TatarTTS: An Open-Source Text-to-Speech Synthesis Dataset for the Tatar Language

Daniil Orel
Institute of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
daniil.orel@nu.edu.kz

Askat Kuzdeuov
Institute of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
askat.kuzdeuov@nu.edu.kz

Rinat Gilmullin
Institute of Applied Semiotics
Tatarstan Academy of Sciences
Kazan, Russia
rinatgilmullin@gmail.com

Bulat Khakimov
Institute of Applied Semiotics
Tatarstan Academy of Sciences
Kazan, Russia
khakeem@yandex.ru

Huseyin Atakan Varol
Institute of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
ahvarol@nu.edu.kz

***Abstract*—This paper introduces an open-source dataset for speech synthesis in the Tatar language. The dataset comprises approximately 70 hours of transcribed audio recordings, featuring two professional speakers (one male and one female). Notably, it is the first large-scale dataset of its kind that is publicly available, aimed at promoting Tatar text-to-speech (TTS) applications in both academic and industrial contexts. The paper describes the procedures for developing the dataset, discusses the challenges faced, and outlines important future directions. To demonstrate the reliability of the dataset, baseline end-to-end TTS models were built and evaluated using the subjective mean opinion score (MOS) measure. The dataset, training recipe, and pre-trained TTS models are publicly available.**

***Index Terms*—Text-to-speech, speech synthesis, low-resource languages, Turkic languages**

I. Introduction

The recent advances in natural language processing (NLP) were achieved due to developments in computational power and accumulation of large amounts of linguistic data [1], which allowed the application of deep neural networks (DNNs) to solve NLP problems. One of the NLP tasks, which has benefited from these changes is text to speech (TTS), also known as speech synthesis. TTS is the transformation of a written text to its audio representation [2]. This technology is of particular value as an assistive technology since it helps to build more inclusive solutions for visually impaired people [3], e.g., giving them easy access to digital resources. Voice assistants in smart devices use TTS as well [4]. For instance, Alexa, Siri, Alisa, and other assistants can interact with users naturally. In addition, TTS is closely related to several other NLP technologies, such as voice cloning [5], speech-to-text translation, and speech-to-speech translation [6].

The increasing amount of data is one of the main drivers for improvements in TTS systems. There are multiple open-source datasets for both mono and multilingual TTS. Such datasets can be constructed based on automatic speech recognition (ASR) data, like LibriTTS [7] or LibriTTS-R [8], a new version of LibriTTS with improved sound quality and 585 hours of transcribed speech from more than 2,400 speakers. Alternatively, the data for TTS datasets can be obtained by recording narration of different texts, as done for LJ-Speech [9]. The major problem for TTS systems is the lack of datasets for low-resourced languages.

Tatar language has more than 5 million speakers [10] worldwide and its usage in digital world is increasing[1]. However, the amount of data for training NLP models for Tatar is still lacking. We aim to tackle this issue in two steps, first by introducing the TatarTTS Corpus, a novel dataset for TTS in the Tatar language and then training TTS models using TatarTTS. The primary contributions of this work include:

- Introduction of a new TTS dataset for the Tatar language.
- Training TTS models for male and female speakers on the TatarTTS dataset.
- Open-sourcing the dataset, source code, and pre-trained TTS models at GitHub[2].

The rest of the paper is organized in the following way: In Section II, the reviews of work in TTS and NLP for low-resourced languages is given; In Section III, the description of Tatar Speech corpus is given; In Section IV, we describe the experimental setup. In Section V we provide analysis of results, obtained in this work and the Section VI concludes the work.

[1] https://www.ethnologue.com/language/tat/
[2] https://github.com/IS2AI/TatarTTS

II. Related Works

A. NLP for Turkic Languages

The Turkic language family includes a wide range of languages such as Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, Uzbek, among others, and are spoken over a vast geographic region [11]. Substantial endeavors have been undertaken to improve NLP resources for these languages. Particularly, Turkish and Kazakh languages have emerged as frontrunners in regard to accessible data and NLP instruments. These languages offer named entity recognition datasets and models [12], [13], along with ASR tools and datasets [14], [15], in addition to other resources.

