ANNOTATION
of Darkhan Kuanyshbay

PhD thesis on speciality 6D070400 - "Computer Systems and Software" on the topic "Development of methods, algorithms of machine learning and mobile applications for Kazakh speech recognition"

Relevance

Automatic speech recognition (ASR) is a dynamic direction in the field of artificial intelligence. Over the past half century, a significant progress has been made in this area - there are many commercial applications that make investments in this area justified and profitable. Among such applications, first of all, it may be noted the introduction of call centers or IVR systems (Interactive Voice Response) - systems for automatic access to information, bypassing the operator. AT modern call centers questions are formulated by the user on natural language, and the answer is synthesized by the computer also in the language user. The introduction of call centers has freed up a huge number of operators and improved the service quality in many airports and train stations. Automatic speech recognition systems are widely used in medical research requiring input when hands the operator is busy (x-ray), or when you want to manage autonomous apparatus for the study of internal organs. Even filling out the medical cards by mid-level personnel in advanced medical facilities is done using voice.

An important area of application of automatic recognition systems and speech synthesis is helping people with disabilities like problems with musculo-skeletal system and impairment of vision (assistive technology).

It should be noted that automatic speech recognition systems in Kazakhstan almost never applied, which leads to an active researches and explorations of the area for developers. The main reason behind poor investigation and research in area of speech recognition for Kazakh language is lack of speech data. As it was observed in popular languages like English, Spanish and Chine, good quality ASR systems require tremendous amount of data. Popular speech corpuses like TIMIT or Switchboard contain massive amount of transcribed audio recordings with various types of speeches, like telephone speeches, conversational speech or clean microphone speeches. In Kazakh language there are almost no decent speech corpuses available on web-sources. The available ones are usually not free for usage and certainly are not enough to obtain powerful and effective ASR models. In order to build the decent speech corpus for ASR system, it requires a lot of time, well-structured environment and reliable monitoring system. However, to completely get

rid of the issue related with the data deficit, the neural network structure and approach should be considered as well.

Speech data in most of the languages that have a low resource doesn't even exist. Therefore, producing speech corpora is very challenging and requires tremendous amount of time. Kazakh language due to its lack of popularity considered to be low-resource language.

The communication between humans can be in different forms like speech, pictorial language, gestures, sign language etc. Among these forms, communication using speech is considered to be more effective and popular. This leads to the fact that human computer interaction should be handled using speech communication. Therefore, this fact highlights the importance of developing Automatic Speech Recognition system.

The performance of ASR system is directly dependent on the quality of speech data. However, speech corpora can be based on general language or domain specific language. Using a dataset which is domain specific has the benefit of growing the capacity of recognition process. Moreover, utilizing the real speech of users as a dataset raises the ASR's overall productivity.

It is very significant and crucial factor to find the adequate speech dataset while constructing the ASR system for low-resource languages like Kazakh. However, the non-existence if required amount of speech data for low-resource languages is obvious. Thus, the tools and instruments for collecting the speech data can play a significant role in building speech recognition systems.

Aims and objectives of research.

In order to construct the proper ASR system avoiding the problem of data deficit our main objectives are the following:

- To construct a well-designed environment for speech data collection using a web platform
- To develop speech synthesis model in order to make an automatic speech collection system for small sentences
- To collect the significant amount of speech data with transcriptions for Kazakh language
- To post-process the collected data and structure the files for neural network operation
- To construct the neural network using Recurrent neural networks based on Connectionist Temporal Classification loss function
- To construct the multilingual technique with Russian language by using transfer learning based approach

Object of research

The research is focused on the methods and techniques of automatic speech data collection for speech recognition systems.

Research methods

Research was conducted by analyzing and interpreting the existing results of state of the art works in speech recognition, speech synthesis, natural language processing fields defining the advantages and disadvantages. The defined objectives were achieved by applying the machine learning algorithms, recent recurrent neural network achievements and sequence to sequence modeling techniques.

Scientific novelty

The novelties that were obtained during the execution of objectives are the followings:

- The application of a novel approach of speech data collection for any language by building reliable website with well-designed monitoring system and control
- The construction of an automatic post-processing step for speech data after collection process, that structures the speech and transcription with respect to training of CTC based neural network
- The training of neural network using multilingual approach by transferring the knowledge obtained from pre-trained Russian language model
- The automatic alignment of speech input sequence to transcription output sequence

The scientific statements are to be defended

- The platform of automatic speech data collection methods based on speech synthesis modeling
- The automatic pre-processing and structuring the speech data with corresponding text transcription
- The method of Connectionist Temporal Classification algorithms to train the recurrent neural network
- The methods and algorithms of building a transfer learning approach using Russian speech recognition model
- Comparative analysis of Long-Short Term Memory (LSTM) and Bidirectional LSTM recurrent neural network based models

