Speech Recognition System in Kazakh language

Capstone Report

Galymzhan Saginbayev

Nazarbayev University
Department of Electrical and Computer Engineering
School of Engineering and Digital Sciences

This project report was created on TexStudio editing platform using LaTeX. All the figures were drawn using draw.io online software tool.

NAZARBAYEV UNIVERSITY

Electrical and Computer Engineering
Nazarbayev University
http://www.nu.edu.kz

Title:
Speech Recognition System in Kazakh language

Theme:
Capstone Project Final Report

Project Period:
Spring Semester 2023

Project Group:
n/a

Participant(s):
Galymzhan Saginbayev

Supervisor(s):
Mehdi Shafiee

Copies: 1

Page Numbers: 10

Date of Completion:
April 26, 2023

Abstract:

Machine learning is one of the most popular and developing fields in the modern world, and it has an increasing impact on people's daily life. It is one of the many study methods designed to teach computers without the external help. One of the most important inventions in the field of machine learning is the Automatic Speech Recognition (ASR). ASR is implemented in almost every smartphone, and they are used as voice assistants. The main objective of this project is to design an effective speech recognition system that will be able to understand Kazakh language and convert it to text based on a machine learning algorithm. It was decided to choose this topic because despite the fact that the field of machine learning is developing rapidly, there have not been any ASRs designed yet in the country that could be used on a daily basis. Nowadays, smart devices have become an integral part of majority people's lives, and the effects of voice assistants are also increasing due to the high development of machine learning. However, these technologies are only available in widely spoken languages, such as Chinese, English, Russian, and Spanish. Therefore, the main aim of this project is to design ASR for specifically Kazakh population.

Contents

Preface

Nazarbayev University, April 26, 2023

Galymzhan Saginbayev
<galymzhan.saginbayev@nu.edu.kz>

Chapter 1

Introduction

According to Fradkov [1], the origin of this field took place in the middle of 20th century, when a group of scientists leaded by Frank Rosenblatt constructed a machine which was able to identify the characters. There is sufficient amount of research conducted in this field. For instance, there is an article written by Ravanelli, Parcollet, and Bengio [2] about a Pytorch-Kaldi toolkit designed specifically for the speech recognition system. Article was published in 2019. Authors claim that their algorithm is the combination of Pytorch and Kaldi toolkits. Basically, acoustic models are taken from Kaldi, whereas operations such as computation and decoding are performed through Pytorch. They also stated that they tried to make the uncomplicated toolkit, so that the ordinary users could easily understand the model. Moreover, it was claimed to be flexible, as it allows users to add various customizations. Another important thing to mention is that the toolkit was written in python language. Even though the project is available on the internet, it is still in the early stages of development.

There is another useful paper written by Mamyrbayev et al. [3]. regarding the end-to-end ASR model based on Recurrent Neural Network Transducer (RNN-T). RNN-T is an end-to-end architecture designed mainly for speech recognition systems [4]. According to Mamyrbayev et al. [3], RNN-T is considered as the most suitable and used model for speech recognition. They found that Recurrent Neural Network Transducers were more effective compared to the hybrid model based on Word Error Rate metrics. They compared their model with other models, and the results were promising. Iakushkin et al. [5] describe a development of ASR in Russian language based on DeepSpeech which is an architecture designed by Mozilla company for speech recognition purposes. Model was trained by using the data taken from YouTube videos. After training the model, authors tested it on a collection of different audio recordings. Before conducting an experiment, authors expected Word Error Rate to be below 30% from the model. According to the test results, the lowest WER was approximately 18% and the average value was

22%. The main reason behind choosing DeepSpeech for the model was extremely low WER of under 7% based on the experiment conducted six years ago by Mozilla. According to Iakushkin et al. [5], in general, model requires a set of data with a total duration of about 1000 hours.

There is another experiment conducted by Mamyrbayev et al. [6] in 2022, where authors constructed an end-to-end ASR system based on a Transformer model. Authors claim that main advantages of this model is that it learns very quickly, and also it lacks the sequential operation, which is similar to RNN. They also found that this model can be highly accurate even with the limited amount of data. Moreover, they made a new model by combining Transformer model with a Connectionist Temporal Classification (CTC), which showed better results.