There have been extensive research aimed at developing unified solutions for all the aforementioned languages. A multilingual approach often proves advantageous for enhancing model performance. For instance, research on ASR across Turkic languages [14] indicates that combined training using all languages improves the resulting model compared to individual monolingual models. Furthermore, Turkic languages find representation in multilingual GPT [16], created through multilingual pre-training of GPT-3. Additionally, major neural machine translation systems, such as Google's [17], also offer support for Turkic languages.

B. NLP for Tatar Language

Tatar is also a Turkic language, and in recent times, it has emerged as a prominent focus in NLP research. Several studies have aimed to compile corpora for classical NLP tasks for this language, delving into tasks like estimating text similarity, analogies, and text relatedness [18]; there are also tools for named entity recognition in Tatar, which show state-of-the-art (SOTA) results [19]. Notably, Tatar has found its place in extensive collections of NLP resources designed for Turkic languages. For instance, it is integrated into the Turkic ASR - a multilingual model catering to speech recognition across 10 Turkic languages [14].

Furthermore, there exist multiple neural machine translation tools facilitating translation to and from Tatar. These tools encompass solutions developed through collaborations between large companies and researchers, such as No Language Left Behind [20], alongside endeavors by smaller teams [21]. Notably, state-of-the-art models are now available for translations involving the Tatar language [22].

Additionally, efforts have been made to curate corpora for Tatar TTS systems. One such initiative aimed at gathering high-quality audio data involves recordings by 2 male and 1 female speakers[3]. However, the limitation of this solution is its relatively small dataset, comprising less than 8 hours of speech. Other attempts to build Tatar TTS models involve transliterating Tatar texts into Kazakh, benefiting from existing TTS models [23]. While this approach exhibits promising results, the generated speech currently lacks intelligibility, rendering the resulting sentences challenging to comprehend [23].

TABLE I
Tatar TTS corpus statistics

Category	M	F
# Segments	20,508	18,274
# Tokens	201,606	168,614
# Unique tokens	37,684	30,329
Duration	36.2 h	33.9 h

III. Dataset Construction

A. Tatar Speech Corpus

The textual content for the dataset was sourced from the Tatar language corpus[4]. This corpus encompasses a wide range of literary genres such as fiction, media texts, official documents, educational literature, and scientific publications. It is important to note that the texts selected for the dataset were carefully chosen to ensure they were free of grammatical errors. This level of quality control guarantees that the dataset contains accurate and linguistically correct content.

The selection process of the speakers for narration was conducted with careful consideration. Speakers were chosen from Tatar National Theatre actors who demonstrated fluency in the Tatar language and possessed experience in narrating TV and radio programs. This ensured that the chosen speakers had the necessary language proficiency and performance skills. To guarantee high-quality audio, a range of professional equipment, including Neumann microphones and the Steinberg Cubase digital audio workstation, was utilized for the recordings, and they were conducted in a controlled studio to minimize external noise interference. The recordings were sampled at a frequency of 44.1 kHz with bit depth of 32 bits, ensuring a sufficient level of detail in the audio.

After the audio collection phase, skilled transcribers were enlisted to manually segment the recordings into sentence-level chunks. This meticulous process involved precise identification and separation of individual sentences within the recorded audio data.

B. Dataset Specifications

The TatarTTS dataset comprises speech recordings from two speakers, one male and one female speaker. The statistics of the dataset is shown in Table I. In total, the dataset contains around 70 hours of audio consisting of over 38,000 segments. There is nearly equal amount of data for both speakers.

The organization of the TatarTTS dataset is structured in the following manner. The data for the two professional speakers is stored separately in two distinct folders. Each folder contains one CSV file and one sub-folder. The CSV file has N rows and 2 columns, where N is the number of audio recordings. The first column represents audio file name, while the second one stores the corresponding text. The sub-folder contains the audio recordings in the WAV format.

