applied sciences

MDPI

Article

Comparative Study for Multi-Speaker Mongolian TTS with a New Corpus

Kailin Liang, Bin Liu, Yifan Hu, Rui Liu *, Feilong Bao and Guanglai Gao

College of Computer Science, Inner Mongolia University, Hohhot 010031, China
* Correspondence: liurui_imu@163.com

Abstract: Low-resource text-to-speech synthesis is a very promising research direction. Mongolian is the official language of the Inner Mongolia Autonomous Region and is spoken by more than 10 million people worldwide. Mongolian, as a representative low-resource language, has a relative lack of open-source datasets for its TTS. Therefore, we make public an open-source multi-speaker Mongolian TTS dataset, named MnTTS2, for related researchers. In this work, we invited three Mongolian announcers to record topic-rich speeches. Each announcer recorded 10 h of Mongolian speech, and the whole dataset was 30 h in total. In addition, we built two baseline systems based on state-of-the-art neural architectures, including a multi-speaker Fastspeech 2 model with HiFi-GAN vocoder and a full end-to-end VITS model for multi-speakers. On the system of FastSpeech2+HiFi-GAN, the three speakers scored 4.0 or higher on both naturalness evaluation and speaker similarity. In addition, the three speakers achieved scores of 4.5 or higher on the VITS model for naturalness evaluation and speaker similarity scores. The experimental results show that the published MnTTS2 dataset can be used to build robust Mongolian multi-speaker TTS models.

Keywords: Mongolian; text-to-speech (TTS); open-source dataset; multi-speakerkeyword

Citation: Liang, K.; Liu, B.; Hu, Y.; Liu, R.; Bao, F.; Gao, G. Comparative Study for Multi-Speaker Mongolian TTS with a New Corpus. *Appl. Sci.* **2023**, *13*, 4237. https://doi.org/10.3390/app13074237

Academic Editors: Ya Li, Kai Yu and Yan Song

Received: 2 March 2023
Revised: 21 March 2023
Accepted: 23 March 2023
Published: 27 March 2023

1. Introduction

Text-to-Speech (TTS) aims to convert the input text into human-like speech [1]. As a standard technology in human–computer interaction, it is commonly used in car navigation, intelligent voice audio, cell phone voice assistant, etc. Compared with traditional speech synthesis methods based on cascading [2] and statistical modeling [3], neural end-to-end TTS models have shown superior performance. This is attributed to the encoder–decoder architecture [4]. Typical models include Tacotron [5], Tacotron2 [1], Transformer TTS [6], Deep Voice [7], etc. Later, in order to improve the disadvantage of the slow inference speed of autoregressive models, non-autoregressive TTS models [8], such as FastSpeech [9], FastSpeech2(s) [10], etc. were proposed and became the mainstream approaches for TTS. Note that armed with the neural network-based vocoder, including WaveNet [11], WaveRNN [12], MelGAN [13], HiFi-GAN [14], etc., the TTS model can synthesize speech sounds that are comparable to human sounds.

We note that the large-scale corpus resources are an essential factor in the rapid development of neural TTS mentioned above. This is especially true for languages, such as English and Mandarin, which are widely spoken worldwide. Mongolian is the official language of the Inner Mongolia Autonomous Region of the People's Republic of China and is also widely spoken in other surrounding provinces and countries. However, low-resource languages such as Mongolian [15] have been making slow progress in related research due to the difficulties in corpus collection. Currently, Mongolian lacks a low-resource language and lacks technical staff and native speakers of Mongolian to annotate and produce high-quality datasets. The lack of publicly available high-quality datasets seriously affects the development of Mongolian speech technology. Therefore, building a large-scale and high-quality Mongolian TTS dataset is necessary. In addition, our lab has previously open-sourced a single-speaker dataset called MnTTS [16], 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 dataset called MnTTS2, which extends to three speakers and increases the data duration to 10 h for each speaker. The textual content has also been further expanded and enriched in the domain. Similar to our MnTTS, the MnTTS2 dataset is freely available to academics and industry practitioners.

To demonstrate the reliability of MnTTS2, we built two baseline models. First, we combined the state-of-the-art FastSpeech2 [10] model with the HiFi-GAN [14] vocoder to build a model. Second, we built a fully end-to-end TTS system based on the VITS [17] model. After that, we conducted listening experiments and reported the results of three evaluations. These include the Naturalness Mean Opinion Score (N-MOS) for the naturalness of speech, the Speaker Similarity Mean Opinion Score (SS-MOS) for speaker similarity, and a robustness analysis to detect model stability. 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 a robust multispeaker TTS system.

