Mongolian Text-to-Speech Challenge under Low-Resource Scenario for NCMMSC2022

Rui Liu[1], Zhen-Hua Ling[2], Yi-Fan Hu[1], Hui Zhang[1], and Guang-Lai Gao[1]

[1]Inner Mongolia University, Hohhot, China
[2]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R. China
liurui_imu@163.com, zhling@ustc.edu.cn, hyfwalker@163.com, {cszh, csggl}@imu.edu.cn

Abstract. Mongolian Text-to-Speech (TTS) Challenge under Low-Resource Scenario is a special session for National Conference on Man-Machine Speech Communication 2022 (NCMMSC2022), termed as NCMMSC2022-MTTSC. A Mongolian TTS dataset was provided to participants this year and a low-resource Mongolian TTS task was designed. Specifically, the task is to synthesize high-quality Mongolian speech with given Mongolian scripts. Thirteen teams submitted their results for final evaluation. Mean opinion score (MOS) listening tests were conducted online to measure the naturalness, intelligibility of the synthetic speech. In addition, the word error rate (WER) of automatic speech recognition was further treated as the objective metric for intelligibility evaluation. The evaluation results show that the top system achieved comparable naturalness and intelligibility with the ground truth speech.

Keywords: Mongolian · Text-to-Speech (TTS) · Low-Resource · NCMMSC2022.

1 Introduction

Text-to-Speech (TTS), that is a standard technology in human-computer interaction, aims to convert the input text to human-like speech [1]. Mongolian TTS Challenge under Low-Resource Scenario is a special session for National Conference on Man-Machine Speech Communication 2022 (NCMMSC2022), termed as NCMMSC2022-MTTSC. The NCMMSC2022-MTTSC was organized by the Inner Mongolia University, the University of Science and Technology of China (USTC) and the other members of the committee [1]. The majority of previous TTS challenges have used speech datasets for mainstream languages, such as Mandarin Chinese and English. For example, Blizzard Challenges [2] 2008-2010 and 2019-2020 adopt Mandarin Chinese data and Blizzard Challenges 2005-2013 and 2016-2018 use English data. Note that the TTS research for the minority language gradually attracted wide attention. To this end, Blizzard

[1] http://mglip.com/challenge/NCMMSC2022-MTTSC/index.html
[2] http://festvox.org/blizzard/index.html

Challenges 2013-2015 use several India languages and 2021 adopt European Spanish data as the official training corpus. Despite the progress, there are also some languages that have not attracted enough attention, such as Mongolian language [2].

Mongolian is the most famous and widely spoken language of the Mongolian language family. In addition, Mongolian is the main national language in the Inner Mongolia Autonomous Region of China and is mainly used in Mongolian-inhabited areas of China, Mongolia, and the Siberian Federal District of the Russian Federation. Recently, the researchers began a series study of Mongolian TTS [3]. Specifically, deep learning techniques were first introduced to Mongolian TTS [4], and DNN-based acoustic models trained on 5-hours training data were used instead of HMM acoustic models to achieve high-quality synthesized speech. Li et al. [5] further implemented a Tacotron-based Mongolian TTS system. To accelerate the inference speed and improve the speech fidelity, Liu et al. [6] proposed a pure non-autoregressive neural Mongolian TTS model, called MonTTS, which consists of the FastSpeech2-based acoustic model and the HiFi-GAN vocoder. The above mentioned works provide a solid foundation for the research of Mongolian TTS technology. However, there is still a lack of in-depth research on Mongolian TTS, especially in low-resource scenario that aims to synthesize high-quality Mongolian speech with limited training data.

The NCMMSC2022-MTTSC is the first time that a minority language, i.e., Mongolian, has been used for TTS challenge in China. This challenge also promotes the development of intelligent information processing in minority languages within China. This paper will present the details of the speech dataset, tasks, participating systems, evaluations and results of challenge.

