KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset

Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov and Huseyin Atakan Varol

Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Nur-Sultan, Kazakhstan

{saida.mussakhojayeva, aigerim.janaliyeva, almas.mirzakhmetov, yerbolat.khassanov, ahvarol}@nu.edu.com

Abstract

This paper introduces a high-quality open-source speech synthesis dataset for Kazakh, a low-resource language spoken by over 13 million people worldwide. The dataset consists of about 93 hours of transcribed audio recordings spoken by two professional speakers (female and male). It is the first publicly available large-scale dataset developed to promote Kazakh text-to-speech (TTS) applications in both academia and industry. In this paper, we share our experience by describing the dataset development procedures and faced challenges, and discuss important future directions. To demonstrate the reliability of our dataset, we built baseline end-to-end TTS models and evaluated them using the subjective mean opinion score (MOS) measure. Evaluation results show that the best TTS models trained on our dataset achieve MOS above 4 for both speakers, which makes them applicable for practical use. The dataset, training recipe, and pretrained TTS models are freely available.

Index Terms: speech corpus, speech synthesis, low resource, Kazakh, end-to-end, TTS, Transformer, Tacotron

1. Introduction

Text-to-speech (TTS) systems are essential in many applications, such as navigation, announcement, smart assistants, and other speech-enabled devices. For any language, it is also necessary to ensure accessibility for the visually-impaired, and availability of human-machine interaction without requiring visual and tactile interfaces [1]. To build a robust TTS system, a sufficiently large and high-quality speech dataset is required. In order to address this, we developed a large-scale open-source speech dataset for the Kazakh language. We named our dataset KazakhTTS, and it is primarily geared to build TTS systems.

Kazakh is the official language of Kazakhstan, and it is spoken by over 13 million people worldwide, including countries, such as China and Russia. It is an agglutinative language with vowel harmony belonging to the family of Turkic languages. Kazakh is a low-resource language and is considered to be endangered due to multiple factors, primarily the dominance of the Russian and English languages in education and administration [2]. Therefore, there is a growing awareness of the importance of increasing the number of Kazakh speakers reflected in many language rescue initiatives launched by the government.

Currently, there is no Kazakh speech dataset of sufficient size and quality for building TTS systems, especially recently proposed end-to-end (E2E) neural architectures [3, 4, 5, 6, 7, 8, 9]. This work aims to fill this gap by introducing the KazakhTTS dataset. To the best of our knowledge, it is the first large-scale open-source dataset developed for building Kazakh TTS systems. Our dataset contains around 93 hours of high-quality speech data read by two professional speakers (36 hours by female and 57 hours by male). The dataset was carefully annotated by native transcribers and covers most of the daily-use Kazakh words. The KazakhTTS is freely available for both academic and commercial use under the Creative Commons Attribution 4.0 International License.

With the help of KazakhTTS, we plan to promote Kazakh language use in speech-based digital technologies and to advance research in Kazakh speech processing. We believe that our dataset will be a valuable resource for the TTS research community, and our experience will benefit other researchers planning to develop speech datasets for low-resource languages. Although the primary application domain of KazakhTTS is speech synthesis, it can also be used to aid other related applications, such as automatic speech recognition (ASR) and speech-to-speech translation.

To demonstrate the reliability of KazakhTTS, we built two baseline Kazakh E2E-TTS systems based on Tacotron 2 [7] and Transformer [9] architectures for each speaker. We evaluated these systems using the subjective mean opinion score (MOS) measure. The experiment results showed that the best TTS models built using our dataset achieve 4.5 and 4.1 in MOS for the female and male speakers, respectively, which assures the usability for practical applications. In addition, we performed a manual analysis of synthesized sentences to identify the most frequent error types. The dataset, reproducible recipe, and pretrained models are publicly available.

The rest of the paper is organized as follows. Section 2 briefly reviews related works for TTS dataset construction. Section 3 describes the KazakhTTS construction procedures and reports the dataset specifications. Section 4 explains the TTS experimental setup and discusses obtained results. Section 5 concludes the paper and highlights future research directions.