Amirgaliyev et al. [7] made a ASR system by using a pre-trained model designed for Russian language. They said that using a pre-trained Russian model is a good idea due to the fact that Kazakh and Russian letters are similar to each other. They also claimed that this method could partially fix the problem regarding the lack of data in Kazakh language.

The main issue that this paper tries to fix is the communication problems caused by monolingualism. There are many people in Kazakhstan who are still not familiar with English and Russian languages. Smaiyl [8] conducted a survey in 2017 where he found that 83% of people spoke Kazakh language, while only 22% of participants were trilingual. In addition, Zhuravleva and Agmanova conducted a survey consisting of only immigrant students, where they found that just 8% of them spoke Russian language. Therefore, this project will not only help Kazakh speaking people, but it might also be extremely useful for people who are willing to learn Kazakh language. Speech recognition system could be used in many aspects of life. For example, it could be used as a voice assistant in smartphones or as virtual assistants for big companies by directing the customers to an appropriate expert of that field who is familiar with a problem. Doctors or lawyers would not have to take notes from meetings, since the computer would automatically be able to convert speech to text. ASR could be extremely useful for disabled group of people by helping them order food, make a call, and get to a destination. Last feature could also be helpful for the drivers. NHTSA [9] provides yearly statistics regarding vehicle crashes and drivers in America. According to their 2020 report [9], nearly 8% of distractions were caused by phones. It means that drivers would be more focused on the road instead of looking at a phone by the help of voice assistants. Car crashes is one of the major problems in Kazakhstan that should not be ignored. Many car accidents tend to happen due to drivers getting distracted. For instance, over 13000 car accidents were reported in 2020 which resulted in nearly 20 000 people getting injured [10].

It is highly important to begin constructing ASR models now in order to catch up with other ones. It would be a huge step in developing a machine learning

field in the country, since Kazakhstan is not one of the most developed countries in terms of technology. One of the main goals of the country is to become a part of the 30 most developed countries in the world. Therefore, progress must be made in all spheres for Kazakhstan to become one of the most developed countries. Moreover, Nursultan Nazarbayev, former president of Kazakhstan, stated that improving the quality of life of the country's population is the main aim of the Kazakhstan 2050 strategy [11]. As it was mentioned earlier, current project can be beneficial in many aspects of life for many people, which means that the project meets the most important national priority according to the former president.

If the model becomes publicly available, it can potentially be implemented into smartphones as an application. It would most likely be used for translating or note taking purposes. However, if enough effort and contributions are put into this project, it can even be used as a voice assistant. This project is important because there are no ASR systems with high accuracy for Kazakh language available right now. Even though there have been some models constructed according to the articles mentioned in this paper, these are still in early development stage, and there is no certainty that they will be widely used in the future. Project has several benefits, including:

• Efficiency and productivity: Having an opportunity to dictate text instead of typing can noticeably increase the work rate. For instance, it will be extremely useful for doctors, lawyers, and experts of the big companies. Doctors could use it for making health records, while lawyers could use it for taking notes during from courtrooms or meetings.

• Accessibility: Project will be helpful for people with disabilities who have vision and hearing issues and mobility limitations by helping with their daily activities. It can also benefit people with dyslexia or other learning disabilities. Moreover, it can be helpful for immigrant people who want to learn Kazakh language, as the model could potentially be used as a translator in the future. Reaching the destination for the drivers would be much easier since the voice assistant would tell them when to turn or stop.

• Safety: Project can be helpful for regular users who usually spend most of the time with the keyboard, as it reduces the risk of strain injuries associated with frequent typing.

• Multitasking: Users will be able to dictate text while performing other tasks, such as driving, cooking, and cleaning. As a result, time can be spent a lot more efficiently.

• Cost saving: Organizations can also benefit from Automatic Speech Recognition by saving money on labor costs. Overall, project can potentially provide many benefits to organizations and users of any age group. Algorithm will most likely become even more accurate and versatile, since technology, especially the field of machine learning keeps improving. Thus, ASRs becoming one of the most

valuable tools can be considered a matter of time.

