MnTTS2: An Open-Source Multi-Speaker Mongolian Text-to-Speech Synthesis Dataset

Kailin Liang[†], Bin Liu[†], Yifan Hu[†], Rui Liu[⋆], Feilong Bao, and Guanglai Gao

Inner Mongolia University, Hohhot, China
liangkailin98@foxmail.com,iframe_liu@163.com,hyfwalker@163.com
liurui_imu@163.com, {csfeilong, csggl}@imu.edu.cn

Abstract. Text-to-Speech (TTS) synthesis for low-resource languages is an attractive research issue in academia and industry nowadays. Mongolian is the official language of the Inner Mongolia Autonomous Region and a representative low-resource language spoken by over 10 million people worldwide. However, there is a relative lack of open-source datasets for Mongolian TTS. Therefore, we make public an open-source multi-speaker Mongolian TTS dataset, named MnTTS2, for the benefit of related researchers. In this work, we prepare the transcription from various topics and invite three professional Mongolian announcers to form a three-speaker TTS dataset, in which each announcer records 10 hours of speeches in Mongolian, resulting 30 hours in total. Furthermore, we build the baseline system based on the state-of-the-art FastSpeech2 model and HiFi-GAN vocoder. The experimental results suggest that the constructed MnTTS2 dataset is sufficient to build robust multi-speaker TTS models for real-world applications. The MnTTS2 dataset, training recipe, and pretrained models are released at: `https://github.com/ssmlkl/MnTTS2`.

Keywords: Mongolian · Text-to-Speech (TTS) · Open-Source Dataset · Multi-Speaker.

1 Introduction

Text-to-Speech (TTS) aims to convert the input text to human-like speech [1]. It is a standard technology in human-computer interaction, such as cell phone voice assistants, car navigation, smart speakers, etc. The field of speech synthesis has developed rapidly in recent years. Different from the traditional methods, which use concatenation [2], statistical modeling [3] based methods to synthesize speech, neural end-to-end TTS models achieve remarkable performance with the help of Encoder-Decoder architecture [4]. Typical models include Tacotron [5], Tacotron2 [1], Transformer TTS [6], Deep Voice [7], etc. To further accelerate

the inference speed, the non-autoregressive TTS models, such as FastSpeech [8], FastSpech2(s) [9], are proposed and become the mainstream methods of TTS. Note that armed with the neural network based vocoder, including WaveNet [10], WaveRNN [11], MelGAN [12], HiFi-GAN [13], etc., the TTS model can synthesize speech sounds that are comparable to human sounds.

We note that an important factor in the rapid development of neural TTS mentioned above is the large scale corpus resources. This is especially true for languages such as English and Mandarin, which are widely spoken worldwide. However, low-resource language such as Mongolian [14] have been making slow progress in related research due to the difficulties in corpus collection. Therefore, building a large-scale and high-quality Mongolian TTS dataset is necessary. In addition, our lab have previously open-sourced a single-speaker dataset called MnTTS [15], which was recorded by a young female native Mongolian speaker and received much attention from academia and industry upon its release. This also shows the necessity of continuing to collect and organize Mongolian speech synthesis datasets and opening the baseline model's source code.

Motivated by this, this paper presents a multi-speaker Mongolian TTS dataset, termed as MnTTS2, which increases the number of speakers to three and increases the data size from 8 to 30 hours, with an average of 10 hours per speaker. The textual content has been further expanded and enriched in the domain. Similar with our MnTTS, MnTTS2 dataset is freely available to academics and industry practitioners.

To demonstrate the reliability of MnTTS2, we combined the state-of-the-art FastSpeech2 [9] model and the HiFi-GAN [13] vocoder to build the accompanied baseline model for MnTTS2. We conduct listening experiments and report the Naturalness Mean Opinion Score (N-MOS) and Speaker Similarity Mean Opinion Score (SS-MOS) results in terms of naturalness and speaker similarity respectively. The experimental results show that our system can achieve satisfactory performance on the MnTTS2, which indicates that the MnTTS2 corpus is practically usable and can be used to build robust multi-speaker TTS system.

