УДК: 004.522, 004.853

М.Б.Абдіқадыр[1], Ф.У.Маликова[1]

[1]Алматинский Технологический Университет, г.Алматы, Республика Казахстан

ИЗУЧЕНИЕ СИСТЕМЫ ФОРМИРОВАНИЯ КАЗАХСКОГО РЕЧЕВОГО КОРПУСА

Аннотация. Основным вопросом, обсуждаемым в данной статье, является система корпуса речи и роль в ней системы распознавания. Одной из самых сложных задач в области автоматического распознавания речи является распознавание речи во время речи. Рассмотрен анализ речевого строя в казахском языке, работа нейронных сетей. Составление баз данных на казахском языке.

Ключевые слова: речевой корпус, распознавание, формирование корпуса, система усиления речи, нейронная модель, система ASR, казахский язык.

M.B.Abdikadyr[1], F.U.Malikova[1]

[1]Almaty Technological University, Almaty, The Republic of Kazakhstan

STUDY OF THE SYSTEM OF FORMATION OF THE KAZAKH SPEECH CORPUS

Abstract. The main issue discussed in this article is the system of the speech corpus and the role of the recognition system in it. One of the most difficult tasks in the field of automatic speech recognition is speech recognition during speech. Analysis of the speech system in the Kazakh language, the work of neural networks were considered. Compilation of databases in the Kazakh language.

Keywords: speech corpus, recognition, corpus formation, speech enhancement system, neural model, ASR system, Kazakh language.

Introduction. Speech corpus automation is one of the areas of dynamic development in the field of artificial neural systems. Over the past half century, the region has made significant progress. Currently, a lot of investment is being made in this area. The most common of these examples are call center inputs or IVR (automatic access to information without an operator) applications. In modern call centers, questions are asked by the user through the language, and the answer is returned by the computer in the appropriate language.[1] With the introduction of such automated call centers, the service sector has grown and the work of many operators has been simplified. The use of automated speech corpus systems is widely used in medical research. For example, the operator's hands are busy, but when it is necessary to enter information or to control autonomous devices for research, and even to fill out medical cards can be done by voice.

An important area of application of the automatic speech recognition system is people with disabilities (problems with vision or musculoskeletal system).

The current reward learning from human preferences could be used to resolve complex reinforcement learning (RL) tasks without access to a reward function by defining a single fixed preference between pairs of trajectory segments. However, the judgment of preferences between trajectories is not dynamic and still requires human input over thousands of iterations. We proposed a weak human preference supervision framework, for which we developed a human preference scaling model that naturally reflects the human perception of the degree of weak choices between trajectories and established a human-demonstration estimator through supervised learning to generate the predicted preferences for reducing the number of human inputs. The proposed weak human preference supervision framework can effectively solve complex RL tasks and achieve higher cumulative rewards in simulated robot locomotion—MuJoCo games—relative to the single fixed human preferences. Furthermore, our established human-demonstration estimator requires human feedback only for less than 0.01% of the agent’s interactions with the environment and significantly reduces the cost of human inputs by up to 30% compared with the existing approaches.[2]

It is important to note that speech automation has never been used in Kazakhstan. Therefore, it is necessary to note the relevance of the topic under consideration.[3] The main reason for the poor development of automation of the Kazakh language is the lack of a Kazakh language database. As can be seen in popular languages such as English, Spanish, and Chinese, a large database is required for the ASR system to work properly.

Popular speech corpora such as TIMIT or the switchboard contain a huge amount of transcribed audio recordings with different types of speeches such as telephone speeches, conversational speeches or clear microphone speeches. In the Kazakh language, there are practically no decent speech corpora in web sources.[4] The available ones are generally not free to use and certainly not sufficient to produce powerful and efficient ASR models. It takes a lot of time, a well-structured environment, and a reliable monitoring system to create a decent speech corpus for an ASR system. However, in order to completely get rid of the problem associated with scarcity of data, one should also consider the structure and approach of the neural network.

Speech data in most low resource languages does not even exist. Therefore, the creation of speech corpora is a very difficult task and requires a huge amount of time. The Kazakh language, due to its low popularity, is considered a low-income language.

