UZBEK AUTOMATIC SPEECH RECOGNITION MODELS USING DEEP LEARNING TECHNIQUES

Salaeva Makhliyo
Master student at Urgench State University,
makhliyo.salaeva@urdu.uz
Kuriyozov Elmurod
Teacher at Urgench State University,
elmurod1202@urdu.uz
Salaev Ulugbek
Ph.D. Student at Urgench State University,
ulugbek.salaev@urdu.uz

Annotatsiya. O‘zbek tili kabi resurslari kam bo‘lgan tillar uchun nutqni avtomatik aniqlash (ASR: Automatic Speech Recognition) tizimlarini yaratish yuqori sifatli nutq ma’lumotlari yo‘qligi sababli qiyin vazifa hisoblanadi. Ushbu maqolada biz audio kitoblardan ularning transkriptlari bilan ma’lumotlar to‘plamini va o‘zbek tili uchun ASR tizimini yaratish uchun ma’lumotlar manbasini yaratish uchun ochiq manba loyihasini oldik. Ma’lumotlar to‘plamiga asoslanib, u nutqni aniqlash uchun ikkita Deep Learning modeli bo‘yicha treyning o‘tkazildi. Bizning tajribalarimiz shuni ko‘rsatadiki, model ASR ko‘rsatkichlarida 18,2% WERga erishgan.

Kalit so‘zlar: *Nutqni avtomatik aniqlash, o‘zbek tili, takrorlanuvchi neyron to‘rlar, konvolyutsion neyron to‘rlar.*

Abstract: Building Automatic Speech Recognition (ASR) systems for less-resourced languages such as Uzbek is a challenging task due to the lack of high-quality speech data. In this article, we obtained a dataset using from audiobooks with their transcripts and the open-source project to create source of data for building ASR systems for Uzbek language. Based on the dataset we performed training by two Deep Learning models for speech recognition. Our experiments show that the model achieved 18.2% WERs in ASR performance.

Keywords: *Automatic Speech Recognition, Uzbek language, Recurrent Neural Networks, Convolutional Neural Networks*

Аннотация: Создание систем автоматического распознавания речи (ASR) для менее ресурсоемких языков, таких как узбекский, является сложной задачей из-за отсутствия высококачественных речевых данных. В этой статье мы получили набор данных, используя аудиокниги с их транскрипциями и проект с открытым исходным кодом для создания источника данных для построения систем ASR для узбекского языка. На основе набора данных мы провели обучение по двум моделям глубокого обучения для распознавания речи. Наши эксперименты показывают, что модель достигла 18,2% частота ошибок слов (WER) в производительности ASR.

Ключевые слова: *Автоматическое распознавание речи, узбекский язык, рекуррентные нейронные сети, сверхточные нейронные сети*

Introduction.

Linguistic resources in the field of natural language processing mainly correspond to widely spoken languages spoken by more than half of the world’s population. The availability of language resources has enhanced the progress of Text-to-Speech (TTS), Automatic Speech Recognition (ASR), and AI-based systems. However, the development and research of technologies for low-resource natural languages requires the development of sufficient resources in these types of languages.

Automatic speech recognition refers to the task of recognizing human speech and presenting it in text form. Advanced technologies have been developed in this field of research in recent decades. ASR is typically seen in user-facing applications such as virtual agents, live captioning, and clinical notes, where accurate speech transcription is critical. ASR is an important component of speech AI, which is a set of technologies that help people talk to computers by voice. Recently, great progress has been made in the field of ASR using various deep learning approaches. Although technology giants have created advanced speech recognition engines for English, European and Asian languages, there is little research on the development of ASR systems such as Uzbek which is in the preliminary stage for creating language resources. This is due to the absence of a standard speech corpus of the Uzbek language, as well as dialectal differences [1]. A major challenge in developing an

ASR system is collecting sufficient speech corpus and data to train and test of the system. Other obstacle of building speech recognition models for Uzbek language is that it is an agglutinative language where the addition of numerous suffixes to the root of a word can create entirely new words[2], leading to a substantial increase in vocabulary size. With this problem in mind, a dataset consisting of approximately 11 hours of transcribed audio recordings of 16 speakers of different genders and ages was first collected. This data set is intended for the ASR task.