The main contributions are summarized as follows. 1) We developed a multi-speaker TTS dataset, termd as MnTTS2, containing three speakers. The total audio duration is about 30 hours. The transcribed text covers various domains, such as sports and culture, etc. 2) We used the state-of-the-art non-autoregressive FastSpeech2 model to build the baseline model and validate our MnTTS2. 3) The MnTTS2 dataset, source code, and pre-trained models will be publicly available to academics and industry practitioners.

The rest of the paper is organized as follows. Section 2 revisits the related works about the Mongolian TTS corpus. In Section 3, we introduce the details of MnTTS2, including the corpus structure and statistical information. Section 4 explains and discusses the experimental setup and experimental results. Section 5 discusses the challenges faced by Mongolian speech synthesis and the future research directions. Section 6 concludes the paper and summarizes the work and research of this paper.

2   Related Work

For mainstream languages such as English and Mandarin, there are many free and publicly available TTS datasets. For example, LJSpeech [16] is a single-speaker dataset for English. To rich the speaker diversity, some multi-speaker TTS dataset are released, such as Libritts [17] for English and Aishell [18] for Chinese.

For the low-resource language such as Mongolian, the available resources are pretty limited. We note that some attempts tried to improve the effect of TTS synthesis under low resource data with unsupervised learning [19], semi-supervised learning [20], and transfer learning [21] methods, etc. However, due to the lack of large-scale training data, all the mentioned methods are difficult to achieve the effect that meets the requirements of practical scenarios.

In order to promote the development of Mongolian TTS, some works built their own Mongolian TTS corpus and designed various models to achieve good results. For example, Huang et al. established the first emotionally controllable Mongolian TTS system and achieved eight emotional embeddings by transfer learning and emotional embedding [22]. Rui Liu et al. introduced a new method to segment Mongolian words into stems and suffixes, which greatly improved the performance of the Mongolian rhyming phrase prediction system [23]. Immediately after that, Rui Liu proposed a DNN-based Mongolian speech synthesis system, which performs better than the traditional HMM [24]. Also, he introduced the Bidirectional Long Term Memory (BiLSTM) model to improve the phrase break prediction step in the traditional speech synthesis system, making it more applicable to Mongolian [25]. Unfortunately, none of the Mongolian TTS dataset from the above works have been released publicly and are not directly available to the public. We also found that some dataset in related fields, such as M2ASR-MONGO [26] for Mongolian speech recognition, are public recently. However, the speech recognition corpus cannot be applied in the TTS filed due to the environment noise and improper speaking style issues etc.

We previously released the single-speaker MnTTS dataset [15], called MnTTS. The total duration of the MnTTS is 8 hours, and it was recorded in a studio by a professional female native Mongolian announcer. However, the duration and speaker diversity still needs to be further expanded. In a nutshell, it is necessary to construct a high-quality multi-speaker Mongolian TTS dataset to further promote the Mongolian TTS research, which is the focus of this paper. We will introduce the details of the MnTTS2 at the following subsection.

3   MnTTS2 Dataset

In this section, we first revisit the MnTTS dataset briefly and then introduce our MnTTS2 by highlighting the extended content.

3.1   MnTTS

In the preliminary work, we presented a high-quality single-speaker Mongolian TTS dataset, called MnTTS [15]. The transcription of the dataset was collected from a wide range of topics, such as policy, sports, culture, etc. The Mongolian script was then converted to Latin sequences to avoid as many miscoding issues as possible. A professional female native Mongolian announcer was invited to record all the audio. A Mongolian volunteer was invited to check and re-align the alignment errors. The audio containing ambient noise and mispronunciation was removed to ensure the overall quality.

MnTTS has received much attention from researchers in the same industry upon its release. Furthermore, the subset was used in the Mongolian Text-to-Speech Challenge under Low-Resource Scenario at NCMMSC2022 [1]. The organizers provided two hours of data for all participants to train their models. This competition also promotes the development of intelligent information processing in minority languages within China.

3.2   MnTTS2

The construction pipeline of MnTTS2 consists of "Text collection and narration", "Text preprocessing" and "Audio recording and audio-text alignment". We will introduce them in order and then report the corpus structure and statistics.