[3] https://github.com/egorsmkv/qirimtatar-tts-datasets

[4] https://tugantel.tatar/

IV. TTS Experiments

A. Experimental Setup

In order to prove the efficacy of our dataset, we utilized end-to-end variational inference text-to-speech (VITS) models [24]. These models employ conditional variational autoencoders that incorporate normalizing flows and an adversarial training strategy. The VITS method also incorporates a stochastic duration predictor, enabling the synthesis of speech from input text. This approach allows for adaptability to different speech pitches and rhythms, enhancing the overall flexibility of the models.

For training and preparing the TTS models for inference, we utilized the Piper framework[5]. This framework offers a straightforward and convenient approach to train models and prepare them for inference using the ONNX format. Piper offers VITS models in four distinct quality tiers: x-low, low, medium, and high. This tiered approach allows users to choose the appropriate model based on their specific quality requirements. The tiers range from the lowest quality tier (x-low) to the highest quality tier (high), providing flexibility in selecting the desired level of audio fidelity and speech synthesis performance. We employed a high quality TTS models with 83M trainable parameters.

As part of the preprocessing phase, the text files in the dataset underwent a transformation to include only 39 Cyrillic letters and five specific symbols: ’.’, ’,’, ’-’, ’?’, and ’!’. This simplification of the character set ensures consistency and compatibility for further processing and modeling tasks. By limiting the characters to these designated symbols, the dataset becomes more focused and tailored to the specific language and speech synthesis objectives.

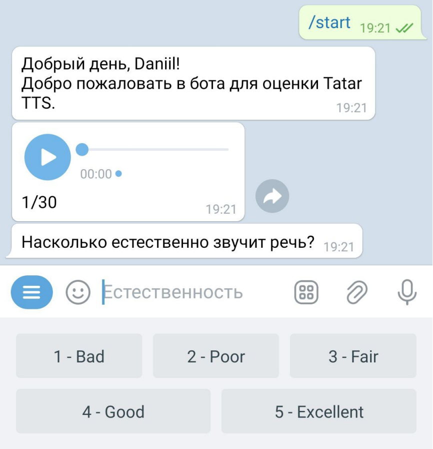
We trained separate TTS models for each speaker, creating single-speaker models. Each model was trained on A100 graphics processing units (GPUs) on an NVIDIA DGX server.

B. TTS Model Training

Given the substantial volume of data within the TatarTTS dataset, exceeding 30 hours per speaker, we adopted a straightforward approach in training the TTS models. Our methodology involved the random initialization of VITS models for both female and male speakers. Subsequently, these models underwent training on the given dataset for 1,000 epochs per speaker. This comprehensive training process aimed to harness the extensive available data for optimal model development. The details of training can be found the GitHub repository.

The Pytorch checkpoint (PTH) of each TTS model has a size of 998MB. After conversion to the ONNX format, the size significantly reduced to 113.8MB. The conversion process from PTH to ONNX resulted in a considerable decrease in file size, making the models more compact and efficient for deployment and inference purposes.

[5] https://github.com/rhasspy/piper



/start 19:21

Добрый день, Daniil!
Добро пожаловать в бота для оценки Tatar TTS. 19:21

00:00
1/30 19:21

Насколько естественно звучит речь? 19:21

Fig. 1. The interface of the Telegram messenger bot used for evaluation. **Translation:** ***Good afternoon, Daniil! Welcome to the bot for Tatar TTS evaluation. Please evaluate the naturalness of the speech ...***

C. Model Evaluation

We evaluated the models using the Mean Opinion Score (MOS) metric, which is widely used within the TTS domain. This assessment involved a survey conducted with native speakers to gauge the models performance based on different criteria. In an effort to minimize bias, we randomly selected a set of 200 utterances narrated by speakers. These utterances formed our “testing set,” deliberately excluded from the model training process. Following the completion of training, we utilized the TTS models to generate voice outputs for these utterances, thereby providing both machine-generated and human narrations for comparative analysis.