Our main contributions are as follows. (1) We developed a multi-speaker TTS dataset, termed MnTTS2, containing three speakers. The total audio duration is about 30 h. The transcribed text covers various domains, such as sports and culture, etc. (2) We use the state-of-the-art non-autoregressive FastSpeech2 and the fully end-to-end VITS [17] to build the baseline models and validate our MnTTS2. (3) The MnTTS2 dataset, source code, and pre-trained models will be publicly available to academics and industry practitioners.

We also highlight some differences between this work and our conference version [18]: (1) We add a new baseline model based on the latest and most powerful fully end-to-end VITS system and report the training details; (2) we conduct more comparative study with the additional baseline; and (3) we further conduct robustness evaluation on the basis of naturalness and speaker similarity experiments to validate our corpus and model more comprehensively.

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 concludes and summarizes the work.

2. Related Work

For mainstream languages, such as English and Mandarin, there are many free and publicly available TTS datasets. For example, LJspeech [19] is a single-speaker dataset for English. To enrich the speaker diversity, some multi-speaker TTS datasets have been released, for example Libritts [20] and VCTK [21] with different accents for English and AiShell [22] for Chinese.

For low-resource languages, such as Mongolian, the available resources are pretty limited. We note that there are some methods for learning by generating additional annotated data that can improve TTS synthesis with low-resource data, such as data augmentation [23,24] and cross-linguistic mapping [25]. Other approaches, such as unsupervised learning [26], semi-supervised learning [27], and transfer learning [28], have also achieved good results.

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 [29]. 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 [30]. Immediately after that, Rui Liu proposed a DNN-based Mongolian speech synthesis system, which performs better than the traditional HMM [31]. In addition,

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 [32]. Unfortunately, none of the Mongolian TTS datasets from the above works have been released publicly and are not directly available to the public. We also found that some datasets in related fields, such as M2ASR-MONGO [33] for Mongolian speech recognition, have been made public recently. However, the speech recognition corpus cannot be applied in the TTS field due to environmental noise and improper speaking style issues, etc.

We previously released the single-speaker MnTTS dataset [16], called MnTTS. The total duration of the MnTTS is 8 h, and it was recorded in a studio by a professional female native Mongolian announcer. However, the duration and speaker diversity still need to be further expanded. In a nutshell, it is necessary to construct a high-quality multi-speaker Mongolian TTS dataset to further promote Mongolian TTS research, which is the focus of this paper. We will introduce the details of the MnTTS2 in the following subsection.

3. MnTTS2 Dataset

In this section, we first briefly revisit the MnTTS dataset 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 [16]. 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 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 (http://mglip.com/challenge/NCMMSC2022-MTTSC/index.html, accessed on on 13 October 2022).

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.

3.2.1. Text Collection and Narration

Similar to MnTTS [16], the first step in building the MnTTS2 dataset was to collect a large amount of transcription. The natural idea for collecting such text materials is to search for text information from websites and electronic books. The most basic requirements for our text are that: (1) the content of the text should cover a variety of scenarios that people use every day, (2) the subject matter should be as rich as possible, and (3) the content should not involve inappropriate content, such as those involving sensitive political issues, religious issues, or pornography. These are the most basic requirements for our texts. Following this, we searched 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. These contents were removed in the hope that our dataset could make a positive contribution to the development of the Mongolian language, which was the original intention of our work.

3.2.2. Text Preprocessing

Compared to mainstream languages, such as Mandarin and English, traditional Mongolian exhibits agglutinative characteristics. This makes Mongolian letters express different styles in different contexts and brings a serious harmonic phenomenon [16]. In order to solve this problem, we transformed the texts into a Latin alphabet, instead of a traditional Mongolian representation, for TTS training. The entire pipeline of converting Mongolian texts into Latin sequences was divided into three steps: encoding correction, Latin conversion, and text regularization. A detailed description can be found in our previous work MnTTS [16].

3.2.3. Audio Recording and Audio–Text Alignment

Different from the MnTTS [16], we invited three native Mongolian-speaking announcers to record the audio. Each announcer volunteered to participate and signed an informed consent form to agree to 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. All recordings were made in a standard recording studio at Inner Mongolia University. We chose Adobe Audition (https://www.adobe.com/cn/products/audition.html, accessed on 16 November 2021) as the recording software.

