

KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics

Saida Mussakhoyeva, Yerbolat Khassanov, Huseyin Atakan Varol

Institute of Smart Systems and Artificial Intelligence (ISSAI),

Nazarbayev University, Nur-Sultan, Kazakhstan

{saida.mussakhoyeva, yerbolat.khassanov, ahvarol}@nu.edu.kz

Abstract

We present an expanded version of our previously released Kazakh text-to-speech (KazakhTTS) synthesis corpus. In the new KazakhTTS2 corpus, the overall size has increased from 93 hours to 271 hours, the number of speakers has risen from two to five (three females and two males), and the topic coverage has been diversified with the help of new sources, including a book and Wikipedia articles. This corpus is necessary for building high-quality TTS systems for Kazakh, a Central Asian agglutinative language from the Turkic family, which presents several linguistic challenges. We describe the corpus construction process and provide the details of the training and evaluation procedures for the TTS system. Our experimental results indicate that the constructed corpus is sufficient to build robust TTS models for real-world applications, with a subjective mean opinion score ranging from 3.6 to 4.2 for all the five speakers. We believe that our corpus will facilitate speech and language research for Kazakh and other Turkic languages, which are widely considered to be low-resource due to the limited availability of free linguistic data. The constructed corpus, code, and pretrained models are publicly available in our GitHub repository.

Keywords: text-to-speech, TTS, speech synthesis, speech corpus, open-source, Kazakh, Turkic, agglutinative

1. Introduction

Text-to-speech (TTS), also known as speech synthesis, is the automatic process of converting written text into speech (Taylor, 2009), which has wide application potential and a substantial social impact, including digital assistants, improved accessibility for people with reading disabilities, speech and vision impairments, to name a few. For visually impaired people, in particular, it enables voice-controlled access to Internet-of-things devices, on-demand access to books and websites, and access to other vocalized assistive technologies. In turn, these enhance the overall quality of life, consumption of information, and access to knowledge. In addition, TTS can complement other important language and vision technologies, such as speech recognition (Tjandra et al., 2017), speech-to-speech translation (Wahlster, 2013), face-to-face translation (Prajwal et al., 2019), and visual-to-sound (Zhou et al., 2018). Considering the aforementioned benefits, TTS is undoubtedly an essential speech processing technology for any language.

In recent years, TTS research has progressed remarkably thanks to neural network-based architectures (Tan et al., 2021), regularly organized challenges (Black and Tokuda, 2005; Dunbar et al., 2019), and open-source datasets (Ito and Johnson, 2017; Zen et al., 2019; Shi et al., 2020). Especially, impressive results have been achieved for commercially viable languages, such as English and Mandarin. However, there is still a lack of research into the development of TTS technologies for low-resource languages. To address this problem in regard to Kazakh, Mussakhoyeva et al. (2021a) have recently developed the first open-source Kazakh

text-to-speech (KazakhTTS) corpus, which contains 93 hours of manually transcribed audio from two professional speakers (one female and one male) reading news articles. The developed corpus has generated substantial interest and has been downloaded over 200 times in less than a year by academia and industry, both from local and global organizations. This demonstrates high demand for open-source and high-quality transcribed speech data in the Kazakh language.

Motivated by this, in this paper, we present a new version of the KazakhTTS corpus called KazakhTTS2, which adds more data, speakers, and topics to our corpus. Specifically, we have increased the data size from 93 hours to 271 hours. We have added three new professional speakers (two females and one male), with over 25 hours of transcribed data for each speaker. In addition to news, we have diversified the topic coverage with a book and Wikipedia articles. Like the first version, KazakhTTS2 is freely available to both academic researchers and industry practitioners in our GitHub repository¹.