Text collection and narration Similar with MnTTS [15], the first step in building the MnTTS2 dataset is to collect a large amount of transcription. The natural idea for collecting such a text materials is crawl text information from websites and electronic books. The text topics should cover human daily usage scenarios as much as possible. Following this, we crawled 23, 801 sentences, that are rich in content and have a wide range of topics (e.g., politics, culture, economy, sports, etc.), to meet our requirements well. At the same time, we manually filtered and removed some texts with unsuitable content, which may involve sensitive political issues, religious issues or pornographic content. These contents are removed in the hope that our dataset can make a positive contribution to the development of Mongolian language, which is the original intention of our work.

Text preprocessing Compared to mainstream languages, such as Mandarin and English, traditional Mongolian performs agglutinative characteristic [27]. This makes the Mongolian letters express different styles in different contexts and brings a serious harmonic phenomenon [15]. In order to solve this problem, we transformed the texts into a Latin alphabet, instead of traditional Mongolian representation, for TTS training. The entire pipeline of converting Mongolian texts into Latin sequences is divided into three steps: encoding correction, Latin conversion and text regularization. The detailed description can be found in our previous work MnTTS [15].
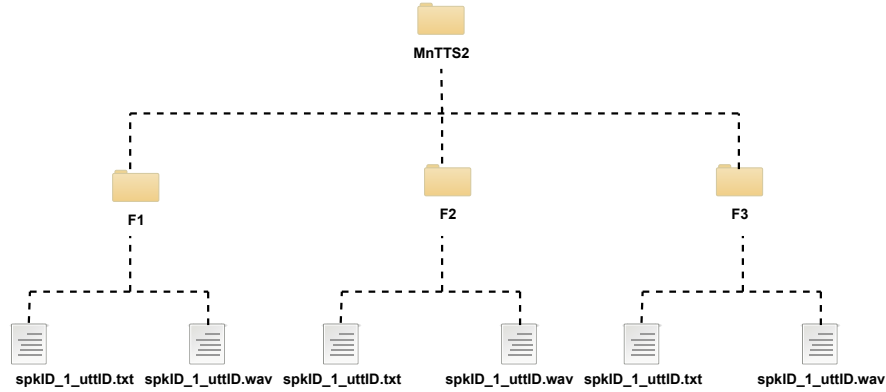
---

[1] `http://mglip.com/challenge/NCMMSC2022-MTTSC/index.html`

Fig. 1: The folder structure of the MnTTS2 corpus.

Audio recording and audio-text alignment Different with the MnTTS [15], we invited three native Mongolian-speaking announcers to record the audio. Each announcer volunteered to participate and signed an informed consent form to be informed of the data collection and use protocol. F1, F2 and F3 are three native Mongolian speaking females, with F2 being a little girl and F1 and F3 being slightly older in grade. All recordings were done in a standard recording studio at Inner Mongolia University. We choose Adobe Audition [2] as the recording software.

During the recording process, we asked the announcer to keeps a 0.3s pause at the beginning and end of each audio segment, keeps a constant distance between the lips and the microphone, performs a slight pause at the comma position, and performs an appropriate pitch boost at the question mark position.

To ensure the quality of the recording data , we rechecked the recording data after completing the recording work. Specifically, we invited three volunteers to check each text against its corresponding natural audio. These volunteers are responsible for splitting the recorded audio file into sentences and aligning the split sentences with the text. The Mongolian text is represented by a Latin sequence, where each Latin word in the sequence becomes a word and each letter that makes up the word is called a character. Characters also include punctuation marks, such as commas (','), periods ('.') , question mark ('?') , exclamation mark ('!') etc. Finally, we obtained about 30h of speech data, which were sampled at 44.1kHz with a sampling accuracy of 16bit.

Corpus structure and statistics The file structure of the MnTTS2 corpus is shown in Fig. 1. Each speaker's recording file and the corresponding text collection are saved in a folder named after the speaker. All audios are stored in

---

[2] `https://www.adobe.com/cn/products/audition.html`

Table 1: The statistics results of MnTTS2 dataset.