The collected and processed texts were divided into three parts without duplicate content. The three announcers completed the recording according to their respective texts. During the recording process, we asked the announcer to keep a 0.3 s pause at the beginning and end of each audio segment, keep a constant distance between the lips and the microphone, perform a slight pause at the comma position, and perform 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 were 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 marks (‘?’), exclamation marks (‘!’), etc. Finally, we obtained about 30 h of speech data, which were sampled at 22.05 kHz with a sampling accuracy of 16bit.

3.2.4. Corpus Structure and Statistics

The file structure of the MnTTS2 corpus is shown in Figure 1. Each speaker’s recording file and the corresponding text collection were saved in a folder named after the speaker. All audio was stored in WAV format files with a sampling rate of 22.05 kHz and a sampling accuracy of 16 bits. All text was 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 Figure 2. As shown in Table 1, the entire corpus has a total of 23,801 sentences. For example, F1, with a total of 572,016 Mongolian characters, has an average of 79 characters per sentence, 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 Figure 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 is concentrated in 4–5 s. In comparison, we found that the word numbers of sentences for F2 were not particularly concentrated, and the duration was 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.

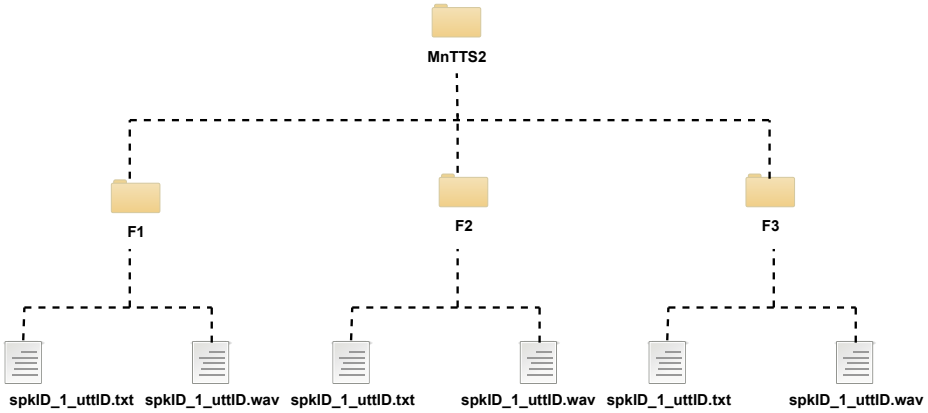

Figure 1. The folder structure of the MnTTS2 corpus.

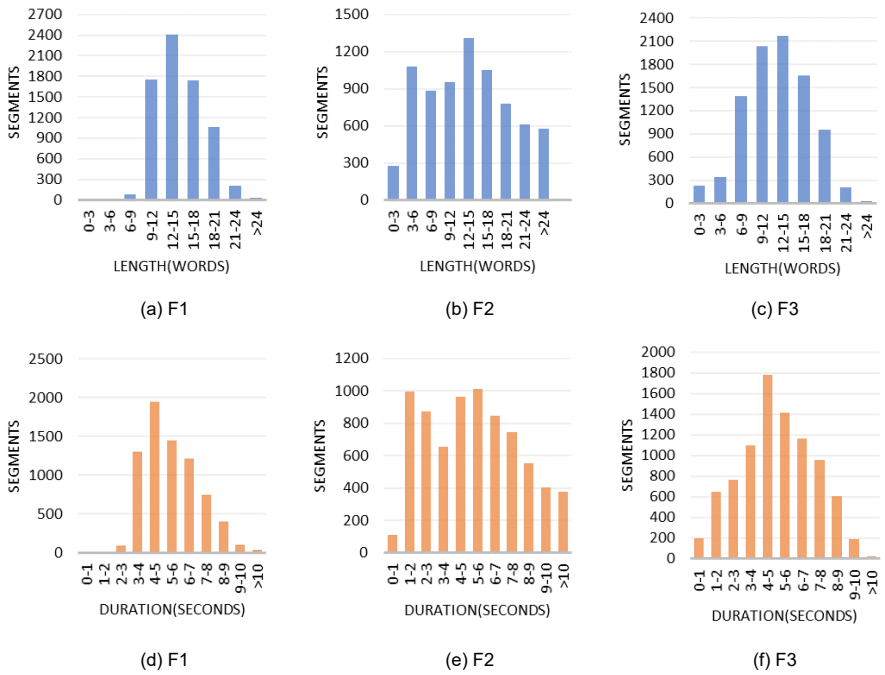

Figure 2. Word number distributions (**a**–**c**) and sentence duration distributions (**d**–**f**) for all speakers of MnTTS2.