To validate the KazakhTTS2 corpus, we built a state-of-the-art TTS system based on the Tacotron 2 (Shen et al., 2018) architecture. The constructed TTS system was evaluated using the subjective mean opinion score (MOS) measure. The obtained MOSs for all the speakers ranged from 3.6 to 4.2, which indicates the utility of the KazakhTTS2 corpus for building robust TTS systems suitable for real-world applications. We believe that our corpus will further facilitate the rapid development of TTS systems in the Kazakh language and thus serve as an enabler for the wide range of applica-

¹https://github.com/IS2AI/Kazakh_TTS

tions mentioned above. We also believe that this work will encourage subsequent efforts in this area to address some of the practical issues that arise when training TTS systems for the Kazakh language. Additionally, our corpus can be employed to bootstrap speech technologies for other similar languages from the Turkic family, for example, by means of cross-lingual transfer learning (Chen et al., 2019) and self-supervised pre-training (Baevski et al., 2020).

To sum up, our main contributions are:

- We developed a text-to-speech synthesis corpus for the Kazakh language containing five speakers (three females and two males) comprising 271 hours of carefully transcribed data from various sources (news, book, and Wikipedia).
- We validated the efficacy of the corpus, by training state-of-the-art neural TTS models, which achieved a sufficient subjective MOS for most practical applications.
- The KazakhTTS2 corpus, code, and pretrained models were made publicly available¹ for both commercial and academic use.

The rest of this paper is organized as follows: Section 2 reviews the work on Kazakh language corpus creation. In Section 3, we briefly summarize the previous release of the corpus and explain the changes made in KazakhTTS2, including the corpus structure and statistics. The experimental setup and evaluation results are described in Section 4. Section 5 discusses the challenges of Kazakh speech synthesis and future research directions. Section 6 concludes this work.

2. Related Work

Despite its under-resourced status, Kazakh language research is emerging as an evolving field with an increasing number of recently released open-source corpora. For example, Khassanov et al. (2021) developed the first large-scale publicly available corpus for automatic speech recognition (ASR). The corpus was collected by means of crowdsourcing, with over 2,000 people contributing around 330 hours of audio recordings. Similarly, Yeshpanov et al. (2021) developed an open-source Kazakh named entity recognition dataset consisting of over 100,000 sentences annotated for 25 entity classes. Linguistic corpora development has also been observed in neighboring countries with languages similar to Kazakh, such as Uzbek (Musaev et al., 2021). Additionally, there are other large-scale projects aimed at collecting open-source corpora for various languages, including Kazakh, such as Common Voice (Ardila et al., 2020). However, all these datasets are unsuitable for building robust Kazakh TTS systems, which require a large number of high-quality audio recordings of a single speaker.

The first attempt to collect a large-scale open-source TTS dataset for the Kazakh language was made by Mussakhoyeva et al. (2021a). The collected dataset was called KazakhTTS and consisted of 93

hours of carefully transcribed audio from two professional speakers. Specifically, the speakers were assigned to read local news articles. The recorded articles were manually segmented into sentences and then aligned with the corresponding text with the help of native Kazakh transcribers. The TTS systems developed using KazakhTTS achieved an MOS of above 4.0, demonstrating the high quality of the collected data. This work further extends the KazakhTTS corpus, as described in the following sections.

The other existing corpora dedicated to Kazakh TTS are either proprietary or have been collected by leveraging unsupervised and semi-supervised approaches. For example, Black (2019) extracted readings of the Bible in hundreds of languages, including Kazakh. The extracted recordings were automatically segmented and aligned with the corresponding text. Although a TTS system built using this corpus is sufficient to deploy in some use-cases, its overall quality is unsatisfactory for most real-world applications. Specifically, in the evaluation experiments, the Kazakh TTS system achieved a mel-cepstral distortion score of more than 6, which is considered low quality². In another work, Khomitsevich et al. (2015) developed a Kazakh TTS system using a female voice. However, the authors did not provide any information on their corpus, such as its size, how the recordings were acquired, and how to download it. Additionally, the authors did not describe the evaluation procedures performed to assess the developed TTS system.

3. KazakhTTS2 Corpus

In this section, we describe the curation procedures for the KazakhTTS2 corpus. The KazakhTTS2 corpus collection was approved by the Institutional Research Ethics Committee of Nazarbayev University. We first briefly summarize the previous version of the corpus (i.e., KazakhTTS) and then systematically explain the changes made to extend it.