Approbation of the work

The results of the research were reported and discussed at international conferences: “III International Scientific Conference “Informatics and Applied Mathematics” dedicated to the 80th anniversary of professor R.G. Biyashev and 70th anniversary of professor M.B. Aidarkhanov”(Kazakhstan, Almaty, 2018); “IV International Scientific and Practical Conference “Informatics and Applied Mathematics” dedicated to 70th anniversary of processor Biyarov T.N., Waldemar V. And 60th anniversary of professor Amirgaliyev E.N.” (Kazakhstan, Almaty, 2019); “Trends of Modern Science” (UK, Sheffield, 2019)

Publications

On the dissertation topic 14 published works are presented, including:

- 1 monograph
- 2 articles in international journal indexed by Scopus
- 4 articles that meet the requirements of the higher Attestation Commission of the Ministry of Education of Science of the Republic of Kazakhstan
- 4 articles published in International Conferences
- 2 certificated of authorship/patents
- 1 foreign journal publication

1. Амиргалиев Е.Н., Куанышбай Д.Н., “Вербальный робот”, Алматы: ИИВТ, 2020. -143 с, ISBN 978-601-08-0090-8
2. Kuanyshbay D.N., Amirgaliyev Y, Shoiynbek A., Yedilkhan D., “Automatic speech recognition system for kazakh language using connectionist temporal classifier”, Journal of Theoretical and Applied Information Technology, Vol. 98 No.04, PP 703-713, ISSN: 1992-8645(print), ISSN: 1817-3195(online) (SCOPUS: CiteScore:1.2, Quartile: Q3, percentile: 37)
3. Kuanyshbay D.N., Amirgaliyev B., Kutubayeva M., Baimuratov O., “Development of Automatic Speech Recognition for Kazakh Language using Transfer learning”, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9 No.04, ISSN: 2278-3091(online) (SCOPUS: CiteScore:1.2, percentile: 36, Quartile: Q3)
4. Kuanyshbay D.N., Kozhakhmet K, Shoiynbek A., “Comparison of two speech signal features for deep neural networks to classify emotions”, Вестник КазНИТУ. Том 132 №2 2019, с.343-350 ISSN 1680-9211 (Импакт-фактор по Казахстанской базе цитирования за 2017 год = 0,045)
5. Kuanyshbay D.N., Amirgaliyev Y.N., Musabaev T.R., Kenshimov Sh., “Разработка голосовой идентификации диктора с использованием статистик контура основного тона для применения в робото-вербальных системах”,

Вестник КазНИТУ. Том 132 №2 2019, с.424-433 ISSN 1680-9211 (Импакт-фактор по Казахстанской базе цитирования за 2017 год = 0,045)

6. Куанышбай Д.Н., Амиргалиев Е, Едилхан Д., Баймуратов О., “Об одном улучшенном подходе автоматического распознавания казахской речи с использованием трансферного обучения”, Вестник КазНИТУ. Том 141 №5 2019, с.183-189 ISSN 1680-9211 (Импакт-фактор по Казахстанской базе цитирования за 2017 год = 0,045)
7. Kuanyshbay D.N., Kozhakhmet K, Shoiynbek A., “Various Languages impact on the problem of emotion recognition in speech”, Вестник КазНИТУ. Том 141 №5 2019, с.182-185 ISSN 1680-9211 (Импакт-фактор по Казахстанской базе цитирования за 2017 год = 0,045)
8. Kuanyshbay D.N., Amirgaliyev Y.N., Kozhakhmet K, Shoiynbek A., “Comparison of optimization algorithms of connectionist temporal classifier for speech recognition system”, "Informatyka, Automatyka, Pomiary w GospodarceiOchronieŚrodowiska" - IAPGOS, Vol. 9 No.3, PP 54-57, ISSN: 2083-0157(print), ISSN: 2391-6761 (online)
9. Куанышбай Д.Н., Кожахмет К.Т., Шойынбек А., “Создание корпуса эмоций на казахском и русском языке для голосового распознавания эмоций”, MATERIALS OF XV INTERNATIONAL RESEARCH AND PRACTICE CONFERENCE «TRENDS OF MODERN SCIENCE - 2019» Vol. 14, PP 9-11, Science and Educational LTD, Sheffield, UK ,2019 ISBN 978-966-8736-05-6
10. Kuanyshbay D.N., Kozhakhmet K, Shoiynbek A., “Comparison of classification algorithms svm vs logistic regression for detecting crime”, МАТЕРИАЛЫ III Международной научной конференции «Информатика и прикладная математика», посвященная 80-летнему юбилею профессора Бияшева Р.Г. и 70-летию профессора Айдарханова М.Б. 2018 года, Алматы, Казахстан, Часть 2 , с 37-41, ISBN 978-601-332-165-3
11. Kuanyshbay D.N., Amirgaliyev Y, Shoiynbek A., “Speech recognition preprocessing, background removal”, МАТЕРИАЛЫ III Международной научной конференции «Информатика и прикладная математика», посвященная 80-летнему юбилею профессора Бияшева Р.Г. и 70-летию профессора Айдарханова М.Б. 2018 года, Алматы, Казахстан, Часть 2 , с 7-18, ISBN 978-601-332-165-3
12. Куанышбай Д.Н., Амиргалиев Е.Н., Шойынбек А., “Построение языковой модели для казахского языка на основе рекуррентных нейронных сетей”, МАТЕРИАЛЫ IV международной научно -практической конференции "Информатика и прикладная математика", посвященной 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию

профессора Амиргалиева Е.Н. 2019, Алматы, Казахстан Часть 1 с. 401 – 406, ISBN 978-601-332-384-8
13. Куанышбай Д.Н., Баймуратов О., “Алгоритм синтеза речи казахского языка”, Свидетельство о внесении сведений в государственный реестр прав на объекты охраняемые авторским правом. Вид объекта авторского права: программа для ЭВМ. Запись в реестре №15029 от 10 февраля 2021 года
14. Куанышбай Д.Н., Амиргалиев Е.Н, Козбакова А.Х., “Автоматизированная система формирования корпуса речевых данных”, Свидетельство о внесении сведений в государственный реестр прав на объекты охраняемые авторским правом. Вид объекта авторского права: программа для ЭВМ. Запись в реестре №15236 от 17 февраля 2021 года

Structure of the dissertation

The thesis consist of Introduction, 5 chapters, conclusion and reference list. The volume of the thesis consist of 120 pages, 30 figures, 124 references and 10 appendixes.

The first chapter consist of the overview of a speech signal processing itself and speech signal processing techniques. It provides an overall information about types of noises like wide band noise, interfering speeches and periodic noises. Moreover, chapter shows and illustrates the speech enhancement system (Figure 1) and provides an analysis of various types of speech enhancement techniques.

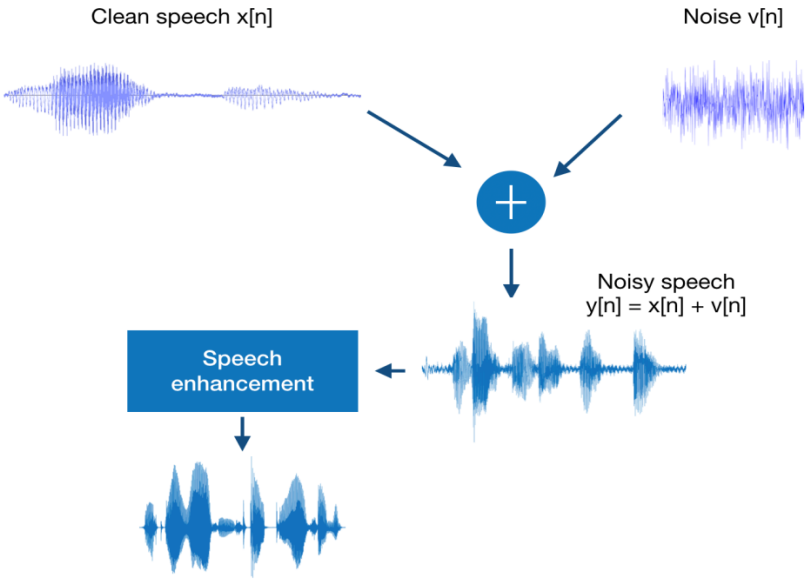


Figure 1. Basic speech enhancement system

Different types of noise cancellation and speech improvement methods were presented like Linear Predictive Coding, Signal subspace method, DFT based methods etc.

The second chapter provides an analysis of an existing automatic speech recognition systems. It first covers the Hidden Markov Model based acoustic model built using probabilistic methodology. In recognition process, the unseen words are recognized by comparing it with already seen templates and finding the closest one. But these templates cannot match the acoustical changes. Therefore, acoustical models are built by probability distribution over the acoustics. The most basic way is by using Gaussian distribution which finds the representation parameters.

Speech recognition task initially considered as a statistical classification task. Classes defined as words sequence W from the large vocabulary set and input defined as the features of the speech X. The equation looks like as following:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{8}$$

Secondly, chapter overviews the ANN based acoustic models. One of the main advantages of ANN over other modeling methods in speech recognition it can approximate the nonlinear dynamic systems. Speech is a nonlinear signal made by nonlinear system.

ANN is basically an interconnection of computational elements (neurons) and this nonlinear system is distributed through the network. The basic nonlinear neuron model is shown in Figure 2.

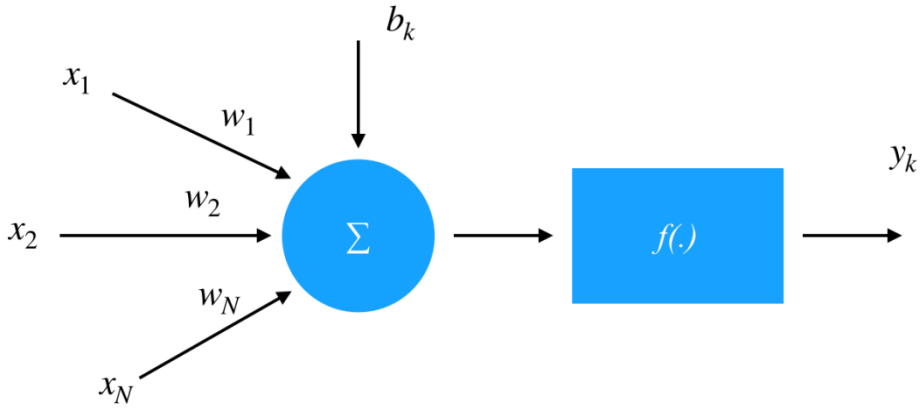


Figure 2. Nonlinear neural model

At the early stage, usage of ANN for speech recognition was successfully attempted on recognizing simple digits, few phonemes and words using multilayer perceptron.