Related work.

There are several preliminary studies aimed at identifying speech in the context of the Uzbek language, of which [1] developed an initial Uzbek speech corpus consisting of 3500 sentences. A proposed ASR system based on HMM and DNN for the Uzbek language [3] allowed to obtain about 105 hours of transcribed speech data through the process of training the model on the basis of data consisting of 108,000 words. Similarly, the authors of [4] developed a system for recognizing spoken speech using 10 hours of transcribed audio. The works of [5] focus on voice command recognition systems based on a finite number of dictionaries.

It should be noted that the datasets used in these works are very limited and specialized for narrow application domains. Other existing Uzbek datasets are either

too expensive or not publicly available [6]. As a result, it limits the possibility of creating and testing an ASR system using a sufficiently large open-source Uzbek speech corpus. That is why it is of the first importance to develop an open-source Uzbek speech corpus of sufficient size.

Regarding the NLP research on Uzbek language, there has been a recent sharp increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for parts-of-speech tagging [7], removal of stopwords [8], Latin-Cyrillic transliteration [9], datasets for performing sentiment analysis [10], [11], stopwords [12], and text classification [13], as well as methodology for word game modeling [14].

Data collection.

Collecting high-quality speech data is crucial for building accurate speech recognition models. One approach to collecting speech data is to use platforms such as Mozilla Common Voice [15], which is a crowdsourcing project for collecting and validating speech data in multiple languages. Another approach is to collect data from audiobooks and their transcripts. Audiobooks are a good source of speech data as they provide a diverse range of speakers, speaking styles, and accents. Additionally, their transcripts provide a reliable source of text for aligning with the speech data, which is essential for training speech recognition models. By using a combination of these approaches, obtained diverse dataset that can improve the accuracy of their speech recognition models.

Methodology.

The ASR system aims to convert an input audio signal $x = (x_1, x_2, \cdots, x_T)$ of certain length T into a sequence of words or characters (i.e., labels) $y = (y_1, y_2, \cdots, y_N)$, $y_N \in \boldsymbol{V}$ from a specific vocabulary $\boldsymbol{V}$. These labels can be either words or characters. The output sequence is determined by finding the most probable sequence of labels $\acute{y} = \arg\max_{y \in V} p(y|x)$. The ASR system typically involves several steps such as pre-processing, feature extraction, classification, and language modeling (Figure 1).

The initial step of ASR is pre-processing which focuses on enhancing the quality of the audio signal by eliminating noise and filtering the signal. Additionally, in ASR, feature extraction is a critical step, where different techniques are employed to generate a set of coefficients or values. It is necessary for this step to be resistant to quality-related factors like noise or the echo effect. The objective of the classification model is to identify the spoken text from the input signal. It takes the pre-processed features as input and produces the corresponding output text. The objective of the classification model is to identify and interpret the spoken language in the input signal. After extracting the relevant features in the pre-processing stage, the model produces the corresponding text as output. On the other hand, the language

model (LM) plays a crucial role in capturing the semantic information and grammatical rules of a language. LM is vital for recognizing the output tokens generated by the classification model and for correcting the output text.

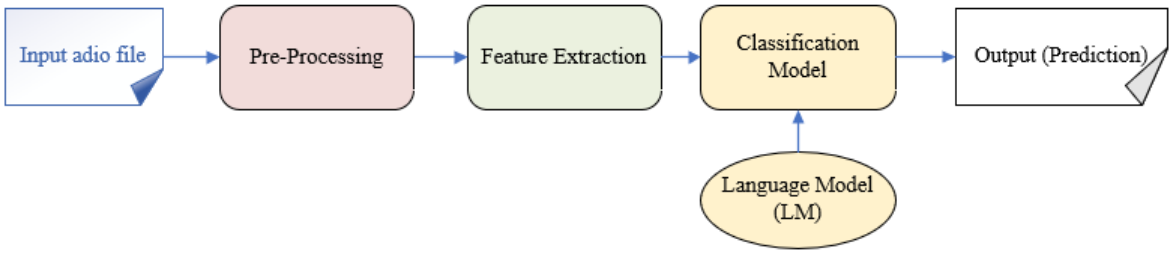

Figure 1. General overview of ASR system process.

Recurrent Neural Networks (RNNs) have been widely used in Automatic Speech Recognition (ASR) tasks due to their ability to model sequential data. Unlike feedforward neural networks, RNNs can capture the temporal dependencies between inputs by passing information from one time step to the next through a hidden state. This makes RNNs effective in modeling the dynamic nature of speech signals, which often exhibit complex temporal patterns.

In ASR, RNNs are typically used as acoustic models, taking in a sequence of acoustic features and outputting a sequence of probabilities over phoneme, grapheme, or word labels. One popular variant of RNNs is the Long Short-Term Memory (LSTM) network, which has been shown to effectively capture long-term dependencies in speech signals. However, RNNs are also known to have some limitations, such as difficulty in capturing long-term dependencies in sequences and the vanishing gradient problem. This has led to the development of other neural network architectures such as Convolutional Neural Networks (CNNs) and more recently, Transformer Networks, which have shown significant improvement in ASR tasks.

The series of experiments conducted on the ASR dataset using RNNs and CNNs models. After training the models, they evaluated their performance based on the word error rate (WER) metric. The results showed that the RNNs model achieved a WER of 21.6%, while the CNNs model achieved a WER of 23.5%. The lower WER score of the RNNs model suggests that it was more accurate in transcribing speech to text (Table 1). The experiments demonstrate the potential of both RNNs and CNNs models in ASR applications and highlight the importance of selecting the appropriate model for the specific task.

Table 1. Training results on the test dataset

Model Name	WER	Number of Layers	Number of Neurons	Batch size	Activation function
RNNs	21.6%	4	512	64	ReLU
CNNs	23.5%	5	256	32	Tanh

Conclusion.

In conclusion, we have developed a high-quality but smaller Uzbek speech corpus containing transcribed audio recordings spoken by numerous speakers. The corpus has been thoroughly checked by native speakers to ensure accuracy and reliability. Our ASR experiments demonstrated the effectiveness of both the RNNs and CNNs in transcribing Uzbek speech and model trained on the Uzbek speech corpus achieved 19.4% and 18.2% WERs on the validation and test sets respectively, using both RNNs and CNNs. We plan to continue increasing the size and quality of our dataset and conduct further ASR experiments by using Deep Learning approaches in the future. Our findings suggest that our Uzbek speech corpus and ASR models can serve as a valuable resource for future research in the field of Uzbek speech processing.

References

1. M. Musaev, I. Khujayorov, and M. Ochilov, "Automatic Recognition of Uzbek Speech Based on Integrated Neural Networks," in *Advances in Intelligent Systems and Computing*, 2021. doi: 10.1007/978-3-030-68004-6_28.
2. A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, "Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language," *Sensors*, vol. 22, no. 10, 2022, doi: 10.3390/s22103683.
3. M. Musaev, S. Mussakhojayeva, I. Khujayorov, Y. Khassanov, M. Ochilov, and H. Atakan Varol, "USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-87802-3_40.
4. M. Musaev, I. Khujayorov, and M. Ochilov, "Development of integral model of speech recognition system for Uzbek language," in *14th IEEE International Conference on Application of Information and Communication Technologies, AICT 2020 - Proceedings*, 2020. doi: 10.1109/AICT50176.2020.9368719.
5. M. Musaev, I. Khujayorov, and M. Ochilov, "The Use of Neural Networks to Improve the Recognition Accuracy of Explosive and Unvoiced Phonemes in Uzbek Language," in *2020 Information Communication Technologies Conference, ICTC 2020*, 2020. doi: 10.1109/ICTC49638.2020.9123309.
6. M. Sharipov and U. Salaev, "Uzbek affix finite state machine for stemming," *arXiv preprint arXiv:2205.10078*, 2022.
7. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekTagger: The rule-based POS tagger for Uzbek language," *arXiv preprint arXiv:2301.12711*, 2023.

8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.

9. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com

10. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com

11. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.

12. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.

13. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

14. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.

15. R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020.