2 Voices to build

2.1 Speech dataset

A Mongolian speech dataset kindly provided by Inner Mongolia University was released for voice building. The dataset contains recorded speech from a professional female native Mongolian speaker together with text transcriptions. The texts were from various domains, including daily life, sport, education, travelling, etc. The speech was recorded in a studio of school of computer science with quiet environment. The total duration of the waveform files, which were sampled at 44.1 kHz with a sampling accuracy of 16bit, amounts to around 2 hours.

2.2 Task

We design a Mongolian TTS task under a low-resource scenario using the released dataset. Each participating team should build a voice from the provided 2-hours Mongolian data to synthesize the given Mongolian text, following the challenge rules [1]. The submitted synthetic speech should be 16 bit depth, and at any

standard sampling rate (e.g., 16 kHz, 22.05 kHz, 44.1 kHz, or 48 kHz). For evaluation, teams were required to synthesize 200 test sentences (disjoint from the training data) that contained only Mongolian text.

Regarding the use of external data, the NCMMSC2022-MTTSC allowed that each participant can use any data of any language, whether freely-available or not, to conduct the pre-training etc. Participants were asked to report their data usage instruction when submitting synthetic speech and in their paper.

Table 1: The participating teams and their institutions. The system identifier of natural speech (the first row) is letter A. The method descriptions are summarised based on the questionnaires and the workshop papers from participants.

Team Name	**Institution**	**Input Type**	**Acoustic Model**	**Vocoder**	**Transfer Learning**
Natural Speech	N/A	N/A	N/A	N/A	N/A
all u need	University of Science and Technology of China, Hefei	Latin	FullConv	Griffin-Lim	No
Mnemosyne	Microsoft Azure Speech, Beijing	Latin	Conformer based FastSpeech2	HiFinet2	Yes
sigma	VXI Global Solutions, Shanghai	Latin	VITS	N/A	Yes
TJUCCA_TTS	Tianjin University, Tianjin **(no paper submission)**	Phoneme	Tacotron2	HiFi-GAN	Yes
RoyalFlush	Zhejiang Hithink RoyalFlush AI Research Institute, HangZhou	Latin	Tacotron2	HiFi-GAN	Yes
火之源	Inner Mongolia University, Hohhot	Latin	FastSpeech2	HiFi-GAN	Yes
IOA-THINKIT	Institute of Acoustics, and University of Chinese Academy of Sciences, Beijing **(no paper submission)**	Latin	VITS	N/A	Yes
DBLAB	OPPO	Latin	VITS	N/A	Yes
在线_AI特工队	China Mobile Online Services Co., Ltd., Luoyang	Latin	VITS	N/A	No
qdreamer	Suzhou Qimeng People Network Technology Company, Suzhou	Latin	VITS	N/A	Yes
FlySpeech	Audio, Speech and Langauge Processing Group, Northwestern Polytechnical University, XiAn	Phoneme	Delightful TTS	HiFi-GAN	Yes
Cyber	Chengdu Rongwei Software Service Co., Ltd. Chengdu **(no paper submission)**	Phoneme	Flow-based model	HiFi-GAN	Yes
Y	Mobvoi **(no paper submission)**	Phoneme	VITS	N/A	No

3 Participants

There are 13 teams submitted their results. Note that there is no benchmark system for NCMMSC2022-MTTSC. Following Blizzard challenges, all systems are identified using letters in these published results. Specifically, letter A denotes natural or ground truth speech. Letters B to T were assigned to the systems submitted by participants. Each participating team is free to choose whether to reveal their system identifier in their workshop paper.

We summarized the detailed structure of all systems in Table 1. We see that all systems adopted a neural approach, and the great majority employed VITS, which is a state-of-the-art fully end-to-end model. The classic Tacotron2 and FastSpeech2 models were also favored by many teams. Just one team build the acoustic model with convolutional neural network (CNN). Neural vocoder was also adopted by many teams, of which the majority (6 out of 13) was HiFi-GAN.