3.1. KazakhTTS

KazakhTTS is the first version of our corpus, which contains around 93 hours of transcribed audio consisting of over 42,000 sentences. The audio was recorded by two professional speakers, both of whom had had over ten years of narration experience in local television and radio stations. The speakers were assigned to read news articles covering various topics, such as sports, business, politics, and so on. The recorded audio was manually segmented into sentences, with defective segments (e.g., mispronunciation and external noise) filtered out. The correspondence between audio and text was verified by native Kazakh transcribers. The statistics for the first and second versions of the corpus are provided in Table 1.

²http://festvox.org/cmu_wilderness/index.html

Category	KazakhTTS	KazakhTTS2
# Speakers	2	5
# Segments	42,082	136,196
# Tokens	565.6k	1.7M
# Unique tokens	54.9k	107.9k
Duration	93.2 h	271.7 h

Table 1: The comparison of statistics for the KazakhTTS and KazakhTTS2 corpora

3.2. Text Collection

We began by collecting additional news articles from four local news websites. To further broaden the topic coverage, we added a book from the public domain and Wikipedia articles. From Wikipedia, we extracted articles on science, computer technology, countries, and history. All the articles were manually extracted to eliminate defects peculiar to web crawlers and saved in the DOC format for the professional speakers’ convenience (i.e., font size, line spacing, and typeface could be adjusted to the preferences of the speakers). In total, over 2,500 additional news articles, one book, and 159 Wikipedia articles were extracted.

3.3. Recording Process

To narrate the collected text, we auditioned several candidates and, as a result, hired three professional speakers (two females and one male). Each speaker participated voluntarily and was informed of the protocols for data collection and use through an informed consent form. All the hired speakers were tasked to read news articles only. In addition, we rehired the male speaker (speaker M1) from the previous corpus creation process, because of his extensive experience in narrating documentaries. He was subsequently tasked to read the book and Wikipedia articles. The speaker specifications, including gender, age, professional experience as a narrator, and recording device information, are provided in Table 2. Speakers F1 and M1 were part of KazakhTTS, whereas F2, F3, and M2 are newly hired speakers.

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 in their makeshift studios that they had set up to work from home. The speakers were instructed to read the texts in a quiet indoor environment with neutral tone and pace. They were also asked to follow orthoepic rules, to maintain a constant distance between the microphone and lips, to pause at commas, and to intonate sentences ending with a question mark appropriately. In total, each newly hired speaker read around 1,400 news articles, and Speaker M1 read one book entitled *Abai Zholy* (The Path of Abai) and 159 Wikipedia articles.

Speaker ID	Gender	Age	Work experience	Recording device
F1*	Female	44	14 years	AKG P120
F2	Female	39	15 years	RØDE
F3	Female	52	25 years	Behringer C-1
M1*	Male	46	12 years	Tascam DR-40
M2	Male	33	11 years	Mi

Note. Speakers F1 and M1 are from KazakhTTS.

Table 2: The KazakhTTS2 speaker information

3.4. Segmentation and Alignment

For audio segmentation and audio-to-text alignment, we employed the same approach as in the KazakhTTS corpus construction. We hired five native Kazakh transcribers with different backgrounds and thorough knowledge of Kazakh grammar rules. The transcribers manually segmented the recordings into sentence-level chunks and aligned them with the corresponding text using the Praat toolkit (Boersma, 2001). All the texts were represented using a Cyrillic script consisting of 42 letters³ and other punctuation marks, such as period, comma, hyphen, question mark, and exclamation mark. The transcribers were instructed to remove segments with mispronunciation and background noise, to trim long pauses at the beginning and end of segments, and to convert numbers and special characters (e.g., ‘%’, ‘\$’, ‘+’, etc) into the written form. To ensure the uniform quality of work among the transcribers, we assigned a linguist to randomly check the completed tasks and to organize regular “go through errors” sessions. To ensure the correctness of the audio-to-text alignment process, the segmented recordings were inspected using our internal ASR system trained on the KSC dataset (Khassanov et al., 2021). Specifically, the ASR system was used to generate segment transcriptions, which were then compared to the corresponding manually annotated transcripts. Segments with a high character error rate (CER) were regarded as incorrectly transcribed, and therefore rechecked by the linguist.