In 2012, former president of Kazakhstan announced the Kazakhstan 2050 strategy [12]. As it was mentioned earlier, even though Kazakhstan aims to be part of the 30 most developed countries in the world, the most important objective of this strategy for Kazakhstan is to improve the quality of life for people, and this project can help to achieve this strategy. It can be used almost in any area. Moreover, despite its possible low effectiveness caused by the lack of data required for training, it does not have any drawbacks that could threaten human life or prevent to achieve the national priorities. The success of this project is most likely a matter of time due to the rapid development of ML and increasing amount of data on the internet. Human resource development is also one of the priorities of the country, which includes education and health care [12]. As it was noted before, this model can be helpful for immigrants who are willing to learn the national language and for people who work in health care industry. Its speech to text and text to speech features can be used for taking health records. According to Maydirova et al. [13], there every country should focus on six global trends which include the technological development. ASR could bring a huge impact by digitalizing the country.

Chapter 2

Methodology

2.1 Methods and Procedure of Data Collection

Model was designed in a traditional way, and its architecture can be seen in Figure 2.1 [14]. Firstly, input data is entered to the model through microphone or other recording device. Then, the audio is pre-processed to remove background noise and transformed into a set of features. After that, a statistical model is used to map these features. Usually, a language model is used to predict the next sequence of words. The next step is the decoding, where the system selects the words based on the language model. Finally, model gives the output in a text format.

The data used for the project was collected from the Institute of Smart Systems and Artificial Intelligence (ISSAI). Dataset contains more than 600,000 audio recordings with a total duration of approximately 1200 hours [15]. Recordings include speech segments in which Kazakh speakers engage in the practice of switching between Kazakh and Russian languages [15]. It is noteworthy that the dataset can be openly accessed by researchers and industry professionals on their website [15]. ISSAI also included the demographic information, such as age, gender, and region, which can possibly improve the accuracy of the system.

2.2 Methods and Procedure of Data Analysis

There was no need in preprocessing data since the dataset already contained high-quality audio recordings with removed background noise and converted to a suitable format for training. Then, the data was used to train the model by using a Hidden Markov Model (HMM). Moreover, there are various techniques that can improve the accuracy of the model, such as data augmentation and transfer learning. Word Error Rate (WER) was used to evaluate the performance of the system. The model was tested on other data in order to check its accuracy.

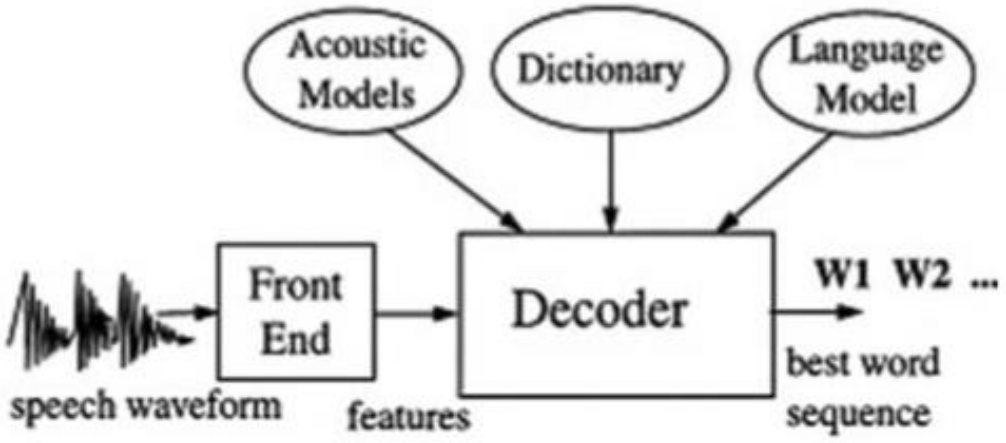

Figure 2.1: Speech Recognition System design [14]

2.3 Ethical Issues

Project does not violate any ethical standards. In case if more data is decided to be collected in a real-world scenario, the privacy and confidentiality of all participants will be ensured. According to the figures provided on ISSAI website, collected data for the model was diverse and representative, and the speech recognition is not biased towards groups of any age, gender, and region. [15].

Overall, this methodology involved collecting preprocessed data from a public source. Standard metrics was used during the process of evaluating the accuracy of the model. Ethical issues were carefully considered throughout the study.