To conduct the evaluation process, we developed a Telegram bot, as shown in Fig. 1, and shared the bot’s link with native Tatar speakers. This bot facilitated the assessment of 30 recordings per session, focusing on qualitative criteria such as audio quality, pronunciation, naturalness, and overall comprehensibility. Participants were requested to rate each criterion on a scale from 1 (bad) to 5 (excellent). They were allowed to listen to the recordings multiple times but were restricted to submitting a single grade for each criterion. It is worth noting that the user interface of the bot was in Russian. This decision was influenced by the prevalent Tatar-Russian bilingualism observed within the target audience of the bot, as highlighted in the work of Wigglesworth-Baker [25].

To ensure balance and eliminate listener bias, we maintained consistency while presenting the audios by providing five audios per source. This was accomplished by incorporating two TTS models each for male (M) and female (F) speakers, along with real recordings. Altogether, there were four distinct sources: real recordings by M and F speakers; synthesized recordings by M and F models. It was ensured that within the same evaluation batch, narrations of the same text from different sources were not included, thus diminishing potential listener bias.

TABLE II

AGGREGATED STATISTICS BY SPEAKERS. FOR CRITERIA THE SCORE IS REPRESENTED AS MEAN ± STANDARD ERROR FOR NATURALNESS (N), PRONUNCIATION (P), COHERENCE (C), AND OVERALL QUALITY (Q).

Speaker	**N**	**P**	**C**	**Q**
M_S	4.85 ± 0.04	4.66 ± 0.06	4.61 ± 0.06	4.54 ± 0.08
M_O	4.84 ± 0.04	4.81 ± 0.05	4.79 ± 0.05	4.76 ± 0.05
F_S	4.80 ± 0.05	4.66 ± 0.05	4.69 ± 0.06	4.65 ± 0.06
F_O	4.89 ± 0.03	4.87 ± 0.03	4.92 ± 0.02	4.96 ± 0.02

V. RESULTS & DISCUSSION

When each audio received at least one rating, we gathered statistics on them, as presented in Table II. In the Speaker column, the initial letter signifies the speaker's gender, while the section after "_" denotes specific parameters. "S" refers to synthesized, and "O" signifies the original audio.

Analyzing Table II, it's evident that in case of naturalness of speech, both male and female models fall within confidence interval of original audios, which means that our models produce speech, which sounds natural.

Both M and F original audios achieved scores above 4.70, indicating high-quality content within the dataset. When considering pronunciation, coherence, and overall quality, the models lag behind the originals by approximately 0.2 points. Nevertheless, this slight difference still positions them as high-quality TTS models with MOS score above 4.5.

VI. CONCLUSION

This study presents the TatarTTS corpus, which consists of 70 hours of transcribed speech from two speakers (one male and one female). Furthermore, we have created preliminary TTS models as a baseline for the Tatar language. Although these models are not without their imperfections and have some limitations in terms of coherence and clear pronunciation, they demonstrate notable naturalness and sound quality. Despite the need for further refinement, these initial TTS models represent a significant step forward in the development of Tatar language synthesis. Our research findings reveal that even with a relatively limited amount of training data, specifically 30 hours of speech data, reasonably good performance in terms of MOS can be achieved for TTS models.

As we move forward, our future plans involve enhancing the TatarTTS dataset by expanding the number of audio samples for the existing speakers and introducing new speakers to further diversify the dataset. Additionally, we acknowledge the increasing prevalence of loanwords in natural speech, which is a consequence of globalization. To address this linguistic aspect, future versions of TatarTTS will incorporate sentences that include loanwords. This expansion aims to strengthen our models and make them more robust in handling the challenges posed by globalization, ensuring accurate and natural synthesis of Tatar speech.