Communication between people can be carried out in various forms, such as speech, visual language, gestures, sign language, etc. Among these forms, communication using speech is considered more effective and popular.[5] This leads to the fact that human interaction with a computer should be carried out using speech communication. Therefore, this fact emphasizes the importance of developing an automatic speech recognition system.

The performance of an ASR system is directly dependent on the quality of the voice data. However, speech corpora may be based on a general language or the language of a specific subject area. Using a dataset specific to a particular area has the advantage of empowering the recognition process. Moreover, using real user speech as a dataset improves the overall performance of ASR.[6]

This is a very important and decisive factor for finding an adequate set of speech data when building an ASR system for low resource languages such as Kazakh. However, the lack of the necessary volume of speech data for languages with a low level of resources is obvious. Thus, tools for collecting speech data can play a significant role in the creation of speech recognition systems.

Goals and objectives of the research

In order to build a proper ASR system avoiding the problem of data scarcity, our main goals are as follows:

- Create a well-designed environment for collecting speech data using a web platform;
- Develop a speech synthesis model to create an automatic speech collection system for small sentences;
- Collect a significant amount of speech data with transcriptions for the Kazakh language;
- For subsequent processing of the collected data and structuring files for the operation of the neural network;
- Build a neural network using recurrent neural networks based on CTC loss function;
- Build a multilingual methodology with Russian using a knowledge transfer approach.

Object of research

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

Research methods

The study will be carried out by analyzing and interpreting the existing results of modern work in the field of speech recognition, speech synthesis, natural language processing, which determine the advantages and disadvantages. The goals will be achieved through the application of machine learning algorithms, the latest advances in recurrent neural networks and sequence modeling techniques.

In this topic, we consider an overview of the processing of the speech signal itself and methods for processing the speech signal. It provides general information about types of noise such as broadband noise, interfering speech, and periodic noise. In addition, the chapter shows and illustrates the speech enhancement system (Fig. 1) and provides an analysis of various types of speech enhancement methods[7].

For systems with irregular (asymmetric and positively-negatively alternating) constraints being imposed/removed during system operation, there is no uniformly applicable control method. In this work, a control design framework is established for uncertain pure-feedback systems subject to the aforementioned constraints. Unknown nonlinearity is approximated by neural networks (Nsn) with not only natural weight updating but also activation online adjustment. The resultant control scheme is able to deal with constraints imposed or removed at some time moments during system operation without the need for altering control structure. When applied to high-speed trains, the developed control scheme ensures position tracking under speed constraints, simulation demands, and confirms the effectiveness of the proposed method.[8]

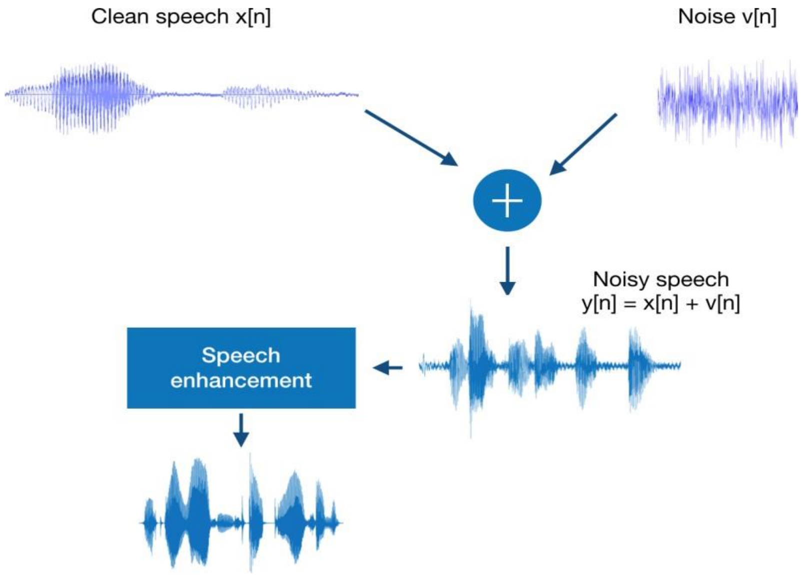

Figure 1. Speech enhancement system

There will be different types of methods of noise suppression and speech improvement, such as linear predictive coding, method of subspace signals, methods based on DFT, etc.