Statistical Unit ╲ Speaker ID		F1	F2	F3
Character	Total	572016	459213	601366
	Mean	79	61	67
	Min	12	2	2
	Max	189	188	190
Word	Total	88209	71245	92719
	Mean	12	9	10
	Min	3	1	1
	Max	29	30	29

(a) F1   (b) F2   (c) F3

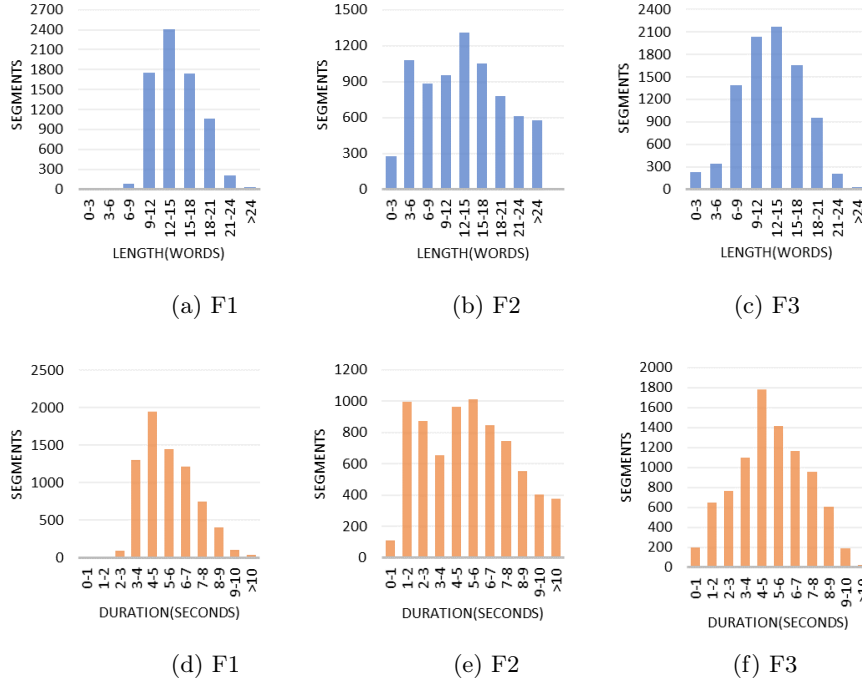(d) F1   (e) F2   (f) F3

Fig. 2: Word number distributions (a, b, c) and sentence duration distributions (d, e, f) for all speakers of MnTTS2.

WAV format files , sampled at 44.10kHz, and coded in 16 bits. All text is saved in a TXT file encoded in UTF-8. The file name of the audio is the same as the corresponding text file name, and the name of each file consists of the speaker, document ID, and corpus ID.

The statistical results of the MnTTS2 data are shown in Table 1 and Fig. 2. As shown in Table 1, the entire corpus has a total of 23, 801 sentences. Take the speaker F1 for example, F1 with a total of 572,016 Mongolian characters, and the average number of characters per sentence is 79, with the shortest sentence

having 12 characters and the longest sentence having 189 characters. If words are used as the statistical unit, the total number of words in this dataset for F1 is 88,209, the mean value of words in each sentence is 12, the minimum value is 3, and the maximum value is 29. As shown in Fig. 2, we also counted the sentence duration to draw a histogram. Take speaker F1 for example, the word numbers of the sentences are concentrated in 12-15, and duration are concentrated in 4-5 seconds. In comparison, we found that the word numbers of sentences for F2 was not particularly concentrated, and the duration were relatively scattered. F3, on the other hand, is more similar to F1, with a more obvious concentration. The statistics of all three speakers are in line with the normal distribution.

4 Speech Synthesis Experiments

To verify the validity of our MnTTS2, we conducted Mongolian TTS experiments based on the FastSpeech2 model and HiFi-GAN vocoder and evaluated the synthesized speech using Mean Opinion Score (MOS) metric in terms of naturalness and speaker similarity.

