FeruzaSpeech: A 60 Hour Uzbek Read Speech Corpus with Punctuation, Casing, and Context

Anna Povey
Redmond High School, WA, USA
annapovey@gmail.com

Katherine Povey
University of Washington, USA
katherinepovey@gmail.com

Abstract

This paper introduces FeruzaSpeech, a read speech corpus of the Uzbek language, containing transcripts in both Cyrillic and Latin alphabets, freely available for academic research purposes. This corpus includes 60 hours of high-quality recordings from a single native female speaker from Tashkent, Uzbekistan. These recordings consist of short excerpts from a book and BBC News. This paper discusses the enhancement of the Word Error Rates (WERs) on CommonVoice 16.1's Uzbek data, Uzbek Speech Corpus data, and FeruzaSpeech data upon integrating FeruzaSpeech.

1 Introduction

The Uzbek language, the official language of Uzbekistan, boasts upwards of 31 million native speakers across Central Asia. [1] Advancement in neural network models and deep learning have significantly improved automated speech recognition (ASR) and text-to-speech (TTS) technology in recent decades. Large freely available English datasets, such as LibriSpeech (Panayotov et al., 2015), Libriheavy (Kang et al., 2023b), and GigaSpeech (Chen et al., 2021) are now more robust then ever, however, datasets for training these models in Uzbek are scarce.

In January 2023, the Uzbek government fully transitioned from using the Cyrillic alphabet to using the Latin alphabet [2], yet the country continues to use both alphabets. FeruzaSpeech is the first dataset to offer both Cyrillic and Latin transcription. To the best of our knowledge, FeruzaSpeech is also the only corpus to provide Cyrillic transcriptions at all. Datasets originally using Latin transcription cannot yet be accurately converted into Cyrillic text using online conversion calculators because there are a few discontinuities between the two alphabets. An example is when conversion calculators

like this one[3] are used on Cyrillic text that include the soft sign ь, it is either lost or can be incorrectly reproduced becoming a hard sign ъ.

Table 1: Conversion Calculator on Польша

Cyrillic to Latin	Latin to Cyrillic
Польша -> Pol'sha	Pol'sha -> Полъша

FeruzaSpeech aims to promote the development of speech recognition and speech synthesis technologies for the use of Uzbek speakers. Because this is a single speaker dataset with an absence of environmental noise it is better for STT when used in addition to other available speech corpuses. The dataset may be suitable for TTS applications, but such experiments are beyond the scope of this paper. It complements existing ASR datasets such Uzbek Speech Corpus (USC) (Musaev et al., 2021), consisting of 105 hours from 958 speakers, and the Common Voice Uzbek Dataset (Ardila et al., 2019) [4], with 265 hours from over 2,000 speakers. We chose these two corpuses because they were the only two other published datasets. When combined with these datasets, FeruzaSpeech enhances ASR model training.

2 FeruzaSpeech Corpus

This section describes the layout of the FeruzaSpeech corpus metadata, transcription, and audio format. Instructions for downloading and utilizing the data can be found on HuggingFace. [5]

2.1 Dataset Type

FeruzaSpeech consists of audio-book recordings from the texts of the book Choliqushi, a classic romance novel, and BBC Uzbek News read by our

---

[1] https://www.worlddata.info/languages/uzbek.php
[2] https://interfax.az/view/826747

[3] https://uzlatin.com/
[4] Download Page: https://commonvoice.mozilla.org/en/datasets
[5] https://huggingface.co/datasets/k2speech/FeruzaSpeech

voice actress, Feruza. Table 2 shows the duration of each type of recording within the dataset. Initially read in the Cyrillic alphabet, the texts were converted to Latin using online tools[6], with some grammatical errors being manually fixed after the use of the conversion calculator. The final transcription provides Uzbek text in both the Cyrillic and Latin alphabets.