Also on this topic are considered acoustic models on the basis of ANN. One of the main advantages of ANN among other methods of modeling in speech recognition is the possibility of approximation of nonlinear dynamic systems. Speech is a nonlinear signal created by a nonlinear system.[9]

ANN basically represents the interrelationship of computational elements (neurons), and this nonlinear system is distributed over a network. The basal nonlinear model of the neuron is shown in Fig. 2.

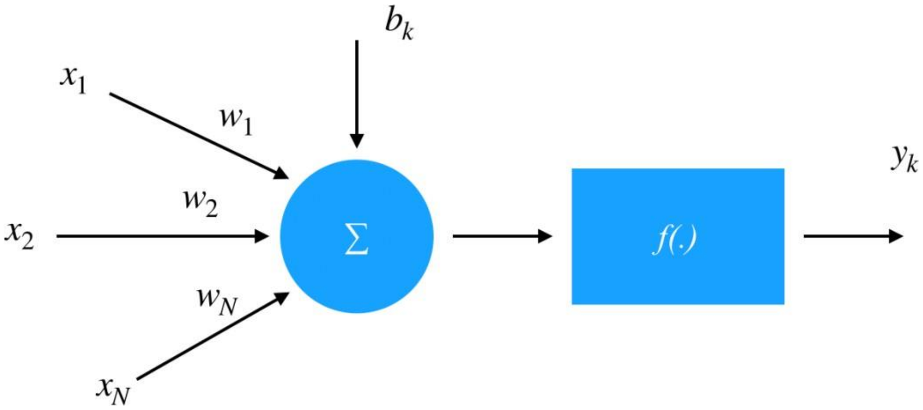

Figure 2. Nonlinear neural model

This topic also discusses new approaches to neural network training, such as CTC (Connectionist Temporal Classifier) and end-to-end models based on RNN transformers. Trainings on CTC-based speech recognition models were analyzed and explained. CTC introduces an empty character to match two sequences together, calculating the probability of a path. After that, the path aggregation algorithm is executed (Fig. 3).

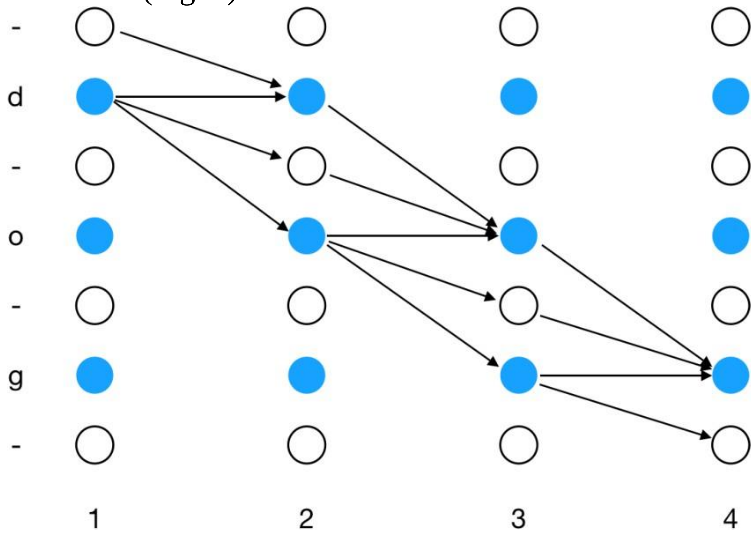

Figure 3. Path example for the word "dog"

The RNN Transformer model contains three important components: the Transcription Network (F(x)); prediction network ((P(y, g)) joint network (J(f,g)) The construction of the RNN transformer model is illustrated in Figure 4.