4 Evaluations and Results

4.1 Evaluation Materials

We released 200 sentences as the testing data for the listening test. All participants used their own system to synthesize 200 sentences for the final subjective and objective rating.

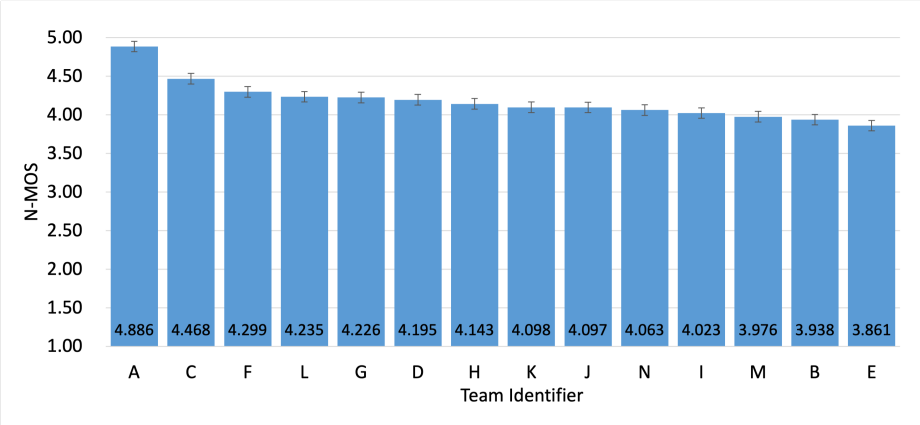
4.2 Evaluation Metrics

The evaluation results for NCMMSC2022-MTTSC consisted of three metrics in terms of naturalness and intelligibility respectively, as follows:

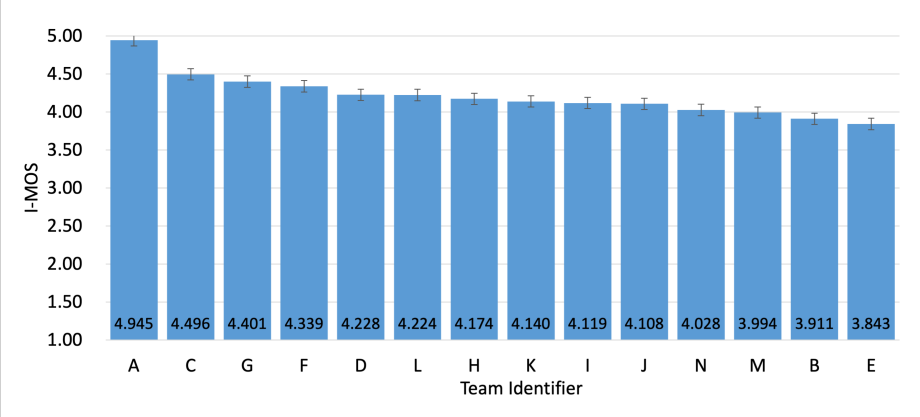
- **Naturalness mean opinion score (N-MOS)** was used to test the speech quality for all teams in terms of naturalness subjectively.
- **Intelligibility mean opinion score (I-MOS)** was used to test the speech quality for all teams in terms of intelligibility subjectively.
- **Word error rate (WER)** was used to test the speech quality for all teams in terms of intelligibility objectively.

For N-MOS and I-MOS, the organizers recruited 20 listeners that were all native speakers of Mongolian and all instructions and other text on the listening test webpages were includes Chinese and Mongolian. For WER, the organizers calculated the WER by leveraging a Mongolian speech recognition interface [3] from Inner Mongolia University.

