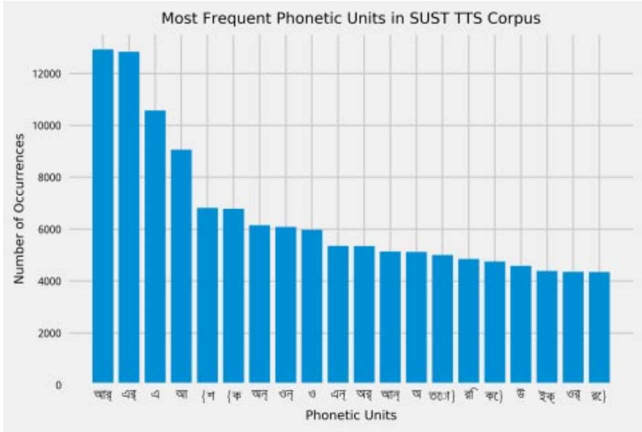

Fig. 1 List of the most frequent 20 phonetic units in *SUST TTS Corpus* with their frequencies.

The process of creating text corpus will be discussed in Sect. 3.2. Figure 1 shows the most frequent 20 units of the SUST TTS corpus.

3.2. Preparing Text Corpus

We gathered Bangla text data from various sources, such as Wikipedia [20], online newspapers, dramas, and novels from Bangla literature. After collecting the texts, we perform some preprocessing such as normalizing the texts, removing sentences with non-standard words (NSWs), changing the abbreviations, acronyms, etc. into their full-forms. The processed texts are then stored as stand-alone sentences. To create the *SUST TTS Corpus*, at first we pick 1,000 random sentences from our collection. Then we apply a greedy algorithm [18] to select phonetically rich sentences among the rest of the 100,000 sentences of the collection. We have selected a total of 17,357 sentences for our corpus. We organize the sentences in 22 scripts. Each of the scripts contains around 800–1,000 sentences. The sentences are placed in the scripts in the following format:

```
[ID] [TEXT]
```

Here `ID` refers to the unique identification number for each of the utterance. `TEXT` is the actual sentence written in `utf-8` format. `ID` and `TEXT` are separated by a `TAB` character. Section 3.4. shows some statistics of the corpus.

3.3. Speaker Selection and Recording Process

After preparing the text corpus, the next step is to record the audio corresponding to each of the sentences of the corpus. We choose a 24-year old male voice talent from a pool of speakers for this purpose. He is a professional voice artist and has a very clear pronunciation.

The recording was done in our audio recording lab using noise-reduction hardware. The Audacity [21] software was used for taking the audio clips. The speech data were sampled at 48 kHz and were stored in 16-bit PCM WAV format. The data were collected in several recording sessions. Each of the sessions usually lasted for 2 to 3 hours. The speaker was allowed to take small breaks in regular intervals. The sentences were recorded in one go, and the speaker tried to articulate them as naturally as possible. Some sentences were rerecorded if any error or unnaturalness were observed. The total duration of the recorded speech after clipping additional silences is around 30 hours.

Table 3 List of the most frequent 20 words of the *SUST TTS Corpus*.

Word	Pronunciation	Frequency
ও	/o/	1887
করে	/kɔre/	1678
না	/na/	1633
এই	/ei/	1513
এবং	/ebɔng/	1356
করা	/kɔra/	1135
থেকে	/t^{h}eke/	925
হবে	/hɔbe/	884
একটি	/ekʈi/	881
এর	/er/	848
এ	/e/	819
জন্য	/ʤɔnmɔ/	789
হয়	/hɔe̯/	725
তিনি	/tini/	678
তার	/tar/	655
কিন্তু	/kintu/	631
আর	/ar/	622
কি	/ki/	611
কী	/ki/	585
যে	/ʤe/	579

Table 4 Summary of *SUST TTS Corpus*.

Total sentences	17,357
Total words	193,826
Total unique words	36,831
Minimum words in a sentence	3
Maximum words in a sentence	20
Average words in a sentence	11.05
Total duration of speech (hours)	30 : 05 : 07
Average duration of each sentence (seconds)	6.24

3.4. Corpus Statistics

Table 4 gives a summary of the *SUST TTS corpus*. As the table indicates, our corpus has a large coverage of text with more than 17,000 sentenses.

The *SUST TTS Corpus* contains a total of 193,826 words. It has more than 36,000 unique words. Table 3 shows the most frequent 20 words of the corpus with their frequency.

Our corpus contains sentences of varying lengths. We did not pick sentences that contain less than 3 words or more than 20 words, since those would not be efficient for training a neural network. Figure 2 shows the distribution of words in the sentences in our corpus. The *mode* of the distribution is 8, and the *mean* is about 11.

We plan to release the SUST TTS corpus for academic and commercial use. Since it is a part of an ongoing Ph.D. research, we have released the corpus for academic use. After the completion of the research, we will make it

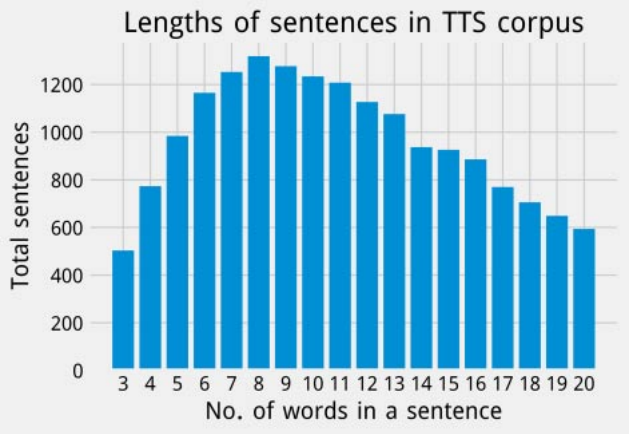

Fig. 2 Distribution of words across the sentences of *SUST TTS Corpus*.

available for both commercial and academic use. For the time being, the corpus can be obtained by filling out the following Google Form.

`https://forms.gle/uAv4hJaFjPR8X3976`

4. EXPERIMENTS

We use a Bangla neural speech synthesizer [22] which is based on Merlin [23], an open-source speech synthesis toolkit. The core of the model is two deep neural networks (DNNs), one for duration modeling, and the other for acoustic modeling. Figure 3 shows the functional diagram of the model. The complete architecture of the model can be found in the corresponding paper [22], but a brief description is presented here for quick referece.

The raw text is fed to a text normalizer. Although our dataset already contains the normalized texts, we employ the normalizer to be able to synthesize any text given by the users. The normalized text then goes to the language-independent front-end processor, called Ossian [24], to get the linguistic features from the text. It produces a feature vector, which is then fed to the duration model. The duration model consists of a deep neural network containing 3 layers of hidden units. Each of the hidden layers contains 512 neurons. The network uses the gradient descent optimizer with a learning rate of 0.002. The output of this network is then fed to the acoustic model. The acoustic DNN was trained to map the input linguistic features and the associated duration features into acoustic features. This network consists of 6 layers of hidden units where each layer contains 1,024 neurons. The output of the acoustic model is normalized appropriately so that it can be used by the WORLD [25] vocoder to generate the waveforms.

5. RESULTS AND DISCUSSION

To assess the quality of the speech synthesizer trained

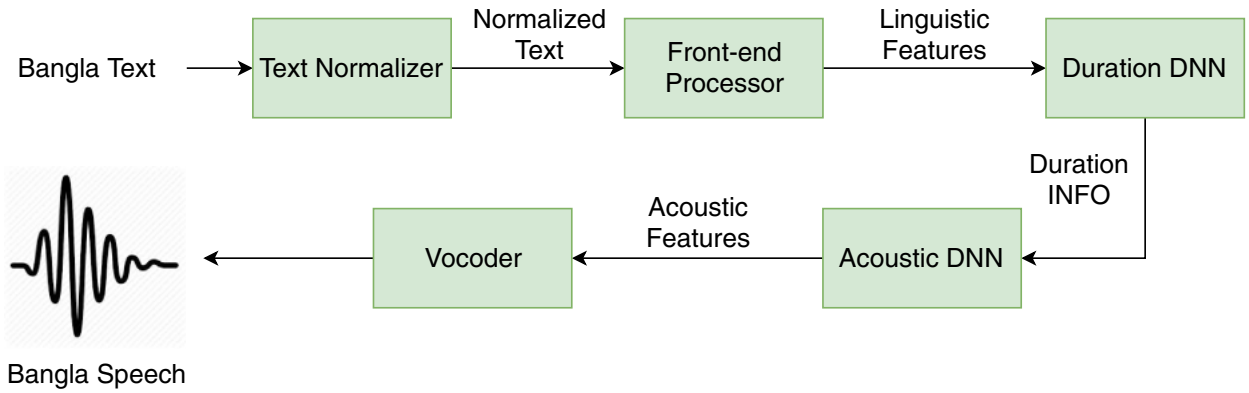

Fig. 3 Block diagram of the Bangla TTS System [22].

on *SUST TTS Corpus*, we have conducted both the objective and the subjective evaluation. For objective evaluation, we choose the Perceptual Evaluation of Speech Quality (PESQ) [26] score. The raw-PESQ scores are mapped to 'Mean Opinion Score-Listening Quality Objective' (MOS-LQO). This is standard practice for measuring the objective listening quality of speech signals. The intervals of raw-PESQ and MOS-LQO are $[-0.5, 4.5]$ and $[1, 5]$, respectively. The experimental set-up for determining the PESQ score is as follows. We train our TTS model [22] with four datasets. The description of the resultant TTS voices are mentioned in Table 5. *SUST TTS-1* is trained on a 10-hour dataset prepared by [4]. *SUST TTS-2* and *SUST TTS-3* are trained on 10-hour and 20-hour datasets, respectively, that are prepared by us. Finally, *SUST TTS-4* is trained on our proposed *phonetically balanced* 30-hour dataset.

Each of the datasets mentioned above is split into 3 (three) parts: training set, validation set, and test set. The ratio of the splitting is as follows: training set 96%, validation set 2% and test set 2%. Table 6 shows the breakdown of the samples for the *SUST TTS Corpus*.

Table 5 Description of the trained TTS voices.

TTS Voice	Speaker	Dataset
SUST TTS-1	Male	10-hour [4]
SUST TTS-2	Female	10-hour
SUST TTS-3	Male	20-hour
SUST TTS-4	Male	30-hour

Table 6 Breakdown of the *SUST TTS Corpus* data.

Type	Ratio	#Samples
Training Set	96%	16,663
Validation Set	2%	347
Test Set	2%	347

For calculating the PESQ scores of the TTS voices, we picked 100 random sentences from the test set and corresponding recordings as the original speech data. Then, we synthesized 100 waveforms of the selected sentences from the corresponding TTS voices. Finally, the original and the synthetic speeches were sent to PESQ algorithm in pairs for determining the PESQ score. The candidate voice samples were aligned in the time domain before being compared. The time-alignment technique is integrated into the PESQ algorithm. Figure 4 shows the average PESQ scores for the four voices mentioned above.