3.5. Corpus Structure and Statistics

The file structure of the KazakhTTS2 corpus is shown in Figure 1. Collections of audio recordings and the corresponding transcriptions are stored in a separate folder for each speaker. Additionally, for Speaker M1, we split the data from different sources into separate folders (i.e., News, Wiki, and Book). All audio recordings were downsampled to 22.05 kHz and stored at 16 bits per sample in the WAV format. All transcripts are stored as TXT files in the UTF-8 encoding. The audio and the corresponding transcript filenames are identical except for the extension. The

³Note that at the time of writing, the Cyrillic alphabet is the official alphabet used for the Kazakh language, though the transition process to the Latin alphabet has already begun.

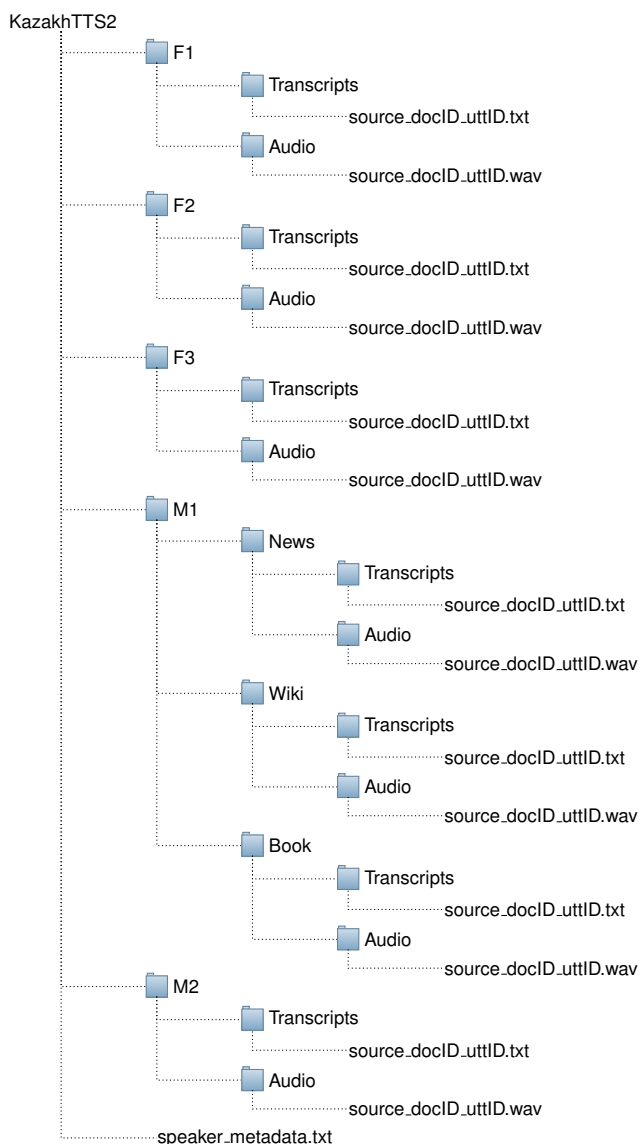


Figure 1: The file structure of KazakhTTS2

name of each file consists of the source name, document ID, and utterance ID (i.e., *source_docID_uttID*). Speaker information, including gender, age, professional experience, and recording device, is provided in the *speaker_metadata.txt* file.

The statistics for the KazakhTTS2 corpus are given in Table 3. The overall corpus size is around 271 hours, with each speaker having at least 25 hours of transcribed audio. The total number of sentences is around 136 thousand, and the total number of tokens is over 1.7 million, with unique token types per speaker ranging from 28.5 thousand to 80.7 thousand. Figure 2 presents the histograms of the distributions of sentence duration and length (in words) for each speaker in KazakhTTS2. For all speakers, the majority of sentence durations are between 3 and 6 seconds. The majority of sentence lengths are between 11 and 15 words for female speakers, and between 6 and 10 words for male speakers.