Thirdly, chapter follows with the Deep Belief neural network based acoustic models.

Deep belief neural network (DBNN) has been discovered as powerful and efficient for many machine learning problems as well as for acoustic modeling in HMM/ANN based speech recognition system. At first, DBNN was presented for acoustic modeling tasks, because it had much higher capacity for modeling than regular GMM. Moreover, DBNN also showed very effective training ability that merges the unsupervised learning for feature discovery and supervised learning for optimization and fine-tuning of the features. Another reason that DBNN is efficient is that the low-level feature characteristics are computed in lower layers and highly nonlinear structure of the input are taken care of in higher layers. It can be similar to human speech recognition that uses a lot of layers for features extractions.

Chapter three reviews the novel approaches on training the neural network like Connectionist Temporal Classification and RNN transducer based end-to-end models. CTC based speech recognition model trainings were analyzed and explained. CTC introduces the blank symbol in order to match two sequences together calculating the path probability. After that path aggregation algorithm is executed (Figure 3).

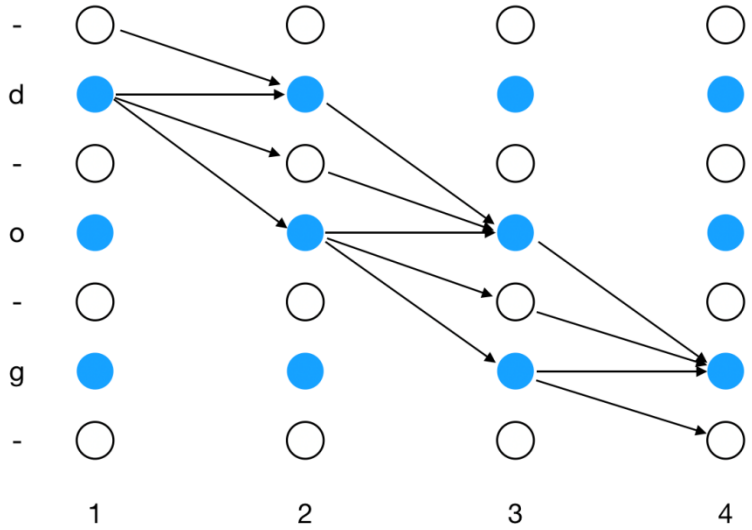


Figure 3. Paths example for label “dog”

Next section introduces the RNN transducer based end-to-end model. RNN-transducer based model in term of the structure share a lot of similarities with CTC

based model. For example, they both share the same loss function; they both solve the problem of manual segmentation between input sequence and output sequence; they use a "black symbol" element; they both estimate the probabilities of all paths and aggregates paths to obtain the label sequence. But, path generation and path probability estimation processes are very different. Chapter also defines the advantages and disadvantages of this algorithm over CTC based model.

RNN-transducer based model contains three important components: Network for transcription ($F(x)$); prediction network ((P(y, g)); joint network ($J(f,\ g)$). The construction of RNN-transducer model is illustrated in Figure 4.

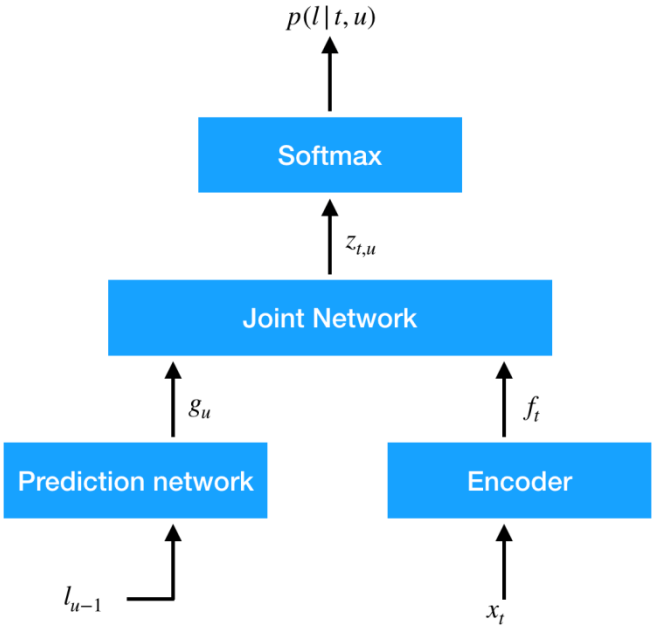


Figure 4. RNN-transducer structure

Chapter four provides an information about constructing an automatic speech collection system built on a web application (Figure 5).