4.1 Experimental Setup

We use the TensorFlowTTS toolkit [3] to build an end-to-end TTS model based on the FastSpeech2 model. The FastSpeech2 model converts the input Mongolian text into Mel-spectrogram features, and then the HiFi-GAN vocoder reconstructs the waveform by the Mel-spectrogram features. FastSpeech2 is a state-of-the-art non-autoregressive [28] speech synthesis model that extracts duration, pitch, and energy directly from the speech waveform and uses these features as input conditions in training. This model can effectively solve errors such as repetition and word skipping, and has the advantage of fast training speed. FastSpeech2 introduces more variance information to alleviate the one-to-many mapping problem. Also, the pitch prediction is improved by wavelet transform. Most of all, FastSpeech2 has the characteristics of fast, robust, and controllable speech synthesis. This is the main reason why we choose FastSpeech2. As shown in Fig. 3, based on FastSpeech2, we implement the multi-speaker FastSpeech2 by adding the speaker encoder module. The speaker encoder includes speaker embedding, dense, and softplus layers [3]. In the network architecture setting, the number of speakers is 3. The dimension of speaker embedding is 384. The hidden side of the text encoder is 384 and the number of hidden layers is 4, the hidden layer size of the decoder is 384 and the number of hidden layers is 4. The number of Conv layers of the variance predictors is 2 and the dropout rate is 0.5. The initial learning rate is 0.001 and the dropout rate is 0.2.

The HiFi-GAN vocoder builds the network through a generative adversarial network to converts Mel-spectrogram into high-quality audio. The generator of HiFi-GAN consists of an upsampling structure, which consists of a one-dimensional transposed convolution, and a multi-receptive filed fusion module,

---

[3] https://github.com/TensorSpeech/TensorFlowTTS
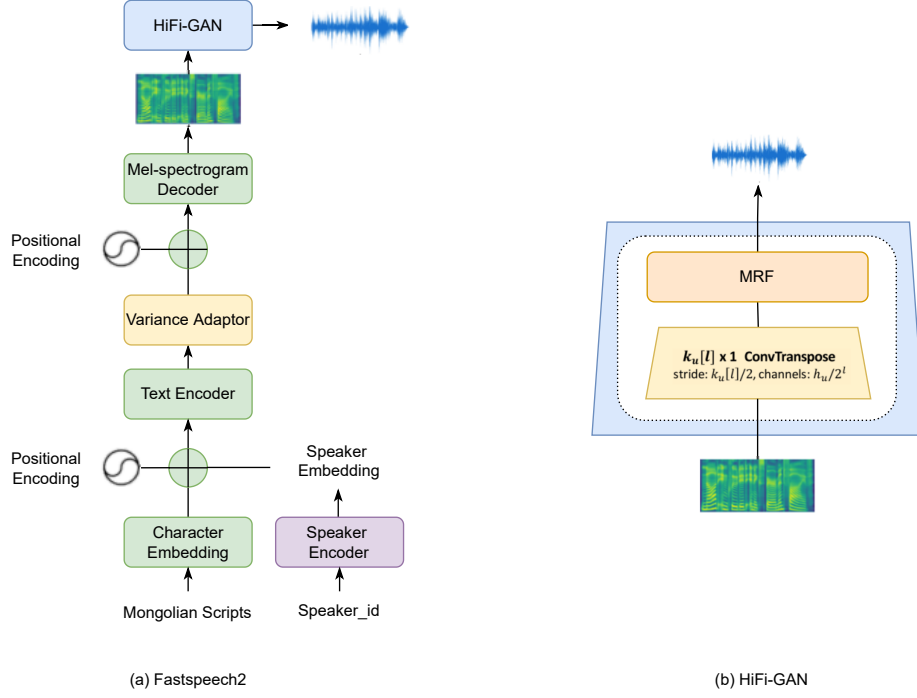
(a) Fastspeech2

(b) HiFi-GAN

Fig. 3: The structure of the FastSpeech2+HiFi-GAN model. We implement the multi-speaker FastSpeech2 by adding the speaker encoder module.

which is responsible for optimizing the upsampling points. HiFi-GAN, as a generative adversarial network, has two kinds of discriminators, including multi-scale and multi-period discriminators. The generagor kernel size of HiFi-GAN is 7 and the upsampling ratio is (8,8,2,2). The list of discriminators for the cycle scale is (2,3,5,7,11). The Conv filters of each periodic discriminator are 8. The pooling type of output downsampling in the melgan discriminator is AveragePooling1D, the kernel size is (5,3), and the activation function is LeakyReLU. HiFi-GAN is trained independently of FastSpeech2. For each speaker, the generator with only stft loss is first trained for 100k steps, and then the generator and discriminator are trained for 100k steps. This gives us the corresponding vocoder for each of the three speakers.