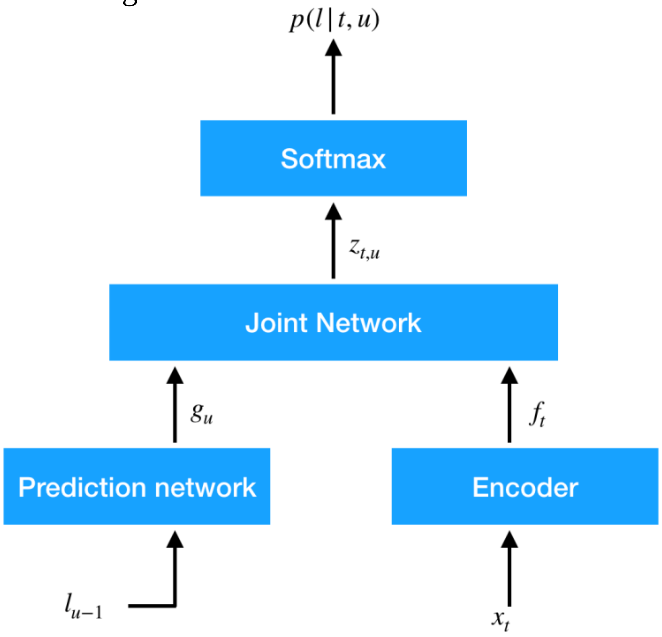

Figure 4. RNN Converter Structure

When building speech synthesis systems, one of the most important tasks is the segmentation and labeling of databases of speech signals into semantically and phonetically significant units of speech, and in our particular case, these are phonemes.[10] The resulting segments are stored in a database and used for machine learning of acoustic models in an integrated system, followed by generation of the speaker's voice in a text-to-speech synthesis system. One of the specific methods of working with trained systems is to configure the system parameters for a specific language and select an alphabet. Since the model of the Kazakh speech synthesizer uses phonemes as input symbols of the alphabet, it was necessary to solve the problem of transcribing texts according to the rules of a grapheme into a 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 was deciphered - a training sample.

Thus, a training experiment base consisting of 3500 sentences will be created, and each sentence corresponds to an audio file in wav format with a sampling rate of 22050 Hz. After that, the system module was activated, which is responsible for receiving spectrograms of audio files, on the basis of which a deep study of networks takes place (Fig. 5).