Table 1. The statistics of the MnTTS2 dataset.

Statistical Unit \ Speaker ID		F1	F2	F3
Character	Total	572,016	459,213	601,366
	Mean	79	61	67
	Min	12	2	2
	Max	189	188	190
Word	Total	88,209	71,245	92,719
	Mean	12	9	10
	Min	3	1	1
	Max	29	30	29

4. Speech Synthesis Experiments

To verify the validity of our MnTTS2 dataset, we conducted TTS experiments in Mongolian using FastSpeech2 and VITS [17], respectively, used the Mean Opinion Score

(MOS) to measure the naturalness of the synthesized speech and the speaker similarity, and finally verified the usability of the dataset by robustness analysis experiments.

4.1. Experimental Setup

There were two models for TTS experiments on MnTTS2: first, the FastSpeech2 model and the HiFi-GAN vocoder as the whole TTS system; and second, the fully end-to-end VITS [17] model.

4.1.1. Experimental Steps of TTS Based on the FastSpeech2 Model and HiFi-GAN Vocoder

We used the TensorFlowTTS toolkit (https://github.com/TensorSpeech/TensorFlowTTS, accessed on 11 November 2022) to build an end-to-end TTS model based on the FastSpeech2 model. As shown in Figure 3, the FastSpeech2 model converts the input Mongolian text into Mel-spectrogram features, and as shown in Figure 4, the HiFi-GAN vocoder reconstructs the waveform from the Mel-spectrogram features. We added a speaker encoder layer to FastSpeech2 to implement a multi-speaker TTS system as a way to match our multi-speaker dataset. The speaker embedding layer inputs the speaker id as an integer code and initializes them with random weights for learning speaker embeddings. The embedding layer does not need to be trained separately and can be directly embedded in the Fastspeech2 model to learn together. The speaker encoder consists of a speaker embedding layer, a dense layer, and a softplus layer. For the network model structure, we set the number of speakers to 3, the dimension of speaker embedding to 384, and the hidden layer to 4. In the decoder, the hidden layer size was 384, and the number of hidden layers was 4.

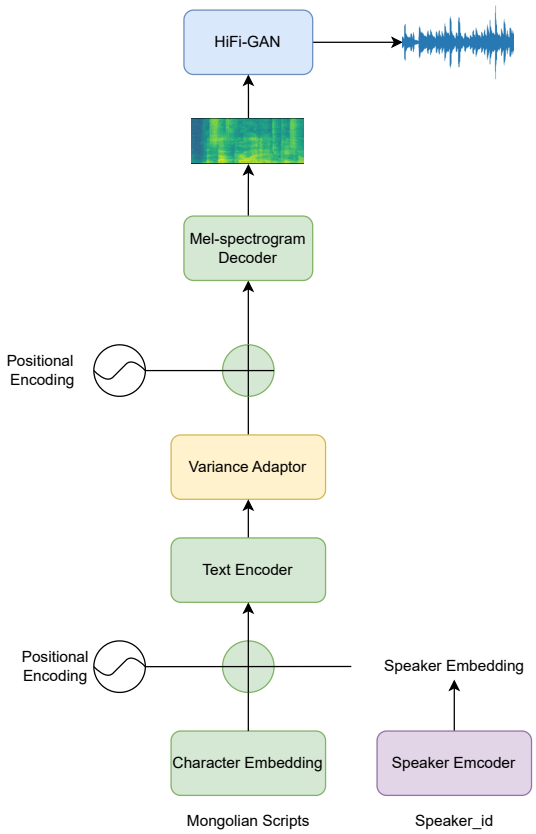

Figure 3. The structure of the FastSpeech2+HiFi-GAN model. We implemented the multi-speaker FastSpeech2 by adding the speaker encoder module.