This application was developed with the latest available technologies that are suitable for any device or computer in the market today. The backend of an application was done using yii2-framework in php7.2 language. The architecture development for this project is based on Model-View-Controller (MVC) template.

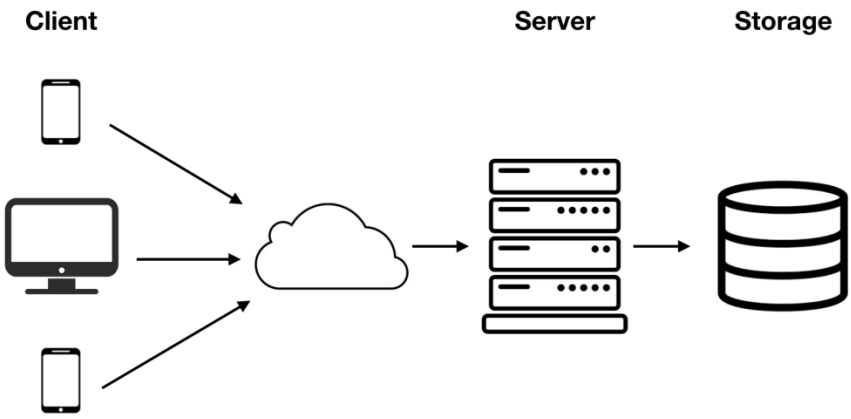


Figure 5. The representation of system architecture

As a HTTP server we utilized an engine NGINX. For managing databases in the application, we have used the language based on MySQL. The control panel was handled by phpMyadmin and heidiSql. For code editing we utilized the phpstorm technology. Each web-page in applications has the ability to change the resolution based on the current device.

The web application consist of a speech synthesis model that automatically pronounces the small sentences extracted from uploaded book by admin user. This speech synthesis model was trained on Kazakh literature books using Convolutional Neural network (CNN) (Figure 6).

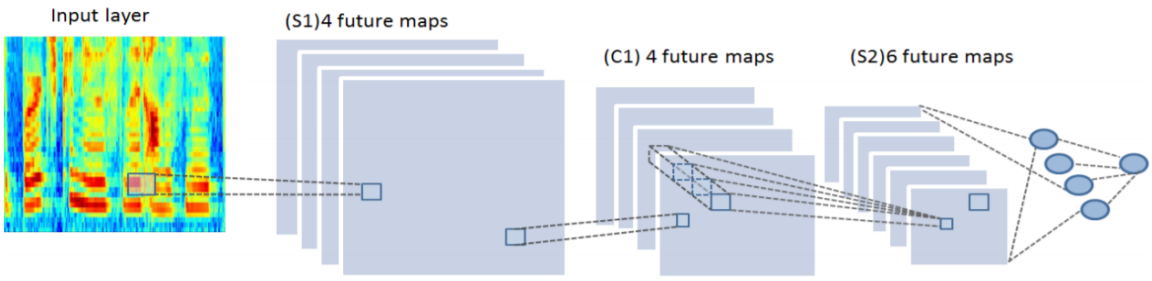
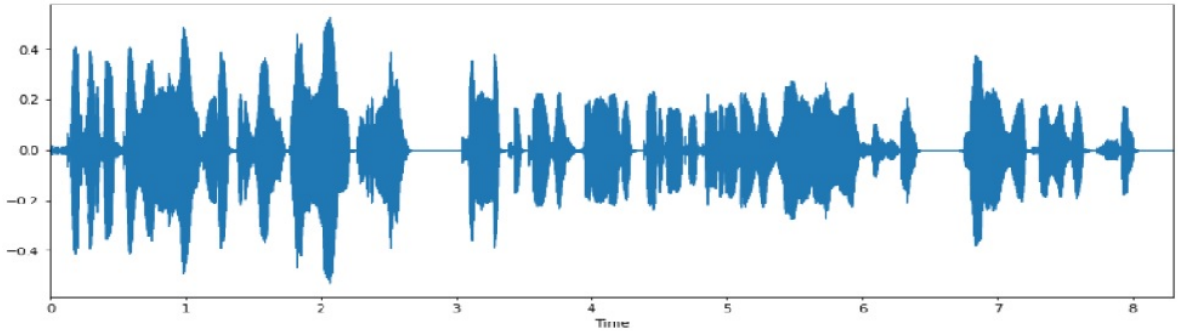


Figure 6. Architecture of a multidimensional CNN network

When building speech synthesis systems, one of the most important tasks is to segment and label databases of speech signals into semantically and phonetically significant units of speech, and in our particular case, these are phonemes. The resulting segments are stored in the database and used for machine learning of acoustic models in the integrated system, with subsequent generation of the speaker's voice in the text-to-speech synthesis system. One of the specific methods of working

with trained systems is to set the system parameters for a specific language and select the alphabet. Since the Kazakh speech synthesizer model uses phonemes as input alphabet symbols, it was necessary to solve the problem of transcribing texts according to the rules of grapheme to phoneme, taking into account the phonetic features of the Kazakh language. The task was solved by creating a phonetic transcription module, and then the prepared text-training sample was transcribed.