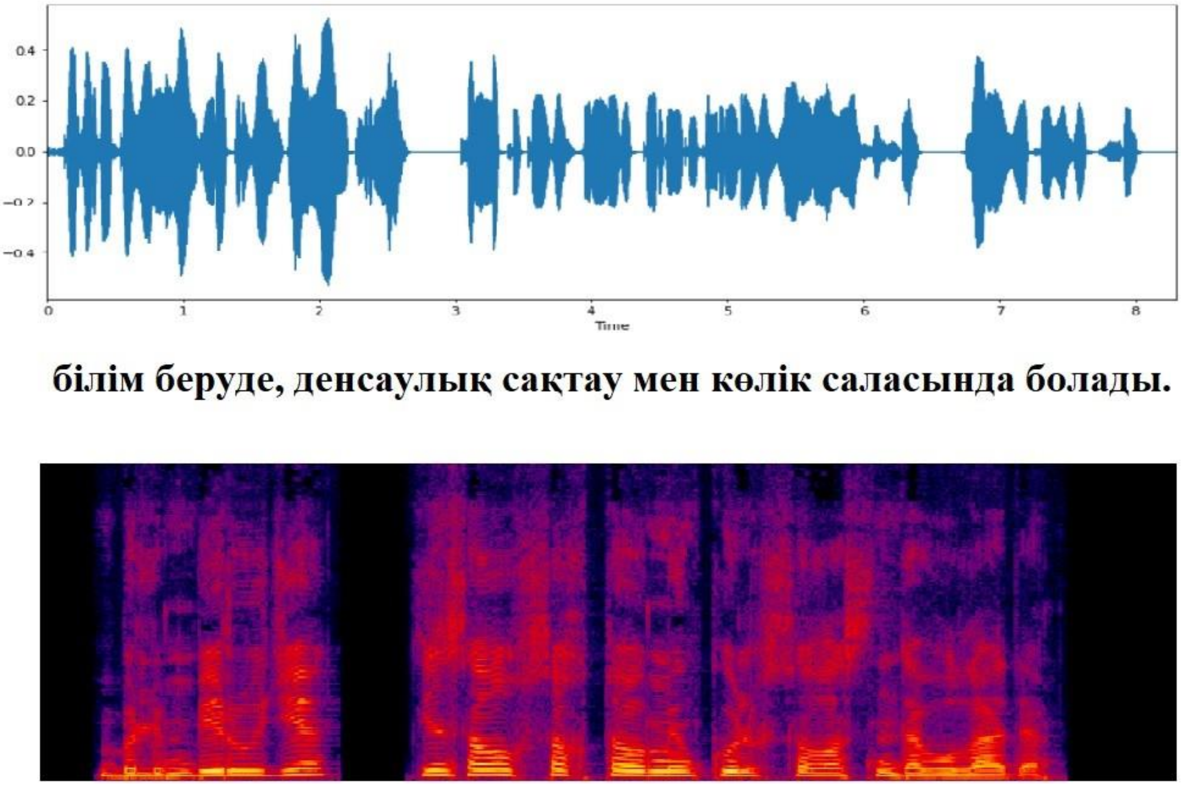

Figure 5. Sample sentence from the training database

Conclusion. In the process of researching the field of speech recognition, all the goals and objectives and their future solution were considered:

- A platform with a well-designed environment and monitoring system for automatic collection of speech data. It is based on a web application containing a speech synthesis model trained using a convolutional neural network;
- More than 50 hours of data will be collected with the voice data collection tool;
- The transmission learning approach was applied to the Kazakh ASR system using a pre-built Russian speech recognition model.[11] By building a neural network based on long-term short-term memory and transferring all weights from the Russian speech recognition model, a multilingual speech recognition model was achieved.

Using the speech collection environment, 65 native speakers will be involved and over 50 hours of pure speech data have been acquired in less than 1.5 months. The Kazakh speech recognition system was trained using a neural network based on LSTM, BiLSTM layers. Applying a multilingual approach using transfer learning (Russian off-the-shelf model) will improve the performance of the Kazakh ASR model by 24% in terms of label error rate.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ И ДИКТОРОВ. LARGE MARGIN AND KERNEL METHODS, Кешет Джозеф, Издательство:

2. В. Мэн, К. Ян, Дж. Си и Ю. Сунь, «Адаптивное нейронное управление классом неаффинных систем с ограничениями по выходу», IEEE Trans. киберн. , том. 46, нет. 1, стр. 85-95, январь 2016 г.Bgf

3. К. Чжао, Ю. Сонг, Т. Ма и Л. Хэ, «Предписанное управление производительностью неопределенных систем Эйлера-Лагранжа с учетом ограничений полного состояния», IEEE Trans. Нейронная сеть. Учиться. Сист. , том. 29, нет. 8, стр. 3478-3489, август 2018 г.

4. Ге С.С., Лю Х., К.-Х. Гох и Л. Сюй, «Управление отслеживанием группировки мультиагентов в ограниченном пространстве», *IEEE Trans. Система управления Технол.* , том. 24, нет. 3, стр. 992-1003, май 2016 г.\

5. Бондарко, Л. В. Спонтанная речь и организация системы языка // Бюллетень фонетического фонда № 8 «Фонетические свойства русской спонтанной речи». – 2015. – С. 17 – 23.

6. Годфри, Дж. Коммутатор: корпус телефонной речи для исследований и разработок [Текст] / Дж. Годфри, Э. Холлиман, Дж. МакДэниел // Proc. Международная конференция IEEE по акустике, обработке речи и сигналов (ICASSP). - 2019. -Т. 1. - С. 517-520.

7. Санкт-Петербургский институт информатики и автоматизации Российской академии наук [Электронный ресурс]. – 2016. – URL: http://www.spiiras.nw.ru/ (дата обращения: 20.11.2017).

8. ООО «ЦРТ» [Электронный ресурс]. - 2016. - URL: http://www.speechpro.ru/ (online; accessed: 20.11.2017).

9. А.А. Шарипбаев, Г.Т. Бекманова. Somequestions of the automatic transcription of kazakh language. The module of transcription of kazakh speech recognition system // Труды II Международной научно-практической конференции Информатизация общества. – Астана, 2010. – С. 543-551.

10. В. Цуй и Ю. Сонг, «Отслеживание управления неизвестными и ограниченными нелинейными системами с помощью нейронных сетей с неявным взвешиванием и обучением активации», в IEEE, Transactions on Neural Networks and Learning Systems , vol. 32, нет. 12, стр. 5427-5434, декабрь 2021 г., doi: 10.1109/TNNLS.2021.3085371.