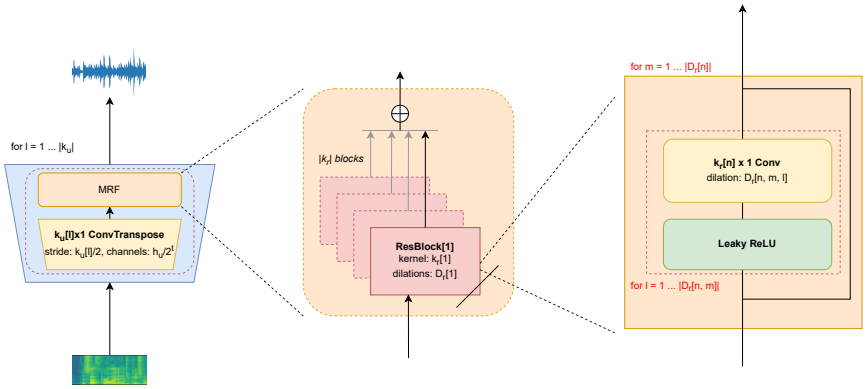

Figure 4. The structure of the HiFi-GAN model.

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

Note that the teacher Tacotron2 model with 80 k steps for each speaker was trained for extracting durations from attentional contrasts for subsequent FastSpeech2 model training. After that, a multi-speaker FastSpeech2 model was trained in 200 k steps to complete the final speech generation. The generator of HiFi-GAN was trained for 100 k steps, and the generator and discriminator were jointly trained for 100 k steps. The above models were trained on 2 Tesla V100 GPUs.

4.1.2. Experimental Steps of TTS Based on the VITS Model

VITS is a generation-based, parallel, fully end-to-end TTS system. We built a fully end-to-end TTS model using code publicly available from the VITS authors. Since the model uses conditional VAE as the generative model and uses normalized flow to enhance the expressiveness of the prior and posterior distributions, it can produce speech that is closer to the real person. The model can synthesize the waveform of natural speech directly from the text. Therefore, there is no need to obtain waveforms from the generated Mel spectrum. The model links two modules, acoustic and vocoder, through VAE, learning hidden variables to achieve a full end-to-end effect. VITS has the features of easy training, fast synthesis, good stability of long text, and rich speech diversity. This is the main reason why we chose VITS. As shown in Figure 5, In the text encoder, the embedding dimension of the text was 192, the number of heads in the attention module was 2, the kernel size was 3, and the dropout was 0.1. To implement a multi-speaker TTS model, the speaker encoder module was added to encode the speaker information, the number of speakers was 3, and the speaker embedding was 256 dimensions. The decoder was the same as HiFi-GAN V1. In the generator, the kernel size was 7 and the upsampling ratio was (8, 8, 2, 2). In the discriminator, the output channel was 1, the kernel size was 5, and the list of discriminators for the cyclic scale was (2, 3, 5, 7, 11).