2. Related works

The recent surge of speech-enabled applications, such as virtual assistants for smart devices, has attracted substantial attention from both academia and industry to the TTS research [3, 4, 5]. Consequently, many large-scale datasets have been collected [10, 11, 12, 13], and challenges have been organized to systematically compare different TTS technologies [14]. However, these datasets and competitions are mostly restricted to resource-rich languages, such as English, Mandarin, and so on.

To create a speech corpus suitable for developing TTS systems in low-resource settings, methods based on unsuper-
An alternative approach is to record and manually annotate audio recordings. This is considered costly, since cumbersome manual work is required. Nevertheless, the produced dataset will be more reliable. Recently, several Kazakh speech corpora have been developed to accelerate speech processing research in this language. For example, the first open-source Kazakh speech corpus (KSC) containing over 332 hours of transcribed audio recordings was presented in [18]. However, the KSC was mainly constructed for ASR applications, and thus, crowdsourced from different speakers, with various background noises and speech disfluencies kept to make it similar to the real-world scenarios. As such, the quality of audio recordings in the KSC is insufficient for building robust TTS models. Additionally, in the KSC, the size of recordings contributed by a single speaker is small. Similarly, the other existing Kazakh speech datasets are either unsuitable or publicly unavailable [19, 20].

3. Dataset construction

The KazakhTTS project was conducted with the approval of the Institutional Research Ethics Committee of Nazarbayev University. Each speaker participated voluntarily and was informed of the data collection and use protocols through a consent form.

3.1. Text collection and narration

We started the dataset construction process with textual data collection. In particular, we manually extracted articles in chronological order from news websites to diversify the topic coverage (e.g., politics, business, sports, entertainment and so on) and eliminate defects peculiar to web crawlers. The extracted articles were manually filtered to exclude inappropriate content (e.g., sensitive political issues, user privacy, and violence) and stored in DOC format convenient for professional speakers (i.e., font size, line spacing, and typeface were adjustable to the preferences of the speakers). In total, we collected over 2,000 articles of different lengths.

To narrate the collected articles, we first conducted an audition among multiple candidates from which one female and one male professional speaker were chosen. The speaker details, including age, working experience as a narrator, and recording device information, are provided in Table 1. Due to the COVID-19 pandemic, we could not invite the speakers to our laboratory for data collection. Therefore, the speakers were allowed to record audio at their own studio at home or office. The speakers were instructed to read texts in a quiet indoor environment at their own natural pace and style. Additionally, they were asked to follow orthoepic rules, pause on commas, and use appropriate intonations for sentences ending with a question mark. The female speaker recorded audios at her office studio, the recordings were sampled at 44.1 kHz and stored using 16 bit/sample. The male speaker recorded audios at home studio, the recordings were sampled at 48 kHz and stored using 24 bit/sample.

In total, female and male speakers read around 1,000 and 1,250 articles respectively, out of which around 300 articles are overlapping.

3.2. Segmentation and audio-to-text alignment

We hired five native Kazakh transcribers to segment the recorded audio files into sentences and align them with text using the Praat toolkit [21]. The texts were represented using the Cyrillic alphabet consisting of 42 letters. In addition to letters, the transcripts also contain punctuation marks, such as period (‘.’), comma (‘,’), hyphen (‘-‘), question mark (‘?’), exclamation mark (‘!’), and so on. The transcribers were instructed to remove segments with incorrect pronunciation and background noise, trim long pauses at the beginning and end of the segments, and convert numbers and special characters (e.g., ‘%’, ‘$’, ‘+’, and so on) into the written forms. To ensure uniform quality of work among the transcribers, we assigned a linguist to randomly check the tasks completed by transcribers and organize regular “go through errors” sessions.

To guarantee high quality, we inspected the segmented sentences using our ASR system trained on the KSC dataset [18]. Specifically, we used the ASR system to transcribe the segments. The recognized transcripts were then compared with the corresponding manually annotated transcripts. The segments with a high character error rate (CER) were regarded as incorrectly transcribed, and thus, rechecked by the linguist. Lastly, we filtered out all segments containing international words written using a non-Cyrillic alphabet, because speakers usually don’t know how to correctly pronounce such words.

3.3. Dataset specifications

The overall statistics of the constructed dataset are provided in Table 2. In total, the dataset contains around 93 hours of audio consisting of over 42,000 segments. The distribution of segment lengths and durations are shown in Figure 1. The whole dataset creation process took around five months, and the uncompressed dataset size is around 15 GB.