Note that a teacher Tacotron2 model trained for each speaker was used to extract duration from the attention contrast for subsequent FastSpeech2 model training. For each speaker, the Tacotron2 model trained with 100k steps of MnTTS was used to extract the duration. After that, the multi-speaker FastSpeech2 model was trained with 200k steps to do the final speech generation. 100k steps were trained for HiFi-GAN's generator and 100k steps for jointly

training the generator and discriminator. All the above models were trained on 2 Tesla V100 GPUs.

4.2 Naturalness Evaluation

For a full comparison of naturalness, we compared our baseline system, **FastSpeech2+HiFi-GAN** with the **Ground Truth** speech. In addition, to verify the performance of HiFi-GAN, we added a **FastSpeech2+Griffin-Lim** baseline model for further comparison. The Griffin-Lim algorithm can directly obtain the phase information of the audio to reconstruct the waveform without additional training. We used the Naturalness Mean Opinion Score (SS-MOS) [29] to assess naturalness. For each speaker, we randomly select 20 sentences as the evaluation set, which are not used for training. The model-generated audio and the ground truth audio were randomly disrupted and distributed to listeners. During the evaluation process, 10 native Mongolian speakers were asked to evaluate the naturalness of the generated 400 audio speeches in a quiet environment.

The N-MOS results are given in Table 2. Ground truth speech gets the best performance without a doubt. FastSpeech2+HiFi-GAN outperforms the FastSpeech2+Griffin-Lim and achieves much closer performance to the ground truth. Each speaker's N-MOS score was above 4.0 on the combination of FastSpeech2 and HiFi-GAN.

Specifically, for the F1, F2, and F3, the N-MOS of FastSpeech2+HiFi-GAN achieved 4.02, 4.15, and 4.29 respectively, which is encouraging. This proves that high-quality Mongolian speech can be synthesized using MnTTS2 and the proposed model. In a nutshell, all results prove that our MnTTS2 dataset can be used to build a robust TTS system for high-quality speech generation.

Table 2: Naturalness mean opinion score (N-MOS) results for all systems with 95% Confidence intervals.

System / Speaker ID	F1	F2	F3
FastSpeech2+Griffin-Lim	3.56±0.18	3.59±0.04	3.86±0.12
FastSpeech2+HiFi-GAN	**4.02±0.18**	**4.15±0.06**	**4.29±0.11**
Ground Truth	**4.73±0.08**	**4.70±0.14**	**4.68±0.09**

4.3 Speaker Similarity Evaluation

We further conduct listening experiments to evaluate the speaker similarity performance for the FastSpeech+HiFi-GAN baseline system. The Speaker Similarity Mean Opinion Score (SS-MOS) results are reported in Table 3.

We synthesized 20 audios for each speaker by FastSpeech2+HiFi-GAN baseline system. Ten native Mongolian-speaking volunteers were also invited

to participate in the scoring. Each volunteer needs to evaluate whether the speaker is the same person or not in the synthesized and the ground truth audio. The SS-MOS scores for F1, F2, and F3 are 4.58, 4.04, and 4.12 respectively, which is encouraging. The results show that the audio synthesized by the FastSpeech2+HiFi-GAN system performs good performance in terms of speaker similarity. The highest SS-MOS score was obtained for speaker F1. Auditioning the audio, we can find that F1's timbre has significant characteristics and the synthesized audio represents the speaker's voice information well. In a nutshell, this experiment shows that the MnTTS2 dataset can be used for speech synthesis work in multi-speaker scenarios.

Table 3: Speaker Similarity Mean opinion score (SS-MOS) results for FastSpeech2+HiFi-GAN system with 95% Confidence intervals.