Table 2: FeruzaSpeech Recordings

Type	Total
Book	21.57h
BBC Uzbek	38.04h
Total	59.61h

2.2 Evaluation and Training Sets

FeruzaSpeech includes "Dev" (development), "Test" (testing), and "Train" (training) sets as detailed in Table 3. Both the Dev and Test sets only include BBC articles, while the Train set also includes the Choliqushi novel.

Table 3: FeruzaSpeech Sets

Sets	Total
Dev	2.93h
Test	4.08h
Train	52.09h

2.3 Audio Format

The corpus contains high-quality, single-channel, 16-bit .wav audio files, available in 16kHz for ASR. The average recording length is 16.39 seconds, the minimum length is 3.78 seconds, and the maximum length is 50.69 seconds. Our segments are recordings of one to two full sentence and are much longer than the segments of USC (Musaev et al., 2021), that are mostly 2 to 3 seconds.

2.4 Sample Text

Table 4 shows example excerpts from the CommonVoice and USC datasets in comparison to our proposed FeruzaSpeech dataset.

We can see that CommonVoice text was normalized but it has punctuation and casing, USC text is normalized to remove casing and punctuation,
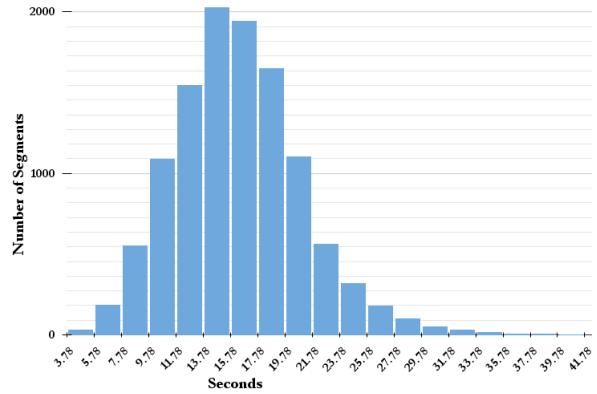

Figure 1: Length of FeruzaSpeech Segments

Table 4: One Sentence of Sample Text from each of the Three Datasets with English Translation That Is Not in the Dataset for Reader

CommonVoice	USC	FeruzaSpeech
— Non dema! — dedi. — Nonni otini atama!	shundan so'ng u sen aytmasang men aytaman degandek qaradi	20 iyul kuni O'zbekistonda 562 holatda kasallik qayd etilgan.
— Don't say bread! — he said. — Don't utter the word bread!	after that he looked like if you don't tell I will	On July 20, 562 cases of the disease were recorded in Uzbekistan.

and FeruzaSpeech retains casing and punctuation. Regarding the choice to not normalize casing or punctuation, deep learning models have recently become powerful enough that for tasks like ASR and TTS it is now feasible to use "natural" text with no normalization. For instance, the recent E2TTS text-to-speech system (Eskimez et al., 2024) is trained on data from Libriheavy (Kang et al., 2024) which is completely un-normalized. The use of un-normalized text for training tends to simplify speech processing systems because it could avoid the need for text normalization and inverse text normalization modules. Table 5 shows how the transcripts were provided in Latin and Cyrillic, but within this paper we only used Latin transcripts for comparison with available Latin datasets.

---

[6]https://www.lexilogos.com/keyboard/uzbek_conversion.htm and https://uzlatin.com/

Table 5: FeruzaSpeech Excerpt in Latin and Cyrillic with English Translation That Is Not in the Dataset for Reader

FeruzaSpeech Latin	FeruzaSpeech Cyrillic
Ayni damda ishlatilib turilgan biometrik pasportlar 2019 yil 1 yanvardan deyarli yaroqsiz holatga keladi.	Айни дамда ишлатилиб турилган биометрик паспортлар 2019 йил 1 январдан деярли яроқсиз ҳолатга келади.
Biometric passports, which are currently in use, will become almost useless from January 1, 2019.	