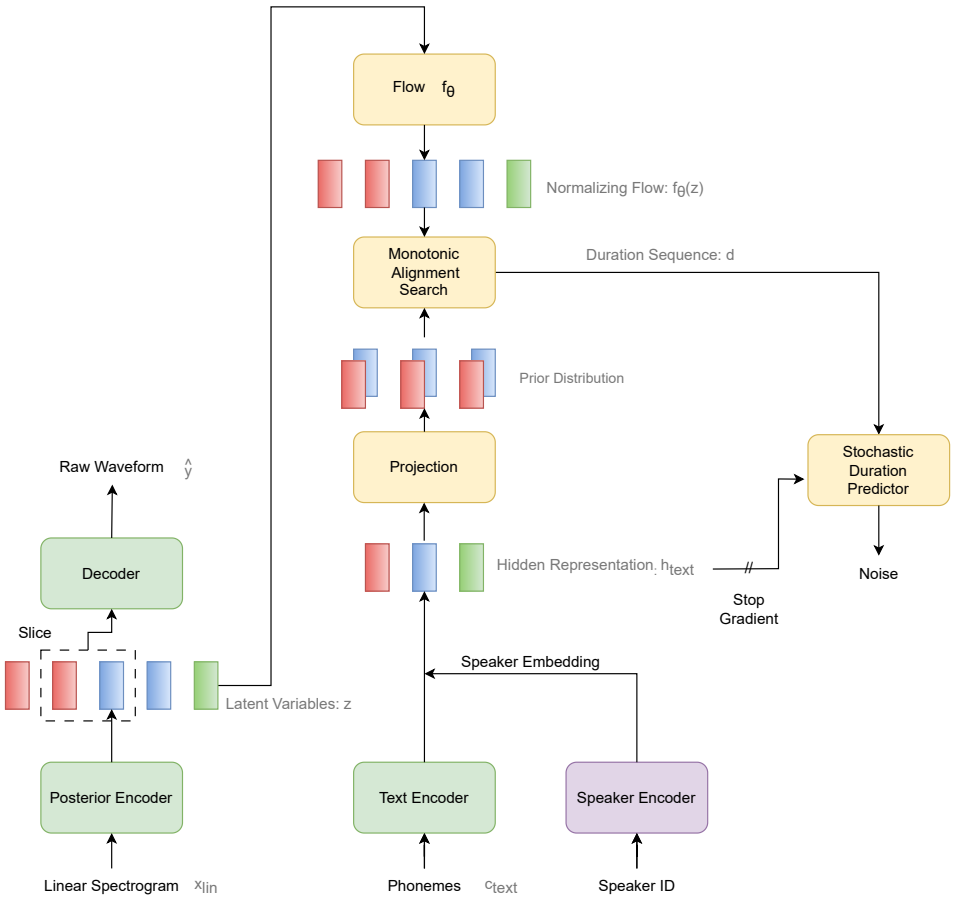

Figure 5. The structure of the VITS model. We implemented the multi-speaker VITS by adding the speaker encoder module.

Since the Mongolian text represented by Latin sequence does not require text processing, the data of each speaker in the dataset was manually divided into a training set, a test set, and a validation set in the ratio of 93:5:2. The entire VITS model was trained on two Tesla P100 GPUs.

4.2. Naturalness Evaluation

To fully compare the naturalness, we compared the baseline systems FastSpeech2 + HiFi-GAN, VITS with ground truth speech. In addition, to verify the performance of the neural vocoder, we added the FastSpeech2 + Griffin-Lim baseline model for further comparison. The Griffin-Lim algorithm can directly obtain the phase information of audio to reconstruct the waveform without additional training. For each speaker, we randomly selected 20 sentences as the evaluation set, which were not used for training. The model-generated and real audio were randomly disrupted and assigned to listeners. During the evaluation process, 10 native Mongolian speakers were asked to evaluate the naturalness of the generated 600 audible speeches in a quiet environment. We report the naturalness results with naturalness mean opinion score (N-MOS) [34].

The main results for N-MOS are given in Table 2. Undoubtedly, the best performance was obtained for real speech. VITS achieved scores extremely close to real speech and had the best results among the two baseline models. FastSpeech2+HiFi-GAN outperformed FastSpeech2+Griffin-Lim and also achieved good results. Specifically, VITS achieved an N-MOS score of 4.62 on F3, which is a small difference compared to the true value. Meanwhile, FastSpeech2+HiFi-GAN also achieved the best score of 4.29 on F3, but there is still a significant difference compared to VITS.

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	$\mathbf{4.02 \pm 0.18}$	$\mathbf{4.15 \pm 0.06}$	$\mathbf{4.29 \pm 0.11}$
VITS	$\mathbf{4.60 \pm 0.07}$	$\mathbf{4.55 \pm 0.08}$	$\mathbf{4.62 \pm 0.09}$
Ground Truth	$\mathbf{4.73 \pm 0.08}$	$\mathbf{4.70 \pm 0.14}$	$\mathbf{4.68 \pm 0.09}$

For F1 and F2, VITS scored 4.60 and 4.55, respectively, while FastSpeech2 + HiFi-GAN achieved 4.02 and 4.15, respectively, which is encouraging. Overall, the VITS model still works better than FastSpeech2 + HiFi-GAN. In short, all our results demonstrate that the MnTTS2 dataset can be used to build a robust TTS system to generate high-quality speech.

4.3. Speaker Similarity Evaluation

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

Table 3. Speaker Similarity Mean Opinion Score (SS-MOS) results for the 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
VITS	4.56 ± 0.09	4.67 ± 0.08	4.54 ± 0.07

We synthesized 20 audios for each speaker with the FastSpeech2+HiFi-GAN and VITS baseline systems, respectively. Ten native Mongolian-speaking volunteers were also invited to participate in the scoring. Each volunteer was asked to assess whether the speaker was the same person in the synthesized audio and the ground truth audio.