VII. ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to the speakers and volunteers who generously devoted their time and effort to evaluating the audio recordings. We would also like to extend special recognition and appreciation to Dias Bakhtiyarov for his significant assistance in the development of the Telegram bot.

REFERENCES

[1] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2021.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021.

[3] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky *et al.*, "Deep voice: Real-time neural text-to-speech," 2017.

[4] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end TTS for voice agent," 2020.

[5] G. Ruggiero, E. Zovato, L. D. Caro, and V. Pollet, "Voice cloning: a multi-speaker text-to-speech synthesis approach based on transfer learning," 2021.

[6] C. Xu, R. Ye, Q. Dong, C. Zhao, T. Ko, M. Wang, T. Xiao, and J. Zhu, "Recent advances in direct speech-to-text translation," 2023.

[7] 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," 2019.

[8] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe *et al.*, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," 2023.

[9] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[10] H. Boeschoten, "The speakers of turkic languages," in *The Turkic Languages*. Routledge, 2021, pp. 1–12.

[11] L. Johanson, *The Turkic Language Family*, ser. Cambridge Language Surveys. Cambridge University Press, 2021, p. 16–40.

[12] E. Okur, H. Demir, and A. Özgür, "Named entity recognition on Twitter for Turkish using semi-supervised learning with word embeddings," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, May 2016, pp. 549–555. [Online]. Available: https://aclanthology.org/L16-1087

[13] R. Yeshpanov, Y. Khassanov, and H. A. Varol, "KazNERD: Kazakh named entity recognition dataset," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 417–426. [Online]. Available: https://aclanthology.org/2022.lrec-1.44

[14] S. Mussakhojayeva, K. Dauletbek, R. Yeshpanov, and H. A. Varol, "Multilingual speech recognition for turkic languages," *Information*, vol. 14, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/2/74

[15] S. Mussakhojayeva, Y. Khassanov, and A. Varol, "KSC2: An industrial-scale open-source kazakh speech corpus," in *Proc. Interspeech*, 2022, pp. 1367–1371.

[16] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova, and T. Shavrina, "mgpt: Few-shot learners go multilingual," 2022. [Online]. Available: https://arxiv.org/abs/2204.07580

[17] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[18] A. Khusainova, A. Khan, and A. R. Rivera, "SART - similarity, analogies, and relatedness for Tatar language: New benchmark datasets for word embeddings evaluation," in *Springer-Verlag*, Berlin, Heidelberg, 2023, p. 380–390.

[19] O. Nevzorova, D. Mukhamedshin, and A. Galieva, "Named entity recognition in tatar: Corpus based algorithm," in *Proceedings of Computational Models in Language and Speech Workshop (CMLS)*, ser. CEUR Workshop Proceedings, A. Elizarov and N. Loukachevitch, Eds., vol. 2303, 2018, pp. 58–68. [Online]. Available: http://ceur-ws.org/Vol-2303/paper4.pdf

[20] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad *et al.*, "No language left behind: Scaling human-centered machine translation," 2022.

[21] A. Valeev, I. Gibadullin, A. Khusainova, and A. Khan, "Application of low-resource machine translation techniques to Russian-Tatar language pair," 2019.

[22] A. Khusainov and D. Suleymanov, “Multilingual neural machine translation system for 7 turkic-russian language pairs,” in *Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on “Integrated Models and Soft Computing in Artificial Intelligence (IMSC-2021)*, 2021.

[23] R. Yeshpanov, S. Mussakhojayeva, and Y. Khassanov, “Multilingual Text-to-Speech Synthesis for Turkic Languages Using Transliteration,” in *Proc. Interspeech*, 2023, pp. 5521–5525.

[24] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, Jul 2021, pp. 5530–5540. [Online]. Available: https://proceedings.mlr.press/v139/kim21f.html

[25] T. Wigglesworth-Baker, “Language policy and post-soviet identities in tatarstan,” *Nationalities Papers*, vol. 44, no. 1, p. 20–37, 2016.