4. Speech Synthesis Experiments

In this section, we describe the experiments conducted to validate the utility of the KazakhTTS2 corpus. We first describe the experimental setup, followed by our evaluation procedures and results.

4.1. Experimental Setup

We used the ESPnet-TTS toolkit (Hayashi et al., 2020) to build end-to-end TTS models based on the Tacotron 2 (Shen et al., 2018) architecture. Specifically, we followed the training recipe of LJ Speech (Ito and Johnson, 2017). All TTS models were trained using Tesla V100 GPUs running on NVIDIA DGX 2 machines. 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 employed WaveGAN (Yamamoto et al., 2020) vocoders.

In the Tacotron 2 model, the encoder module was modeled as a single bidirectional 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 (Kingma and Ba, 2015) with an initial learning rate of 10^{-3} for 200 epochs. To mitigate overfitting, we applied a dropout of 0.5. A separate Tacotron 2 model was trained for each speaker (i.e., a single speaker model). More details on the model specifications and training procedures are provided in our GitHub repository¹.

4.2. Experimental Evaluation

To assess the quality of the synthesized recordings, we performed a subjective evaluation using the MOS measure. We evaluated only the voices developed using the newly collected data⁴ (i.e., F2, F3, M1 Wikipedia and Book, and M2), as the other data had already been evaluated in the previous work (Mussakhojayeva et al., 2021a). The evaluation procedure was similar to that of our previous work, except for the number of sentences selected as a test set. Specifically, in this work, we selected 25 sentences of varying lengths from each speaker, whereas, in the previous work, 50 sentences per speaker were selected. The reason for selecting a smaller number of sentences is based on our observation that raters become exhausted or bored after around 25 sentences and quit the evaluation session. The evaluation sentences were not used to train the models.

The speakers were evaluated in separate sessions, and in each session we compared the ground truth (i.e., natural speech) recordings against the Tacotron 2 synthesized recordings. The ground truth sentences were manually checked to ensure that the speaker read them

⁴For Speaker M1, we trained two separate models from scratch using Wikipedia and Book data.

Speaker ID	Source	# Segments	Segment duration				Tokens				
			Total	Mean	Min	Max	Total	Mean	Min	Max	Unique
F1	News*	17,426	36.1 h	7.5 s	1.0 s	24.2 s	245.4k	14.1	2	42	34.3k
F2	News	12,921	25.7 h	7.2 s	0.8 s	22.0 s	177.9k	13.8	1	42	28.5k
F3	News	23,696	48.5 h	7.4 s	0.7 s	21.4 s	331.3k	14.0	1	48	43.0k
M1	News*	24,608	57.0 h	8.3 s	0.8 s	55.9 s	319.6k	13.0	1	75	42.5k
	Wiki	13,189	29.7 h	8.1 s	0.9 s	29.7 s	166.8k	12.6	1	47	33.6k
	Book	11,453	16.7 h	5.3 s	0.8 s	21.1 s	107.6k	9.4	2	40	23.5k
	All	49,250	103.5 h	7.6 s	0.8 s	55.9 s	594.1k	12.1	1	75	80.7k
M2	News	32,903	57.9 h	6.3 s	0.7 s	28.3 s	393.3k	12.0	1	60	53.3k

Note. The news source data of speakers F1 and M1 are from KazakhTTS.