For subjective evaluation, we used a benchmarking test for the TTS systems, which is called the *Mean Opinion Score (MOS)*. We invited some native Bangla speakers to volunteer for the test. 30 people participated as the listeners

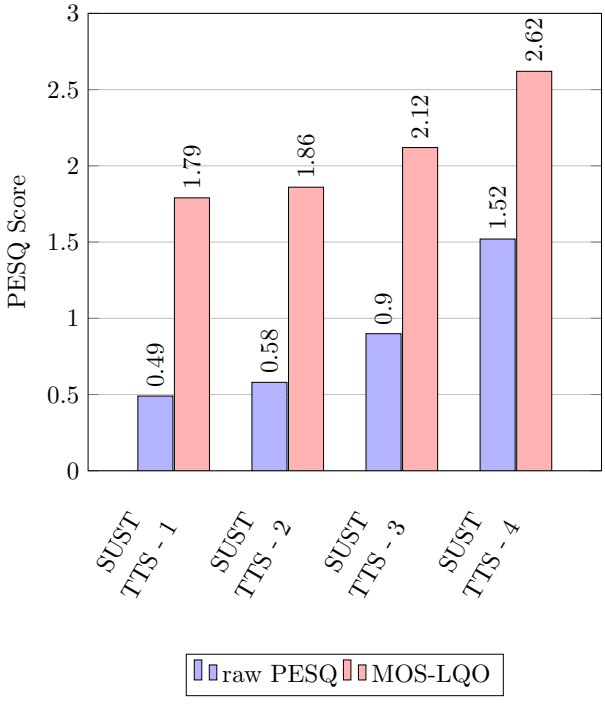

Fig. 4 Average PESQ Scores of Different TTS voices.

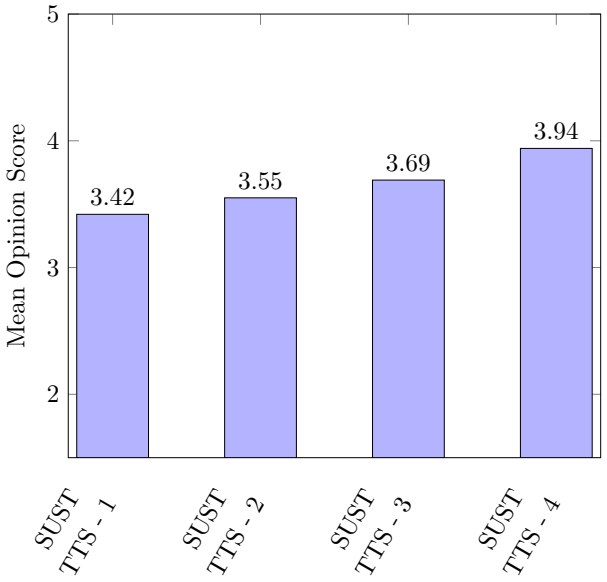

Fig. 5 Naturalness Comparison between Various TTS Voices.

of the MOS test. The listeners are between 20 to 35 years old. 18 of them were male, and the rest were female. We selected 20 random sentences outside of our corpus and synthesized speech with the various TTS voices. The volunteers listened to those 20 sentences and gave a naturalness score between 1 and 5 to each of the systems. A higher score means better naturalness. The scores were then averaged to get the mean score for a system. We excluded the highest and lowest score obtained by each system to calculate a more accurate MOS, namely the Robust-MOS. Figure 5 shows the MOS scores obtained for the various voices.

The results of our MOS test support the objective evaluation tests as well. Both the results confirm that the SUST TTS-4 voice, which is trained on our phonetically balanced 30-hour dataset, has outperformed the other competitors. The carefully curated corpus has helped it achieve a performance comparable to the current state-of-the-art Bangla TTS system [5].

6. CONCLUSION

We have presented a large-scale phonetically-balanced speech corpus for Bangla text-to-speech synthesis in this report. A good dataset is essential for modern deep-learning-based speech synthesis research. For Bangla TTS, there was no such corpus available publicly. We have prepared and are planning to release this dataset with the expectation that it will alleviate the TTS research in Bangla. Our corpus contains 30 hours of clean speech with rich coverage of phonetic units. Empirical results show that this corpus can be used to train a neural TTS system efficiently. Expanding the domains of the corpus and including speeches of varying emotions would be the target of future works.

REFERENCES

[1] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016), http://www.deeplearningbook.org.