Thus, a training experimental base was created, consisting of 3500 sentences, and each sentence corresponds to an audio file in wav format with a sampling frequency of 22050 Hz. After that, the system module was activated, which is responsible for receiving spectrograms of audio files, on the basis of which deep learning of networks takes place (Figure 7).



білім беруде, денсаулық сақтау мен көлік саласында болады.

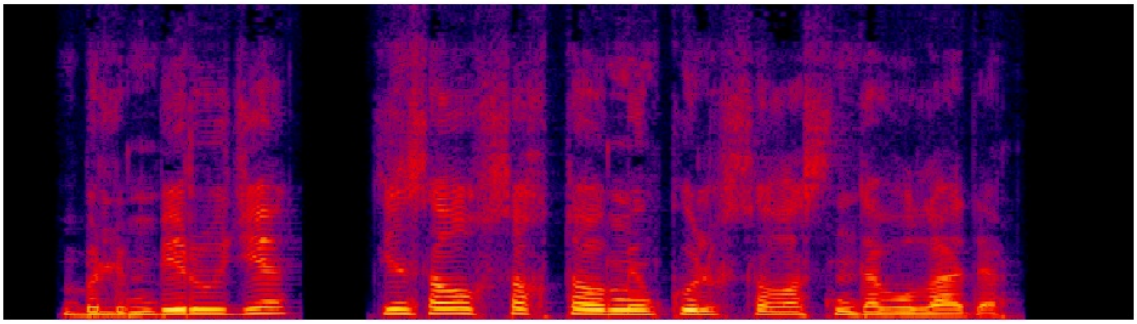


Figure 7. Sample offer from the training database

The next section, overviews the web application itself describing all its functionalities and features.

The system consists of two important roles to complete the whole process of data collection. It is an admin user (only 1-2 users) and large number of clients.Admin is eligible to add any number of books in any format, delete and edit. Admin can view the completeness percentage of a particular book and can listen to all recording by clients. He can list the recordings of any book as well as any speaker. He has the ability to delete the unwanted or corrupted recordings and email the

speakers who make the mistake. After the whole recording process is finished the admin can download all records. The downloaded zip file already structured with folders of speakers and all necessary transcription with corresponded audio files.

Name	Date Modified	Size	Kind
0_Farukh Iskalinov.txt	4/24/20	70 bytes	Plain Text Document
0_Farukh Iskalinov.wav	4/24/20	352 KB	Waveform audio
1_Farukh Iskalinov.txt	4/25/20	13...ytes	Plain Text Document
1_Farukh Iskalinov.wav	4/25/20	639 KB	Waveform audio
2_Baimolda Aray.txt	4/13/20	11...ytes	Plain Text Document
2_Baimolda Aray.wav	4/13/20	524 KB	Waveform audio
3_Farukh Iskalinov.txt	4/25/20	12...ytes	Plain Text Document
3_Farukh Iskalinov.wav	4/25/20	606 KB	Waveform audio
4_Farukh Iskalinov.txt	4/25/20	44 bytes	Plain Text Document
4_Farukh Iskalinov.wav	4/25/20	336 KB	Waveform audio

Figure 8. Files structure

The client can only register to a website by entering name, gender and age and choose any book listed by admin for further recording process. The client has no eligibility to add, delete or edit any book except listening to his own recordings for an improvement purposes. In the recording process client has the ability to rerecord the current sentence after listening his record repeatedly.