Table 3: The KazakhTTS2 dataset specifications

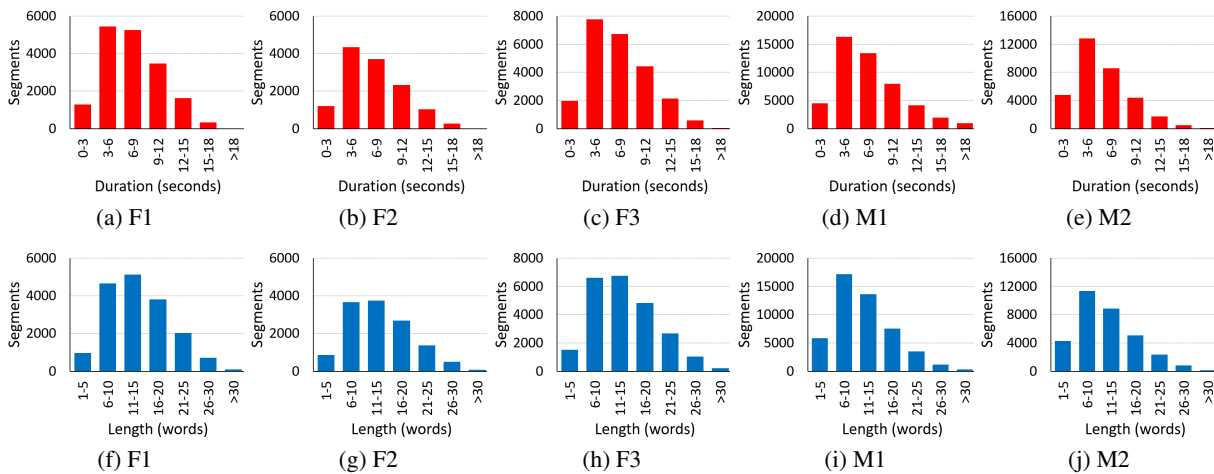


Figure 2: Segment duration (a, b, c, d, e) and length (f, g, h, i, j) distributions for each speaker of KazakhTTS2

well (i.e., without disfluencies, mispronunciations, or background noise).

Evaluation sessions were conducted using the instant messaging platform Telegram (Telegram Messenger Inc., 2013), as it is difficult to find native Kazakh raters on other well-known platforms, such as Amazon Mechanical Turk (Amazon.com Inc., 2005). We developed a separate evaluation Telegram bot for each speaker. The bots first presented a welcome message with instructions and then started the evaluation process. During the evaluation, the bots sent a sentence recording⁵ with the associated transcript to a rater and received the corresponding evaluation score. Recordings were rated using a five-point Likert scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad.

The raters were instructed to assess the overall quality through headphones in a quiet environment⁶. They were allowed to listen to the recordings several times,

⁵Note that in Telegram, to send audio recordings, we had to convert them into MP3 format.

⁶Due to the crowdsourced nature of the evaluation process, we cannot guarantee that all raters used headphones and sat in a quiet environment.

but they were not allowed to alter the ratings once submitted. Additionally, the Telegram bots kept track of the raters' ID, to prevent them from participating in the evaluation session more than once.

The evaluation recordings were presented in the same order and one at a time. However, at each time step, the bots randomly decided which version of a recording to select (i.e., ground truth or synthesized). As a result, each rater heard only one of the versions of a recording, and both systems (i.e., ground truth and Tacotron 2) were presented to all the raters. Each recording was rated at least 24 times for all the three speakers. The numbers of raters were 57, 61, 116, 89, and 53 for speakers F2, F3, M1 Wikipedia, M1 Book, and M2, respectively⁷.

At the end of the evaluation, the bots thanked the raters and invited them to fill in an optional questionnaire about their age, region (where a rater grew up and learned the Kazakh language), and gender. The questionnaire results showed that the raters varied in gender

⁷In fact, the number of raters was higher, but we excluded the ratings of those who did not go through the session to the end, or whose ratings were suspicious (e.g., all scores are "excellent" or all scores are "bad".)

Speaker ID	Source	Ground truth	Tacotron 2
F1	News*	4.726 ± 0.037	4.535 ± 0.049
F2	News	4.449 ± 0.059	4.061 ± 0.070
F3	News	4.178 ± 0.073	4.049 ± 0.076
M1	News*	4.360 ± 0.050	4.144 ± 0.063
	Wiki	4.483 ± 0.040	3.673 ± 0.070
	Book	4.564 ± 0.045	4.057 ± 0.068
M2	News	4.431 ± 0.062	4.200 ± 0.073

Note. The scores of speakers F1 and M1 News are from the previous work (Mussakhoyeva et al., 2021a).

Table 4: Mean opinion score (MOS) results with 95% confidence intervals

and region, but not in age (most of them were under 20). Specifically, the majority of raters were from the south and west of Kazakhstan, and females outnumbered males by a factor of 1.5.