[2] K. Ito and L. Johnson, "The LJSpeech dataset," https://keithito.com/LJ-Speech-Dataset/ (2017).

[3] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *Proc. Interspeech 2019*, pp. 1526–1530 (2019).

[4] S. M. Murtoza, "Phonetically balanced Bangla speech corpus" (2011).

[5] A. Gutkin, L. Ha, M. Jansche, K. Pipatsrisawat and R. Sproat, "TTS for low resource languages: A Bangla synthesizer," *Proc. 10th Int. Conf. Language Resources and Evaluation* (*LREC 2016*), pp. 2005–2010 (2016).

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010 (2017). [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2018*, pp. 4779–4783 (2018).

[8] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta and M. Shoeybi, "Deep Voice: Real-time neural text-to-speech," *Proc. ICML 2017*, pp. 195–204 (2017).

[9] A. Gibiansky, S. Ö. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman and Y. Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," *Proc. NIPS 2017*, pp. 2966–2974 (2017).

[10] H. Tachibana, K. Uenoyama and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2018*, pp. 4784–4788 (2018).

[11] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv: Sound* (2018).

[12] R. Sonobe, S. Takamichi and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," pp. 1–4 (2017). [Online]. Available: http://arxiv.org/abs/1711.00354

[13] L. Gabdrakhmanov, R. Garaev and E. Razinkov, "RUSLAN: Russian spoken language corpus for speech synthesis," in *Speech and Computer*, A. A. Salah, A. Karpov and R. Potapova, Eds. (Springer International Publishing, Cham, 2019), pp. 113–121.

[14] K. Park, "Korean Single Speaker Speech Dataset," Mar. (2020). [Online]. Available: https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset

[15] F. Alam, S. M. Habib, D. A. Sultana and M. Khan, "Development of annotated Bangla speech corpora," *Proc. SLTU 2010*, 7 pages (2010).

[16] N. Srivastava, R. Mukhopadhyay, K. R. Prajwal and C. Jawahar, "IndicSpeech: Text-to-speech corpus for Indian

languages," *Proc. LREC 2020*, pp. 6417–6422 (2020).
[17] G. Mendonça, S. Candeias, F. Perdigão, C. Shulby, R. Toniazzo, A. Klautau and S. Aluísio, "A method for the extraction of phonetically-rich triphone sentences," *Proc. Int. Telecommunications Symp.* (*ITS*) *2014*, pp. 1–5 (2014).
[18] K. Arora, S. Arora, K. Verma and S. S. Agrawal, "Automatic extraction of phonetically rich sentences from large text corpus of indian languages," *Proc. Interspeech 2004*, pp. 2885–2888 (2004).
[19] F. Alam, S. M. Habib and M. Khan, "Acoustic analysis of Bangla vowel inventory" (2008).
[20] A. Khatun, "Bangla Wikipedia dataset," Dec. (2019). [Online]. Available: https://data.mendeley.com/datasets/3ph3n78fp7/2
[21] Audacity Team 2020, "Audacity® Software, version 2.4.2" (2020) [Online]. Available: https://audacityteam.org/
[22] R. S. Raju, P. Bhattacharjee, A. Ahmad and M. S. Rahman, "A Bangla text-to-speech system using deep neural networks," *Proc. Int. Conf. Bangla Speech and Language Processing* (*ICBSLP*) *2019*, pp. 1–5 (2019).
[23] Z. Wu, O. Watts and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW 2016*, pp. 202–207 (2016).
[24] "Ossian: A simple language independent text-to-speech front-end." [Online]. Available: https://github.com/CSTR-Edinburgh/Ossian (accessed 28 Jul. 2020).
[25] M. Morise, F. Yokomori and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, **99-D**, 1877–1884 (2016).
[26] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) — A New Method for Speech Quality Assessment of Telephone Networks and Codecs," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2001* (*Cat. No. 01CH37221*), Vol. 2. IEEE, pp. 749–752 (2001).