Chapter 3

Results and Discussions

The results of this study indicate that the model trained on Kazakh Speech Corpus 2 dataset achieved a promising level of performance in converting Kazakh speech to text. In majority of the times, the model was able to accurately recognize speech segments, such as phrases and short sentences, with a Word Error Rate (WER) of approximately 15%. However, there was a noticeable drop in accuracy of the model after it received a speech in a noisy environment. This means that the model is not yet capable of recognizing a speech in challenging conditions.

One of the biggest benefits of the KSC2 dataset is that it contains speech with switching between Kazakh and Russian, which is a common practice in Kazakh conversations as it was mentioned earlier. Having an ability to recognize such data is an important feature for the speech recognition system.

There are a few limitations of this project that should be addressed. One of them is that there was not any additional equipment used for the model which can limit the model's performance. Additionally, impacts of age, gender, accent, and dialect on the model's accuracy were not explored in the project.

Chapter 4

Conclusion

In conclusion, automatic speech recognition system was constructed for this project which is able to transcribe Kazakh speech into text. Paper describes the benefits and importance of the Automatic Speech Recognition (ASR) system for individuals who prefer speaking in Kazakh. Other studies on this topic were also analyzed during the project. Overall, the model showed the Word Error Rate (WER) of 15%. Future work is required to improve the performance of the model.

Bibliography

[1] Alexander L Fradkov. "Early history of machine learning". In: *IFAC-PapersOnLine* 53.2 (2020), pp. 1385–1390.

[2] Mirco Ravanelli, Titouan Parcollet, and Y. Bengio. "The Pytorch-kaldi Speech Recognition Toolkit". In: May 2019, pp. 6465–6469. DOI: 10.1109/ICASSP.2019.8683713.

[3] Mamyrbayev Orken et al. "End-to-End Model Based on RNN-T for Kazakh Speech Recognition". In: June 2021, pp. 163–167. DOI: 10.1109/ICCCI51764.2021.9486811.

[4] *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021. URL: https://doi.org/10.1109/ICASSP39728.2021.

[5] Oleg Iakushkin et al. "Russian-Language Speech Recognition System Based on Deepspeech". In: Dec. 2018.

[6] Mamyrbayev Orken et al. "A study of transformer-based end-to-end speech recognition system for Kazakh language". In: *Scientific Reports* 12 (May 2022). DOI: 10.1038/s41598-022-12260-y.

[7] Darkhan Kuanyshbay, Yedilkhan Amirgaliyev, and Olimzhon Baimuratov. "Development of Automatic Speech Recognition for Kazakh Language using Transfer Learning". In: *International Journal of Advanced Trends in Computer Science and Engineering* 9 (July 2020), pp. 5880–5886. DOI: 10.30534/ijatcse/2020/249942020.

[8] Meirim Smaiyl. *Skolko grajdan RK vladeyut kazahskim yazyikom*. [Online]. 2018. URL: https://tengrinews.kz/kazakhstan_news/skolko-grajdan-rk-vladeyut-kazahskim-yazyikom-340682/.

[9] NHTSA. *Driver Electronic Device Use in 2020*. [Online]. 2021. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813184.

[10] Committee on Legal Statistics and Special Accounts of the General Prosecutor's Office of the Republic of Kazakhstan. *n 2020, the number of road accidents in Kazakhstan decreased by 18.7%*. [Online]. 2021. URL: https://www.gov.kz/memleket/entities/pravstat/press/news/details/v-2020-godu-kolichestvo-dtp-v-kazahstane-snizilos-na-187?lang=ru.

[11] Aktoty Aitzhanova et al. *Kazakhstan 2050: Toward a modern society for all*. Oxford University Press New Delhi, 2014.

[12] Johannes F Linn. "Kazakhstan 2050: Exploring an ambitious vision". In: *Global Journal of Emerging Market Economies* 6.3 (2014), pp. 283–300.

[13] AB Maydirova et al. "Strategic priorities of Kazakhstan innovative economy development". In: *Opción: Revista de Ciencias Humanas y Sociales* 27 (2020), p. 44.

[14] Jayashree Padmanabhan and Melvin Jose Johnson Premkumar. "Machine learning in automatic speech recognition: A survey". In: *IETE Technical Review* 32.4 (2015), pp. 240–251.

[15] Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. "KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus". In: *Proceedings of the INTERSPEECH, Incheon, Republic of Korea* (2015), pp. 18–22.