4.3. Experiment Results

The subjective evaluation results are given in Table 4. As expected, the ground truth recordings received higher MOS scores than the Tacotron 2 synthesized ones. Nevertheless, all synthesized recordings except M1 Wikipedia scored above 4.0 on the MOS measure and were close to the ground truth (i.e., 8.7%, 3.1%, 18.1%, 11.1% and 5.2% relative MOS reductions for speakers F2, F3, M1 Wikipedia, M1 Book, and M2, respectively). These results demonstrate the utility of our KazakhTTS2 dataset for TTS applications.

Overall, the highest MOS score among the synthesized recordings was achieved by Speaker F1, and the lowest score was achieved by M1 Wikipedia. Presumably, the reason for the poor performance of M1 Wikipedia is the wide variety of topics and the abundance of rare scientific terms (from chemistry, biology, information technology, etc.). We believe that the performance of M1 Wikipedia can be improved by exploiting other data from Speaker M1. For example, by pre-training a model on M1 News and Book data, followed by fine-tuning using M1 Wikipedia.

In addition, we conducted an objective evaluation in which we manually analyzed the synthesized evaluation set recordings. Specifically, we counted the various error types made by the Tacotron 2 systems built using the newly collected data. The objective evaluation results are given in Table 5, which are consistent with the subjective evaluation, with Speaker M2 having the lowest number of errors, followed by F2 and M1 Book, and then F3 and M1 Wikipedia. The most common error types among all speakers are mispronunciation, incomplete words, and word skipping. This analysis indicates that there is still room for improvement and future work should focus on eliminating these errors.

Error types	Speaker ID				
	F2	F3	M1		M2
			Wiki	Book	
# repeated words	0	0	0	0	0
# skipped words	5	6	0	1	1
# mispronounced words	2	1	11	5	4
# incomplete words	2	7	0	3	4
# long pauses	0	0	4	1	0
# nonverbal sounds	1	0	0	0	0
Total	10	14	15	10	9

Table 5: Manual analysis of error types made by Tacotron 2

5. Challenges and Future Work

The Kazakh language presents several challenges to the speech synthesis task. The first one is code-switching, as the majority of Kazakh speakers are bilingual in Kazakh and Russian. While the languages are not mixed in most formal situations (e.g., news, books, law, etc.), intrasentential code-switching often occurs in informal conversations. Moreover, intra-word code-switching is also possible (e.g., Kazakh stem words with Russian suffixes or vice versa), which may further deteriorate TTS quality.

Additionally, Kazakh has a large number of loanwords from Russian, and these words usually retain the orthographic and phonological properties of the source language. This has especially important consequences for TTS applications, as Russian differs from Kazakh in many aspects. For example, in most Kazakh words, the stress is fixed on the final syllable, while in Russian, the stress can be on any syllable of a word (Jouravlev and Lupker, 2014). Furthermore, the spelling of Kazakh words closely matches their pronunciation, which is not the case with Russian words; for example, the letter “o” is sometimes pronounced as /a/. It is important to mention that due to globalization, the number of loanwords from other languages, especially English, is also increasing, which is likely to pose an additional challenge in the near future (Mussakhoyeva et al., 2021b). Another challenge is that Kazakh is an agglutinative language, with a very large vocabulary and many characters per word. It is also susceptible to morphophonemic changes arising during word formation. One of the solutions would be to increase the size of the Kazakh speech corpus to cover more word formation variants. We believe that overcoming these challenges for the Kazakh language will be an interesting direction for future research.

6. Conclusion

We have presented KazakhTTS2, a large-scale open-source Kazakh text-to-speech corpus, which further extends the previous work with more data, voices, and topics. The corpus consists of five voices (three female and two male), with over 270 hours of high-quality

transcribed data. The corpus is publicly available, which permits both academic and commercial use. We validated the corpus by means of crowdsourced subjective evaluation, where all voices synthesized using the Tacotron 2 model achieved an MOS of above 3.6, making it suitable for practical deployment. To enable experiment reproducibility and facilitate future research, we shared our training recipes and pretrained models in our GitHub repository¹. Although the corpus was designed with TTS application in mind, it can be used to complement other speech processing applications, such as speech recognition and speech translation. We hope the TTS corpus construction and evaluation procedures described in this paper will contribute to the burgeoning field of Kazakh speech and language research and help advance the state-of-the-art for other low-resource languages of the Turkic family.