The KazakhTTS dataset is organized as follows. The sources of two professional speakers are stored in two separate folders. Each folder contains a single metadata file and two subfolders containing audio recordings and their transcripts. The audio and corresponding transcript filenames are the same, except that the audio recordings are stored as WAV files, whereas the transcripts are stored as TXT files using the UTF-8 encoding. The naming convention for both WAV and TXT files are as follows source_articleID_segmentID. The audio recordings of both speakers have been downsampled to 22.05 kHz, with samples stored as 16-bit signed integers. The metadata contains speaker information, such as age, gender, working experience, and recording device.

In total, female and male speakers read around 1,000 and 1,250 articles respectively, out of which around 300 articles are overlapping.

<table>
<thead>
<tr>
<th>Speaker ID</th>
<th>Gender</th>
<th>Age</th>
<th>Work experience</th>
<th>Recording device</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Female</td>
<td>44</td>
<td>14 years</td>
<td>AKG P120</td>
</tr>
<tr>
<td>M1</td>
<td>Male</td>
<td>46</td>
<td>12 years</td>
<td>Tascam DR-40</td>
</tr>
</tbody>
</table>

*Note that the Kazakh writing system differs depending on the region where it is used. In Kazakhstan, the Cyrillic alphabet is used.*
Table 2: The KazakhTTS dataset specifications

<table>
<thead>
<tr>
<th>Category</th>
<th>Speaker ID</th>
<th>F1</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td># Segments</td>
<td>Total</td>
<td>17,426</td>
<td>24,656</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>36.1 hr</td>
<td>57.1 hr</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>7.5 sec</td>
<td>8.3 sec</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>24.2 sec</td>
<td>55.9 sec</td>
</tr>
<tr>
<td>Words</td>
<td>Total</td>
<td>245.4k</td>
<td>320.2k</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>14.0</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>42</td>
<td>75</td>
</tr>
<tr>
<td># Unique words</td>
<td>34.2k</td>
<td>42.4k</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Segment duration and length distributions for female (a and b) and male (c and d) speakers in the KazakhTTS.

4. TTS experiments

To demonstrate the utility of the constructed dataset, we built the first Kazakh E2E-TTS systems and evaluated them using the subjective MOS measure.

4.1. Experimental setup

We used ESPnet-TTS toolkit [22] to build the E2E-TTS models based on Tacotron 2 [7] and Transformer [9] architectures. Specifically, we followed the LJ Speech [10] recipe and used the latest ESPnet-TTS developments to configure our model building recipe. The input for each model is a sequence of characters consisting of 42 letters and 5 symbols (‘.’, ‘’, ‘-‘, ‘?’., ‘!’), and the output is a sequence of acoustic features (80 dimensional log Mel-filter bank features). To transform these acoustic features into the time-domain waveform samples, we applied the Griffin-Lim algorithm [23], WaveNet [24], and WaveGAN [25] vocoders. We found WaveGAN to perform best in our case, and it was used in our final E2E-TTS systems. We did not apply any additional speech pre-processing, such as filtering, normalization, and so on.

In the Tacotron 2 system, the encoder module was modeled as a single bi-directional LSTM layer with 512 units (256 units in each direction), and the decoder module was modeled as a stack of two unidirectional LSTM layers with 1,024 units. The parameters were optimized using the Adam algorithm [26] with an initial learning rate of $10^{-3}$ for 200 epochs. To regularize parameters, we set the dropout rate to 0.5.

The Transformer system was modeled using six encoder and six decoder blocks. The number of heads in the self-attention layer was set to 8 with 512-dimension hidden states, and the feed-forward network dimensions were set to 1,024. The model parameters were optimized using the Adam algorithm with an initial learning rate of 1.0 for 200 epochs. The dropout rate was set to 0.1.

For each speaker, we trained separate E2E-TTS models (i.e., single speaker model). All models were trained using the Tesla V100 GPUs running on an NVIDIA DGX-2 server. More details on model specifications and training procedures are provided in our GitHub repository.

4.2. Experimental evaluation

To assess the quality of the synthesised recordings, we conducted subjective evaluation using the MOS measure. We performed a separate evaluation session for each speaker. In each session, the following three systems were compared: 1) Ground truth (i.e., natural speech), 2) Tacotron 2, and 3) Transformer.