In the FastSpeech2+HiFi-GAN baseline system, the SS-MOS scores for F1, F2, and F3 were 4.58, 4.04, and 4.12, respectively, with large differences across speakers. In contrast, the scores of the VITS model were more consistently concentrated, with F1, F2, and F3 achieving scores of 4.56, 4.67, and 4.54, respectively, which is encouraging. The results show that the audio synthesized by the FastSpeech2+HiFi-GAN system and VITS system performs well in terms of speaker similarity. Note that FastSpeech2 model achieved the highest SS-MOS score for speaker F1. One possible reason is that the F1 dataset contains a large number of sentences with fewer words instead of long sentences. Another reason is that the HiFi-GAN vocoder of F2 and F3 did not work well during the FastSpeech2+ HiFi-GAN synthesis experiment. Due to time and funding constraints, we first recruited volunteers to perform SS-MOS scoring on F1. Volunteers were re-recruited to update the data for F2 and F3 after subsequent training of the F2 and F3 vocoders was completed. The scoring of F1 and the division of F2 and F3 into two parts led to a greater subjectivity of the volunteers so that F1 scored higher, while F2 and F3 scored lower, and both scored similarly. We found that VITS can effectively synthesize high-quality speech in the face of longer utterance due to its generative and fully end-to-end structure. In conclusion, this experiment shows that the MnTTS2 dataset can be used for speech synthesis work in multi-speaker scenarios.

4.4. Robustness Analysis

All baseline models achieved good results in both naturalness and speaker similarity assessment. However, the end-to-end TTS model often suffers from the robustness issue, such as repeating, skipping words, etc. In this section, we further conduct robustness analysis to check their performance.

We analyzed the synthesized speech in terms of six classes: (1) repetitive words, words that repeat speech in the sentence; (2) skipped words, words that were missing from the sentence, and skipped directly over words that were not spoken; (3) mispronunciation words, words that were spoken but mispronounced in the sentence; (4) incomplete words, words that are pronounced correctly in a sentence but are incomplete; (5) long pauses, an unusually long pause in a sentence; and (6) non-verbal sounds, other sounds in the sentence that are not verbal. The length of the text to be synthesized was set at 13 words, as either too long or too short sentences would not accurately reflect the synthesis. We synthesized 50 speech items for each speaker from the test set, and randomly selected 20 of them for robustness analysis. Finally, we invited five native Mongolian volunteers to evaluate the synthesized speech for six errors.

The final statistical results are shown in Table 4. For FastSpeech2+HiFi-GAN, the speech samples of F1, F2, and F3 produced six, nine, and three errors in total, respectively. The VITS model also produced 2, 10, and 3 errors for F1, F2, and F3. The difference between the two models was not significant for F2 and F3, except for F1 where the number of errors was significantly lower than that of FastSpeech2+HiFi-GAN. It is worth noting that the speech samples of F2 produce more errors than F1 and F3 for both models. The main reason may be due to the data distribution again. For the F2 dataset, from Figure 2b we can find that the length distribution was not concentrated. There were more sentences with improper length, that were too short and too long, than the other two speakers. Therefore, it will lead to long speech samples, such as 13 words, in the test set of F2, making it easier to produce errors. In conclusion, the robustness analysis shows that our MnTTS2 can be used to build robust TTS systems to some extent.

Table 4. Error types found in the 50-sentence test set (Total number of words is 500).

System	Error Types	F1	F2	F3
FastSpeech2+HiFi-GAN	Repeated words	3	1	1
	Skipped words	0	0	0
	Mispronounced words	1	2	1
	Incomplete words	1	2	1
	Long pauses	0	2	0
	Nonverbal sounds	1	2	0
	Total	6	9	3
VITS	Repeated words	1	2	1
	Skipped words	0	0	0
	Mispronounced words	1	2	1
	Incomplete words	0	1	1
	Long pauses	0	3	0
	Nonverbal sounds	0	2	0
	Total	2	10	3

5. Conclusions

We present a large-scale, open-source Mongolian text-to-speech corpus, MnTTS2, which enriches MnTTS with more duration, topics, and speakers. Releasing our corpus under a Knowledge Attribution 4.0 international license, the corpus allows both academic and commercial use. We described the process of building the corpus and validated the usability of the corpus using the FastSpeech2 with HiFi-GAN vocoder and VITS as the baseline models in detail. The experimental results show that our system can achieve satisfactory performance on MnTTS2, indicating that the MnTTS2 corpus is practically

usable and can be used to build robust multi-lingual TTS systems. In future work, we will introduce sentiment TTS datasets 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.