11. З. Цао, К. Вонг и К.-Т. Лин, «Слабое наблюдение за человеческими предпочтениями для глубокого обучения с подкреплением», в IEEE Transactions on Neural Networks and Learning Systems, vol. 32, нет. 12, стр. 5369-5378, декабрь 2021 г., doi: 10.1109/TNNLS.2021.3084198.

REFERENCES

1. AUTOMATIC SPEECH AND SPEAKER RECOGNITION. LARGE MARGIN AND KERNEL METHODS, Keshet Joseph, Publisher:

2. W. Meng, K. Yang, J. Xi, and Yu. Sun, “Adaptive Neural Control of a Class of Nonaffine Systems with Output Constraints,” IEEE Trans. cybern. , volume. 46, no. 1, pp. 85-95, January 2016 Bgf

3. K. Zhao, Y. Song, T. Ma, and L. He, “Prescribed Performance Control for Uncertain Euler-Lagrange Systems with Full State Constraints,” IEEE Trans. Neural network. To study. Syst. , volume. 29, no. 8, pp. 3478-3489, August 2018

4. Ge S.S., Liu X., K.-H. Goh and L. Xu, "Constrained Space Multi-Agent Grouping Tracking Control", IEEE Trans. Control system Technol. , volume. 24, no. 3, pp. 992-1003, May 2016\

5. Bondarko, L. V. Spontaneous speech and organization of the language system // Bulletin of the Phonetic Fund No. 8 “Phonetic properties of Russian spontaneous speech”. - 2015. - S. 17 - 23.

6. Godfrey, J. Switchboard: Corpus of Telephonic Speech for Research and Development [Text] / J. Godfrey, E. Holliman, J. McDaniel // Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2019. -T. 1. - S. 517-520.

7. St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences [Electronic resource]. – 2016. – URL: http://www.spiiras.nw.ru/ (date of access: 20.11.2017).

8. LLC "CRT" [Electronic resource]. - 2016. - URL: http://www.speechpro.ru/ (online; accessed: 11/20/2017).

9. A.A. Sharipbaev, G.T. Bekmanov. Somequestions of the automatic transcription of the Kazakh language. The module of transcription of Kazakh speech recognition system // Proceedings of the II International Scientific and Practical Conference Informatization of society. - Astana, 2010. - C. 543-551.

10. W. Cui and Y. Song, “Tracking the Control of Unknown and Constrained Nonlinear Systems with Neural Networks with Implicit Weighting and Activation Learning,” in IEEE, Transactions on Neural Networks and Learning Systems, vol. 32, no. 12, pp. 5427-5434, December 2021, doi: 10.1109/TNNLS.2021.3085371.

11. Z. Cao, K. Wong, and K.-T. Lin, “Weak Observation of Human Preferences for Deep Reinforcement Learning,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 12, pp. 5369-5378, December 2021, doi: 10.1109/TNNLS.2021.3084198.

Информация об авторах

Абдіқадыр Мадина Болатқызы – магистрант, образовательный программы «Информационные системы», Алматинский Технологический Университет, г.Алматы, e-mail: madinaabdykadyr@gmail.com,

Маликова Феруза Умирзаховна – PhD доктор, ассоц.профессор, заведующий кафедры «Информационных технологий», Алматинский Технологический Университет, г.Алматы

Authors

Abdikadyr Madina Bolatkyzy - master, educational program "Information systems", Almaty Technological University, Almaty, e-mail: madinaabdykadyr@gmail.com, number phone: +7 708 644 15 08

Malikova Feruza Umirzahovna - PhD doctor, Associate Professor, Head of the Department of Information Technology, Almaty Technological University, Almaty

Для цитирования

Абдіқадыр М.Б., Маликова Ф.У. Изучение системы формирования казахского речевого корпуса // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2022. – №1(13). – С. 16-22 – DOI: 10.26731/2658-3704.2022.1(13).16-22 – Режим доступа: http://ismm-irgups.ru/toma/113-2022, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 26.04.2022)

For citations

Abdikadyr M.B., Malikova F.U. Study of the system of formation of the kazakh speech corpus // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2022. No. 1(13). P. 16-22. DOI: 10.26731/2658-3704.2022.1(13).16-22. [Accessed 26/04/22]