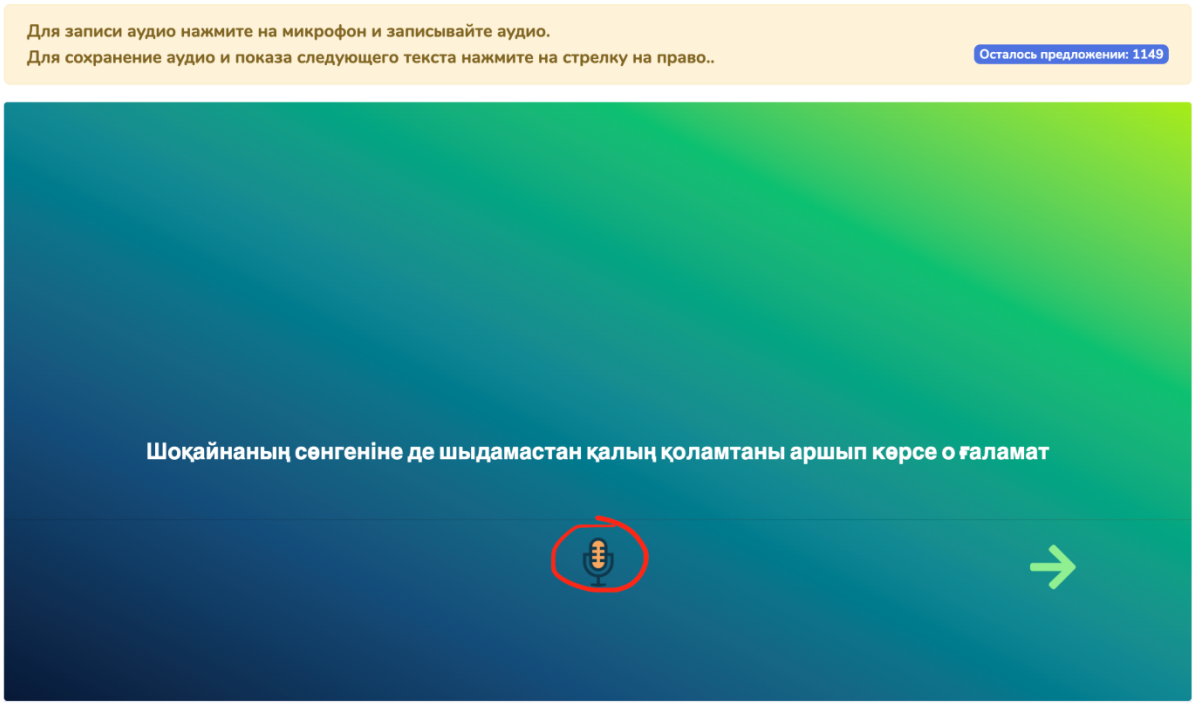


Figure 9. Recording page

The following subsections explains the features of speakers like recording the sentence (Figure 9), rerecording the sentence, manually checking for corruptions, list of recordings (Figure 10), structure of the saved files (Figure 8) etc.

The researchers can benefit from such technology, since it provides big amount of people and short time of data collection. Moreover, it provides good quality of audio speech recording, because of the ability to monitor and control the whole process of recordings. Right now, all the instructions and rules are presented in Kazakh language, but it can be adapted to any low-resourced language. Therefore, this tool will contribute huge effort on the popularity of poop famous languages by helping to collect the speeches in convenient and painless way.

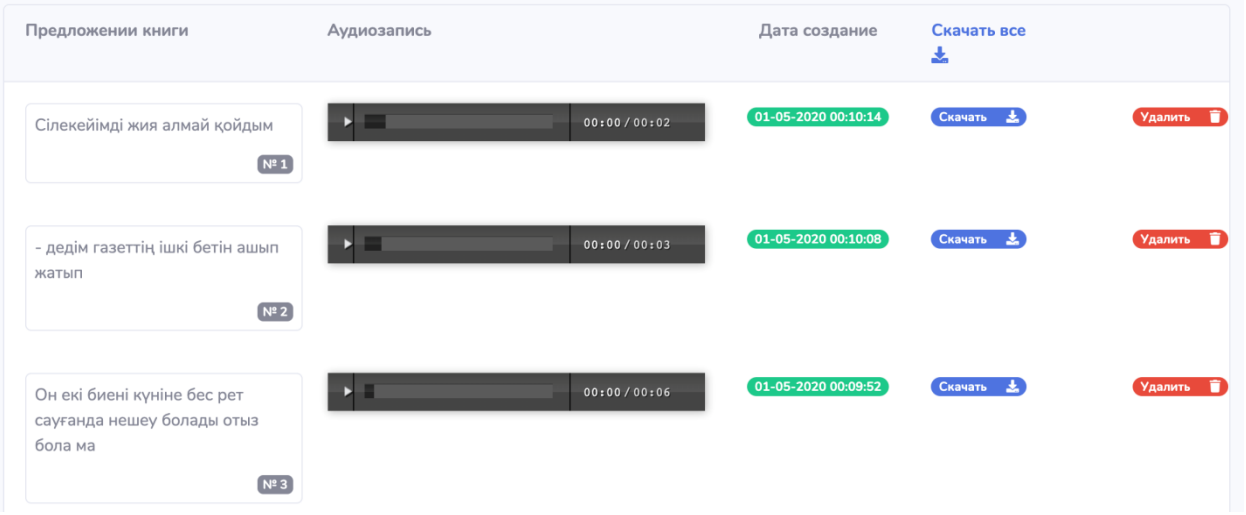


Figure 10. List recordings

Chapter five implements the Automatic speech recognition (ASR) system using the speech data that was collected in previous chapter. Overall data consist of clear 100 hours of speeches. The main idea behind constructing neural network is transfer learning approach (Figure 11).

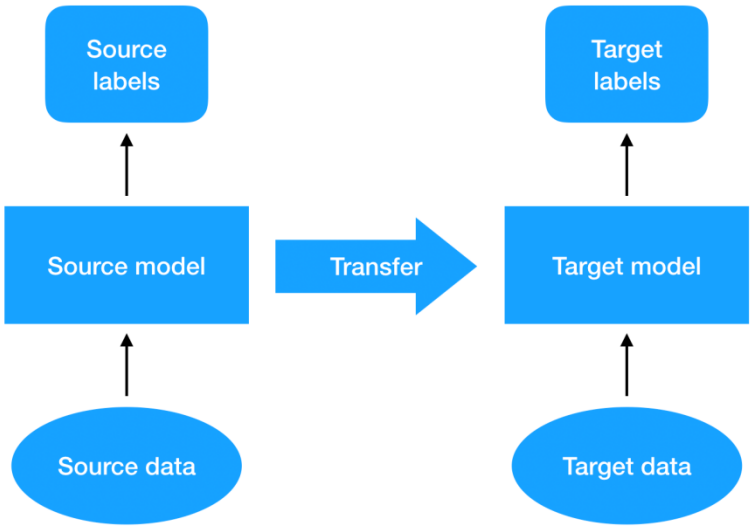


Figure 11. High-level representation of transfer learning

Using a pre-trained Russian speech recognition model all the weights were transferred to our own build neural network. The Russian speech recognition model was trained on VoxForge dataset containing 100 hours of speech data.

We have conducted to our experiment 2 different RNN types: Long-Short Term Memory (LSTM) and bidirectional LSTM. Since BiLSTM in recent researches showed promising results and understands context very well, we have decided to compare it with regular LSTM.As we have mentioned above, as an environment for this experiment we have utilized Jupiter Notebook on Python programming language. Specifically, we have used Python based Tensorflow library to create and train the model for ASR. Moreover, to construct the neural network itself Keras library extension of Tensorflow has been used. The speech dataset is cloned from GitHub repository. The repository consists of three folders: train-set, validation-set and test-set. The training was done on several Graphical Processors (GPU) Tesla K80. The server with these processors has been rented for three month of usage. The list of parameters that were used in our experiment is the following:

- two layers of LSTM and BiLSTM separately
- 128 neuron units in each layer
- 500 epochs of training
- Dropout layer after every LSTM layer with 50% probability
- Batch size is 4
- The value of momentum in Momentum Optimizer is 0.9
- Learning rate is 0.0005
- The loss function is CTC

- The metric is Label Error Rate (LER)

We have considered 4 different scenarios: 1) LSTM neural network without using Russian language model; 2) LSTM neural network using Russian language model; 3) Bidirectional LSTM without using Russian language model; 4) Bidirectional LSTM using Russian language model. They have used the same amount of layers and the same amount of neurons in each layer. The outcomes of the training process can be evaluated in Table 2.The result of each recurrent neural network actually very close, but we see that the architectures with transfer learning clearly make an improvements on everything.

LSTM layered neural network with external model has improved the training cost up to 8%, whereas Label error rate has increased up 4%. Bidirectional LSTM has showed very promising results, improving the training cost up to 24% and decreasing the label error rate down to 32% (Figure 12,13).

The experiment showed that using an external Russian ASR system model to transfer its knowledge to Kazakh language system improves the performance decently.

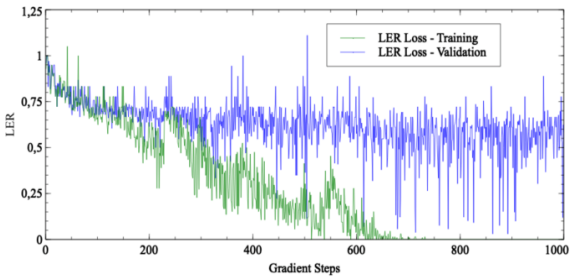


Figure 12. LER illustration

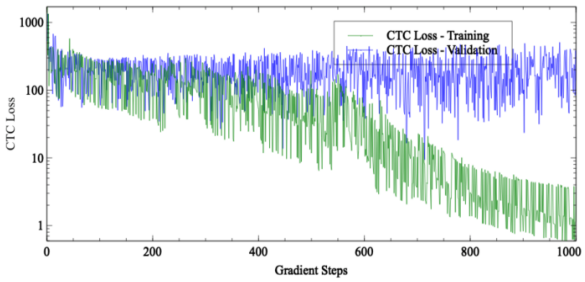


Figure 13. CTC loss illustration

Conclusion.

In the process of researching the speech recognition area all initiated objectives and tasks were obtained:

- The platform with well-designed environment and monitoring system for automatically collecting speech data was constructed. It is based on a web application containing the speech synthesis model trained using convolutional neural network.
- Using the speech data collecting instrument over 50 hours of dataset was collected. In the process of data collection, from overall population of speakers 83% were males and 17% were females.

- Transfer learning approach was applied to Kazakh ASR system using Russian speech recognition pre-built model. By building a Long short term memory based neural network and transferring all the weights from Russian speech recognition model multilingual speech recognition model was achieved.

Using the speech collecting environment 65 native speakers were involved and over 50 hours of clean speech data was obtained in less than 1.5 month.

Kazakh speech recognition system was trained using neural network based on LSTM, BiLSTM layers. The application of multilingual approach using transfer learning (Russian pre-built model) has improved the performance of Kazakh ASR model by 24% in terms of Label Error Rate.