Author Contributions: Project administration: F.B. and G.G.; supervision: R.L.; writing—original draft preparation: R.L., K.L. and Y.H.; dataset collection and organization: R.L., F.B. and Y.H.; baseline model building and training: K.L. and Y.H.; baseline model evaluation: B.L.; model structure visualization: K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the High-level Talents Introduction Project of Inner Mongolia University (No. 10000-22311201/002) and the Young Scientists Fund of the National Natural Science Foundation of China (No. 62206136).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in the paper is now publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
2. Charpentier, F.; Stella, M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86), Tokyo, Japan, 7–11 April 1986; Volume 11, pp. 2015–2018.
3. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
4. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
5. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010.
6. Li, N.; Liu, S.; Liu, Y.; Zhao, S.; Liu, M. Neural speech synthesis with transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33, pp. 6706–6713.
7. Arık, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. Deep voice: Real-time neural text-to-speech. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 195–204.
8. Gu, J.; Bradbury, J.; Xiong, C.; Li, V.O.; Socher, R. Non-Autoregressive Neural Machine Translation. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
9. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech: Fast, robust and controllable text to speech. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
10. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
11. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016; p. 125.
12. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient neural audio synthesis. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 25–31 July 2018; pp. 2410–2419.
13. Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; Courville, A.C. Melgan: Generative adversarial networks for conditional waveform synthesis. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
14. Kong, J.; Kim, J.; Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17022–17033.
15. Bulag, U.E. Mongolian ethnicity and linguistic anxiety in China. *Am. Anthropol.* **2003**, *105*, 753–763. [CrossRef]
16. Hu, Y.; Yin, P.; Liu, R.; Bao, F.; Gao, G. MnTTS: An Open-Source Mongolian Text-to-Speech Synthesis Dataset and Accompanied Baseline. In Proceedings of the 2022 International Conference on Asian Language Processing (IALP), Singapore, 27–28 October 2022; pp. 184–189. [CrossRef]

17. Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 5530–5540.
18. Liang, K.; Liu, B.; Hu, Y.; Liu, R.; Bao, F.; Gao, G. MnTTS2: An Open-Source Multi-Speaker Mongolian Text-to-Speech Synthesis Dataset. *arXiv* **2022**, arXiv:2301.00657.
19. Ito, K.; Johnson, L. The LJ SPEECH dataset. 2017. Available online: https://keithito.com/LJ-Speech-Dataset (accessed on 1 March 2023).
20. Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R.J.; Jia, Y.; Chen, Z.; Wu, Y. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv* **2019**, arXiv:1904.02882.
21. Veaux, C.; Yamagishi, J.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit*; The Centre for Speech Technology Research (CSTR), University of Edinburgh: Edinburgh, UK, 2017.
22. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M. AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines. *arXiv* **2020**, arXiv:2010.11567.
23. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
24. Sheng, P.; Yang, Z.; Hu, H.; Tan, T.; Qian, Y. Data augmentation using conditional generative adversarial networks for robust speech recognition. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 26–29 November 2018; pp. 121–125.
25. Zhou, X.; Tian, X.; Lee, G.; Das, R.K.; Li, H. End-to-end code-switching tts with cross-lingual language model. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7614–7618.
26. Barlow, H.B. Unsupervised learning. *Neural Comput.* **1989**, *1*, 295–311. [CrossRef]
27. Zhu, X.J. *Semi-Supervised Learning Literature Survey*; Department of Computer Sciences, University of Wisconsin-Madison: Madison, WI, USA, 2005.
28. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
29. Huang, A.; Bao, F.; Gao, G.; Shan, Y.; Liu, R. Mongolian emotional speech synthesis based on transfer learning and emotional embedding. In Proceedings of the 2021 International Conference on Asian Language Processing (IALP), Yantai, China, 23–25 October 2021; pp. 78–83.
30. Liu, R.; Bao, F.; Gao, G.; Wang, W. Mongolian prosodic phrase prediction using suffix segmentation. In Proceedings of the 2016 International Conference on Asian Language Processing (IALP), Tainan, Taiwan, 21–23 November 2016; pp. 250–253.
31. Liu, R.; Bao, F.; Gao, G.; Wang, Y. Mongolian text-to-speech system based on deep neural network. In Proceedings of the National Conference on Man-Machine Speech Communication, Lianyungang, China, 11–13 October 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 99–108.
32. Liu, R.; Bao, F.; Gao, G.; Zhang, H.; Wang, Y. Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 57–61.
33. Zhi, T.; Shi, Y.; Du, W.; Li, G.; Wang, D. M2ASR-MONGO: A Free Mongolian Speech Database and Accompanied Baselines. In Proceedings of the 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA 2021), Singapore, 18–20 November 2021; pp. 140–145. [CrossRef]
34. Streijl, R.C.; Winkler, S.; Hands, D.S. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimed. Syst.* **2016**, *22*, 213–227. [CrossRef]