System \ Speaker ID	F1	F2	F3
FastSpeech2+HiFi-GAN	4.58±0.21	4.04±0.16	4.12±0.10

5   Challenges and Future Work

With the development of "Empathy AI", the research of emotional TTS has attracted more and more attention [30]. Speech synthesis in conversational scenarios and emotional speech synthesis are hot research topics nowadays [31]. Furthermore, how to control the emotion category and the emotional intensity during speech generation is an interesting direction [32]. However, our MnTTS2 does not involve information related to emotion category and emotion intensity. In future work, we will carry out a more comprehensive and in-depth expansion of the data to serve the development of emotional Mongolian TTS.

6   Conclusion

We presented a large-scale, open-source Mongolian text-to-speech corpus, MnTTS2, which enriches MnTTS with more durations, topics, and speakers. Releasing our corpus under the Knowledge Attribution 4.0 International License, the corpus allows both academic and commercial use. We describe in detail the process of building the corpus, while we validate the usability of the corpus by synthesizing sounds with the FastSpeech2 model and the HiFi-GAN vocoder. The experimental results show that our system can achieve satisfactory performance on the MnTTS2, which indicates that the MnTTS2 corpus is practically usable and can be used to build robust multi-speaker TTS system. In future work, we will introduce emotional TTS dataset to further enrich our corpus. We also plan to compare the effects of different TTS architectures and model hyperparameters on the results and conduct subsequent analyses.

References

1. Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
2. F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2015–2018, 1986.
3. Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
4. Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103, 2014.
5. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech 2017*, pages 4006–4010, 2017.
6. Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 6706–6713, 2019.
7. Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Y Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *ICML*, 2017.
8. Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174, 2019.
9. Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
10. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA, 2016.
11. Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2415–2424. PMLR, 2018.
12. Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville.

Melgan: Generative adversarial networks for conditional waveform synthesis. *CoRR*, abs/1910.06711, 2019.

13. Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

14. Uradyn E Bulag. Mongolian ethnicity and linguistic anxiety in china. *American Anthropologist*, 105(4):753–763, 2003.

15. Yifan Hu, Pengkai Yin, Rui Liu, Feilong Bao, and Guanglai Gao. Mntts: An open-source mongolian text-to-speech synthesis dataset and accompanied baseline. *CoRR*, abs/2209.10848, 2022.

16. Keith Ito and Linda Johnson. The lj speech dataset, 2017.

17. Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *Proc. Interspeech 2019*, pages 1526–1530, 2019.

18. Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines, 2020.

19. Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

20. Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

21. Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

22. Aihong Huang, Feilong Bao, Guanglai Gao, Yu Shan, and Rui Liu. Mongolian emotional speech synthesis based on transfer learning and emotional embedding. In *2021 International Conference on Asian Language Processing (IALP)*, pages 78–83. IEEE, 2021.

23. Rui Liu, Feilong Bao, Guanglai Gao, and Weihua Wang. Mongolian prosodic phrase prediction using suffix segmentation. In *2016 International Conference on Asian Language Processing (IALP)*, pages 250–253. IEEE, 2016.

24. Rui Liu, Feilong Bao, Guanglai Gao, and Yonghe Wang. Mongolian text-to-speech system based on deep neural network. In *National conference on man-machine speech communication*, pages 99–108. Springer, 2017.

25. Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang. Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model. In *Interspeech*, pages 57–61, 2018.

26. Tiankai Zhi, Ying Shi, Wenqiang Du, Guanyu Li, and Dong Wang. M2ASR-MONGO: A free mongolian speech database and accompanied baselines. In *24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2021, Singapore, November 18-20, 2021*, pages 140–145. IEEE, 2021.

27. Feilong Bao, Guanglai Gao, Xueliang Yan, and Weihua Wang. Segmentation-based mongolian lvcsr approach. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8136–8139, 2013.

28. Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281, 2017.

29. Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

30. Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Decoding knowledge transfer for neural text-to-speech training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1789–1802, 2022.

31. Rui Liu, Berrak Sisman, and Haizhou Li. Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 4648–4652. ISCA, 2021.
32. Rui Liu, Haolin Zuo, De Hu, Guanglai Gao, and Haizhou Li. Explicit intensity control for accented text-to-speech, 2022.