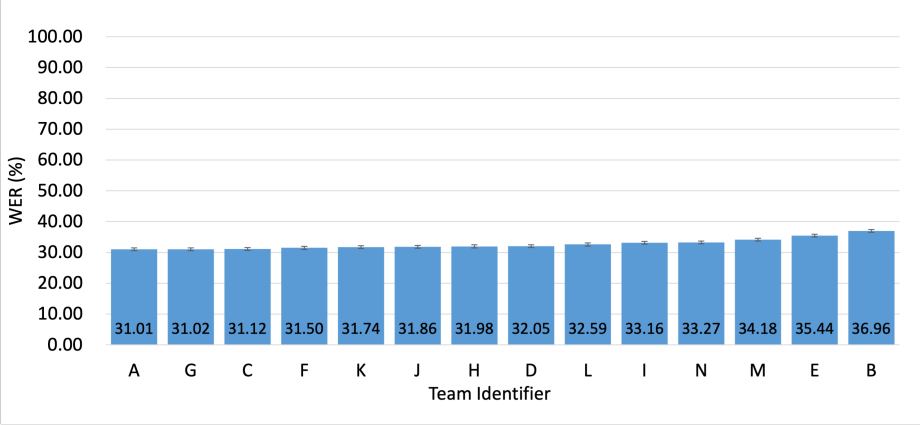
[3] `http://asr.mglip.com/`



(a) N-MOS Results



(b) I-MOS Results



(c) WER Results

Fig. 1: N-MOS, I-MOS and WER results for NCMMSC2022-MTTSC. A is natural speech, the remaining letters denote the systems submitted by participants.

4.3 Results

We report the evaluation results for all metrics at Fig.1. In all subfigures of Fig.1, a consistent system ordering is adopted, which is the descending order of mean values for the corresponding metric. The mean values is calculated from the listeners' scores for each metric. Please note that this ordering only aims to make the plots more readable by using the same system ordering across all plots for each task and can not be interpreted as a ranking, because the ordering does not indicate which systems are significantly better than others.

As shown in Fig.1(a), system C achieved significantly better naturalness than all other submitted systems. The N-MOS ratings of natural speech (system A) and system C were 4.886 and 4.468 respectively, and the difference between them is minimal compared to other systems. All other systems scored below 4.3.

As shown in Fig.1(b), system C still beat other systems and achieve highest I-MOS score. The I-MOS scores of system A and system C were 4.945 and 4.496 respectively.

Regarding objective intelligibility metric, Fig. 1(c) reports the WER. We found that the system G achieved lowest value with 31.02, which is closer to the system A with 31.01 than other systems.

4.4 Acknowledgements

This work was partially supported 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). We wish to thank a number of additional contributors without whom running the challenge would not be possible. Prof. Zhenhua Ling at University of Science and Technology of China provide all instruction related to the challenge selflessly. Yifan Hu at Inner Mongolia University helped to prepare the official challenge website and the training and test data. Hui Zhang at Inner Mongolia University helped to conduct the WER calculation. Pengkai Yin at Inner Mongolia University is responsible for gathering listening test volunteers. Thanks to all participants and listeners.

References

1. Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32(3):35–52, 2015.
2. Juha Janhunen. Mongolic languages. In *The encyclopedia of language & linguistics*, pages 231–234. Elsevier Scientific Publ. Co, 2006.
3. Rui Liu, Berrak Sisman, Feilong Bao, Jichen Yang, Guanglai Gao, and Haizhou Li. Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:274–285, 2020.

4. 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.
5. Jingdong Li, Hui Zhang, Rui Liu, Xueliang Zhang, and Feilong Bao. End-to-end mongolian text-to-speech system. In *2018 11th international symposium on chinese spoken language processing (ISCSLP)*, pages 483–487. IEEE, 2018.
6. Rui Liu, Shiyin Kang, Guanglai Gao, Jingdong Li, and Feilong. MonTTS: A real-time and high-fidelity mongolian tts model with pure non-autoregressive mechanism. *Journal of Chinese Information Processing*, 36(7):86, 2022.