7. Acknowledgements

The authors would like to thank Aigerim Borambayeva, Almas Mirzakhmetov, Dias Bakhtiyarov, and Rustem Yeshpanov for their help in data collection, voice evaluation, and paper revision. The authors would also like to thank the speakers for their recordings and the anonymous raters for their evaluations.

8. Bibliographical References

- Amazon.com Inc. (2005). Amazon Mechanical Turk (MTurk).
- Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In *LREC*, pages 4218–4222. ELRA.
- Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 12449–12460.
- Black, A. W. and Tokuda, K. (2005). The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In *European Conference on Speech Communication and Technology (Interspeech)*, pages 77–80. ISCA.
- Black, A. W. (2019). CMU wilderness multilingual speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5(9):341–345.
- Chen, Y., Tu, T., Yeh, C., and Lee, H. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. In *Interspeech*, pages 2075–2079. ISCA.
- Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S., and Dupoux, E. (2019). The zero resource speech challenge 2019: TTS without T. In *Interspeech*, pages 1088–1092. ISCA.
- Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y., and Tan, X. (2020). ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658.
- Ito, K. and Johnson, L. (2017). The LJ speech dataset. <https://keithito.com/LJ-Speech-Dataset/>.
- Jouravlev, O. and Lupker, S. J. (2014). Stress consistency and stress regularity effects in russian. *Language, Cognition and Neuroscience*, 29(5):605–619.
- Khassanov, Y., Mussakhoyayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M., and Varol, H. A. (2021). A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 697–706. Online, April. Association for Computational Linguistics.
- Khomitsevich, O., Mendelev, V., Tomashenko, N. A., Rybin, S., Medennikov, I., and Kudubayeva, S. (2015). A bilingual kazakh-russian system for automatic speech recognition and synthesis. In *Speech and Computer (SPECOM)*, volume 9319 of *Lecture Notes in Computer Science*, pages 25–33. Springer.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*.
- Musae, M., Mussakhoyayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., and Varol, H. A. (2021). USC: an open-source uzbek speech corpus and initial speech recognition experiments. In *Speech and Computer (SPECOM)*, volume 12997 of *Lecture Notes in Computer Science*, pages 437–447. Springer.
- Mussakhoyayeva, S., Janaliyeva, A., Mirzakhmetov, A., Khassanov, Y., and Varol, H. A. (2021a). KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset. In *Proc. Interspeech 2021*, pages 2786–2790.
- Mussakhoyayeva, S., Khassanov, Y., and Varol, H. A. (2021b). A study of multilingual end-to-end speech recognition for Kazakh, Russian, and English. In *Speech and Computer (SPECOM)*, volume 12997 of *Lecture Notes in Computer Science*, pages 448–459. Springer.
- Prajwal, K. R., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., and Jawahar, C. V. (2019). Towards automatic face-to-face translation. In *Proceedings of the International Conference on Multimedia (MM)*, pages 1428–1436. ACM.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Ryan,

- R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.
- Shi, Y., Bu, H., Xu, X., Zhang, S., and Li, M. (2020). AISHELL-3: A multi-speaker Mandarin TTS corpus and the baselines. *CoRR*, abs/2010.11567.
- Tan, X., Qin, T., Soong, F., and Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.
- Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.
- Telegram Messenger Inc. (2013). Telegram.
- Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking: Speech chain by deep learning. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308.
- Wahlster, W. (2013). *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.
- Yamamoto, R., Song, E., and Kim, J. (2020). Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203.
- Yeshpanov, R., Khassanov, Y., and Varol, H. A. (2021). KazNERD: Kazakh named entity recognition dataset. *arXiv preprint arXiv:2111.13419*.
- Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech*, pages 1526–1530. ISCA.
- Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. L. (2018). Visual to sound: Generating natural sound for videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.