3 Experiments

To build our models we utilized the Next-Gen Kaldi framework and followed two recipes; the Icefall CommonVoice Stateless RNN-T Conformer model taken from the pruned_transducer_stateless7 recipe[7] and the Librispeech zipformer[8], noting significant improvements in WER. These two models were selected because Icefall already contains CommonVoice scripts using pruned_transducer_stateless7 in the French language and Librispeech zipformer is the current state of the art model in Next-Gen Kaldi. In our experiments we utilized three datasets: Common Voice 16.1 (CV), FeruzaSpeech (FS), and Uzbek Speech Corpus (USC). All models were trained for 60 epochs. Table 6 outlines the duration of each training dataset.

Table 6: Training Dataset Duration

Datasets	Training Duration
CV	54.88h
FS	52.09h
USC	90.70h
CV + FS + USC	197.68h

3.1 Pruned-transducer-stateless7 Model

The Stateless RNN-T Conformer model (Kang et al., 2023a) is a stateless transducer (Gulati et al., 2020) with a conformer encoder that reduces memory consumption, and it outperformed the small zipformer model for every test set. All models in Tables 7 and

8 are trained with Casing and Punctuation. Table 7 presents the WERs when the model is scored with Casing and Punctuation, while Table 8 presents the WERs when the model is scored with Uppercase No Punctuation.

Table 7: The WERs of **Stateless RNN-T Conformer model** scored with **Casing and Punctuation (C&P)** Common Voice 16.1 (CV), Uzbek Speech Corpus (USC), FeruzaSpeech (FS)

Method	Dataset	cv-test	fs-test	usc-test
greedy search	CV	33.95	32.9	51.07
	FS	89.54	11.58	85.67
	CV+FS	32.49	9.93	46.89
	CV+FS+USC	29.91	9.79	12.05
modified beam search	CV	31.98	31.88	51.61
	FS	89.10	11.25	85.22
	CV+FS	30.47	9.85	48.6
	CV+FS+USC	27.81	9.56	11.67

Table 8: The WERs of **Stateless RNN-T Conformer** model scored with **Uppercase No Punctuation (UNP)** Common Voice 16.1 (CV), Uzbek Speech Corpus (USC), and FeruzaSpeech (FS)

Method	Dataset	cv-test	fs-test	usc-test
greedy search	CV	21.03	20.15	35.11
	FS	87.18	5.85	77.78
	CV+FS	18.91	4.44	30.53
	CV+FS+USC	12.07	4.17	12.05
modified beam search	CV	20.16	19.34	34.03
	FS	86.26	5.50	76.24
	CV+FS	18.33	4.24	29.67
	CV+FS+USC	11.17	4.05	11.67

3.2 Zipformer Model

We followed a similar procedure for the zipformer model as we did for the Stateless RNN-T Conformer model. This time, we trained a separate model on each of the following three datasets: CV, FS, CV+FS. This differs from the previous section because we excluded the USC training set. Once again, we recorded the WER for each model when tested on each of the following test sets: cv-test, fs-test, and usc- test, sharing results for both the greedy search and modified beam search as methods of decoding. All models in Table 9 and 10 are trained with Casing and Punctuation. Table 9 presents the WERs when the model is scored with Casing and Punctuation, while Table 10 presents the WERs when the model is scored with Uppercase No Punctuation.

Note that the Common Voice recipe with default settings in the Icefall project wasn't converging for zipformer (Kang et al., 2023b), so we used "small zipformer" [9] (Yao et al., 2024) parameters to account for the size of our datasets.

