INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES

https://journals.researchparks.org/index.php/IJHCS e-ISSN: 2615-8159 | p-ISSN: 2615-1898

Volume: 04 Issue: 8 | Aug 2022

Open-Source Speech Corpus and Initial Speech Recognition Experiments

A.A. Rakhmanov

chief specialist of the territorial administration of the Samarkand region of the Ministry for the Development of Information Technologies and Communications of the Republic of Uzbekistan

----------***----------

Abstract: We present a speech corpus and report preliminary results of automatic speech recognition (ASR) using a hidden deep neural network Markov model (DNNHMM) and an end-to-end (E2E) architecture.

Keywords: speech corpus, E2E, USC, ASR, USC, DNN-HMM, WER

We introduce an Open Speech Corpus (USC) in this publication that is intended to advance the study of Automatic Speech Recognition (ASR).

Researchers have long fantasized about utilizing speech to communicate with robots since it is the most natural form of human interaction. As a result, during the past few years. decades, ASR research has garnered a lot of interest. Annotated training datasets and multiple ASR designs have been introduced in particular. Unfortunately, the majority of databases are created for widely used languages like English, Spanish, and Chinese, whereas less widely used languages receive little attention. Consequently, there is a serious paucity of ASR technology research and development for less widely used languages.

Many datasets in less well-known languages have been generated to address the aforementioned challenge. for instance, to encourage study in the area of speech processing. Researchers in Kazakhstan have created open source Kazakh speech corpora that can be used to create speech synthesis and recognition software. to make more speech-enabled applications available and to facilitate speech exploration. Doumbouya et al. provided 150 hours of transcribed audio data for West African languages for illiterate users. Similar to this, other sizable multilingual speech corpora initiatives, including VoxForge, Babel, M-AILABS, and Common Voice, have been launched. The Uzbek language has not yet been included in these efforts, though.

There have been earlier attempts to identify Uzbek speech in the context of the Uzbek language. For instance, using a data set of 3500 utterances, Musaev et al. constructed an ASR system for geographic features. Similar to this, the authors used 10 hours of audio transcription to create a read speech recognition system. Works and are devoted to voice command recognition in situations with restricted vocabulary and colloquial numerals, respectively. It should be mentioned that the data sets utilized in these works were extremely constrained and tailored for certain application domains. Other Uzbek datasets already in existence are either unreasonably expensive or openly accessible. Therefore, it is crucial to create a sufficiently large corpus of public Uzbek speech.

https://journals.researchparks.org/index.php/IJHCS e-ISSN: 2615-8159 | p-ISSN: 2615-1898

Volume: 04 Issue: 8 | Aug 2022

a group of texts. First, we gathered Uzbek text data from a variety of sources. e-books of contemporary Uzbek literature, news portals, and a database of national laws are some examples. The texts, which cover a wide range of subjects like politics, finance, entertainment, and law, were automatically gathered using web crawlers. Additionally, we manually screened the collected texts to remove flaws in search robots, exclude objectionable information, and remove non-Uzbek sentences (for example, user privacy and violence). Sentences with terms taken from other languages, including English, have been saved. Finally, sentences involving numbers and sentences longer than 30 words have been eliminated. Over 100,000 sentences in total were written for the story.

Check the sound. We have created an additional one for examining audio recordings in order to guarantee the high caliber of the material that has been obtained. The checker bot delivers the examiner both the audio recording and the associated sentence, unlike the audio picker bot. The final speech corpus included audio and phrase pairs that had been marked as "correct." The audio recording was taken down and the decision was sent to the audio collection bot for re-reading for couples designated as "incorrect." We manually performed additional quality improvement techniques (such deleting long pauses, breaking the audio into several parts, and normalizing the audio) on pairs marked as "contains long pauses" or "poor quality" before adding the pairs to the final speech corpus. We preserved statements with background noises to better represent real-world occurrences in our data collection.

To show the reliability of the USC data set, we ran speech recognition tests. We created speech recognition models using DNN-HMM and E2E. based on our dataset and assessed them using the rate metrics for symbol error (CER) and word error rate (WER). Other language resources including lexicon, pronunciation patterns, and vocabulary were used instead of any outside data.

The CER and WER experimental findings for the test and validation sets are reported. All ASR models produce outcomes that are competitive. The E2E-Conformer, E2ETransformer, DNN-HMM, and then the E2E-LSTM model, in particular, produce the best results. We have seen that the Uzbek language, where test set absolute WER increases range from 7.7% to 12.6%, benefits from the incorporation of LM into E2E ASR. Additional absolute WER improvements on the test set range from 0.8% to 5.1% when velocity perturbation is applied to the E2E ASR models. While spectral augmentation boosts the performance of the E2E ASR models by 2.0% to 3.8% absolute WERs per test set, it has no effect on the DNN-HMM model.

Overall, E2E-Conformer obtained the lowest WER values, which were 18.1% and 17.4% in the validation and testing kits, respectively. The effectiveness of the USC dataset for developing ASR models is successfully illustrated by these findings..

References

1. Speechocean’s Uzbek speech corpus. http://en.speechocean.com/datacenter/details/1847.html, accessed: 2021-05-21

https://journals.researchparks.org/index.php/IJHCS e-ISSN: 2615-8159 | p-ISSN: 2615-1898

Volume: 04 Issue: 8 | Aug 2022

2. Uzbek language. https://en.wikipedia.org/wiki/Uzbek language, accessed: 2021-05-20
3. Voxforge. http://www.voxforge.org/, accessed: 2021-05-11 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29(6), 82–97 (2012)
4. Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE ACM Trans. Audio Speech Lang. Process. 22(10), 1533–1545 (2014)