As an evaluation set, we randomly selected 50 sentences of various lengths from each speaker. These sentences were not used to train the models. The selected sentences were manually checked to ensure that each of them is a single complete sentence, and the speaker read them well (i.e., without disfluencies, mispronunciations, or background noise). The number of listeners was 50 in both evaluation sessions. The listeners were instructed to assess the overall quality, use headphones, and sit in a quiet environment.

The evaluation sessions were conducted through the instant messaging platform Telegram [27], since it is difficult to find native Kazakh listeners on other well-known platforms, such as Amazon Mechanical Turk [28]. We developed two separate Telegram bots for each speaker. The bots first presented a welcome message with the instruction and then started the evaluation process. During the evaluation, the bots sent a sentence recording with the transcript to a listener and received the corresponding evaluation score (see Figure 2). The recordings were rated using a 5-point Likert scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Note that in Telegram, to send audio recordings, we had to convert them into MP3 format.

We attracted listeners to participate in the evaluation sessions by advertising the project in social media, news, and open messaging communities on WhatsApp and Telegram. The listeners were allowed to listen to recordings several times, but they were not allowed to alter previous ratings once submitted. Additionally, the Telegram bots were keeping track of the listeners’ status and ID. As a result, the listeners could take a break to continue at a later time, and were prevented from participating in the evaluation session more than once.

For all the listeners, the evaluation recordings were presented in the same order and one at a time. However, at each time step, the bots randomly decided from which system to pick.

---

3 Due to the strong correspondence between word spelling and pronunciation in Kazakh, we did not convert graphemes to phonemes.

6 In fact, the number of listeners was higher than 50, however, we excluded the ones who did not complete the session till the end and the suspicious listeners who, for example, rated all recordings as excellent, rated all ground truth recordings as bad, and so on.

7 Due to the crowdsourced nature of the evaluation process, we can not guarantee that all listeners used headphones and sat in a quiet environment.
To identify error types made by E2E-TTS system, we manually analyzed a 50-sentence evaluation set for both speakers. This analysis was conducted only for the Tacotron 2 system which achieved better MOS score than the Transformer. Specifically, we counted sentences with various error types. The analysis results are provided in Table 4, showing that the most frequent error types for both speakers are mispronunciation and incomplete words. The mispronunciation errors are mostly due to incorrect stress, and the incomplete word errors mostly occur at the last word of a sentence, where the last letters of the word are trimmed. Interestingly, the total number of errors in the male speaker’s recordings is considerably higher than the total number of errors in the female speaker’s recordings. This might be one of the reasons for the lower MOS score achieved by the male speaker. This analysis indicates that there is still room for improvement and future work should focus on eliminating these errors.

### 5. Conclusion

This paper introduced the first open-source Kazakh speech dataset for TTS applications. The KazakhTTS dataset contains over 93 hours of speech data (36 hours by female and 57 hours by male) consisting of around 42,000 recordings. We released our dataset under the Creative Commons Attribution 4.0 International License, which permits both academic and commercial use. We shared our experience by describing the dataset construction and TTS evaluation procedures, which might benefit other researchers planning to collect speech data for other low-resource languages. Furthermore, the presented dataset can also aid to study and build (e.g., pretrained) TTS systems for other Turkic languages. To demonstrate the use of our dataset, we built E2E-TTS models based on Tacotron 2 and Transformer architectures. The subjective evaluation results suggest that the E2E-TTS models trained on KazakhTTS are suitable for practical use. We also shared the pretrained TTS models and the ESPnet training recipes for both speakers.

In future work, we plan to further extend our dataset by collecting more data from different domains, such as Wikipedia articles and books, and introduce new speakers. We also plan to explore the optimal hyper-parameter settings for Kazakh E2E-TTS models, compare different TTS architectures, and conduct additional analysis.

### 6. Acknowledgements

We would like to thank our senior moderator Aigerim Boranbayeva for helping to train and monitor the transcribers. We also thank our technical editor Rustem Yeshpanov, PR manager Kuralay Baimenova, project coordinator Yerbol Absalyamov, and administration manager Gibrat Kurmanov for helping with other administrative and technical tasks.
7. References