Table 9: The WERs of **zipformer** model scored with **Casing and Punctuation (C&P)** Common Voice 16.1 (CV) and FeruzaSpeech (FS)

Method	Dataset	cv-test	fs-test	usc-test
greedy	CV	37.00	34.54	53.4
search	FS	93.09	14.32	N/A
	CV+FS	35.90	11.05	52.86
modified	CV	33.96	32.41	54.07
beam	FS	92.61	13.28	N/A
search	CV+FS	33.15	10.75	53.08

Table 10: The WERs of **zipformer** model scored with **Uppercase No Punctuation (UNP)** Common Voice 16.1 (CV) and FeruzaSpeech (FS)

Method	Dataset	cv-test	fs-test	usc-test
greedy	CV	23.01	20.94	38.75
search	FS	91.34	8.97	N/A
	CV+FS	22.34	5.45	37.4
modified	CV	21.92	20.21	37.44
beam	FS	90.54	7.88	N/A
search	CV+FS	21.35	5.10	35.94

4 Results

When adding the FeruzaSpeech dataset to the CommonVoice16.1 dataset while training the Stateless RNN-T Conformer model, WER improved 1.49 to 2.12 percent absolutely on cv-test and 3.01 to 4.58 percent absolutely on usc-test in Tables 7 and 8. And for the Zipformer model, WER improved 0.57 to 1.1 percent absolutely on cv-test and 0.54 to 1.5 percent absolutely on usc-test in tables 9 and 10. This shows that FeruzaSpeech contains quality data and is a useful addition to the current public library of Uzbek speech corpuses for TTS applications. Also, the paper presenting the USC dataset (Musaev et al., 2021) reports that the usc-test had a WER of 17.4%. Our best result for the usc-test WER is 11.67%, which is an improvement of 5.73%. According to Table 6 and 7, when a Stateless RNN-T Conformer model was built using all three datasets combined: CV, FS, and USC, and using modified beam search as the decoding method, the model produced the best WERs for every test.

Our best recorded WER on the Common Voice test set is 11.17%, as shown in Table 7. The best WER for the FeruzaSpeech test set is 4.05%, and the best WER for the Uzbek Speech Corpus test is 11.67%.

5 Conclusion

The development of FeruzaSpeech is a significant step forward in the field of Uzbek speech technology. By offering a dual alphabet corpus, this project bridges the gap between the use of Cyrillic and Latin scripts for Uzbek speakers. Our work also highlights the need for accurate alphabet conversion tools, specifically for more nuanced aspects of the language such as the soft sign (ь), which tends to be lost in translation from Cyrillic to Latin.

Through integrating FeruzaSpeech with existing Uzbek datasets, notable improvements in WERs were demonstrated. In the future, we will provide this same data in a higher sampling rate and bit depth that will be more suitable for TTS. Since we recognize the value of continuity in voice data for TTS applications, our future endeavors will also focus on expanding this corpus with additional recordings from the same native speaker. This strategy aims to enrich the dataset with consistent voice quality and style across the corpus which is essential for developing TTS models.

In sum, FeruzaSpeech is beneficial for ASR model enhancement when used in addition to existing Uzbek language datasets, as observed in WER improvements. Applications of this dataset for TTS will also be explored.

6 Limitations

FeruzaSpeech is not an effective stand alone corpus for STT applications and should be used in compliment with other corpuses such as the Common Voice Uzbek Dataset and Uzbek Speech Corpus explored above. FeruzaSpeech has an average segment length of 16.39 seconds which each contain one or two full sentences which could be segmented into shorter utterances. The audio has no background noise and contains a singular female speaker which is not optimal for STT.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-

multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Wei Kang, Liyong Guo, Fangjun Kuang, Long Lin, Mingshuang Luo, Zengwei Yao, Xiaoyu Yang, Piotr Żelasko, and Daniel Povey. 2023a. Fast and parallel decoding for transducer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2023b. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *arXiv preprint arXiv:2309.08105*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context.

Muhammadjon Musaev, Saida Mussakhojayeva, Ilyos Khujayorov, Yerbolat Khassanov, Mannon Ochilov, and Huseyin Atakan Varol. 2021. Usc: An opensource uzbek speech corpus and initial speech recognition experiments. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, pages 437–447. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition.