ANNOTATION

Dissertation work on the topic:
"Development of an automatic speech recognition system based on an end-to-end approach"
submitted for the degree of Doctor of Philosophy (PhD) in the specialty 8D06103 – "Management information systems"

ORALBEKOVA DINA ORYMBAYEVNA

Relevance of the research topic. The progress of information and computer technologies has led to improvements in processes in many machine learning technologies, particularly in speech recognition. Automatic Speech Recognition (ASR) systems have already become an integral part of our daily lives and play a huge role in the development of other machine learning technologies such as speech synthesis, machine translation, etc. ASR systems have found wide application in various fields of activity, such as voice control of a car, home and household appliances, as well as voice input in various applications, in navigation systems, etc. And these are just some examples of the use of ASR systems. There is a traditional speech recognition system, which usually consists of three main independent elements, and presents the following models, such as acoustic models for predicting the context-dependent states of subphonemes from audio, language models and a lexicon for matching phonemes to words. Models of traditional speech recognition systems are trained independently of each other, so the classical acoustic model can be trained based on Gaussian mixture models and hidden Markov models, and language models based on n-gram. For a long time, in the task of speech recognition, the model based on hidden Markov models (HMM) was widely used, and was the main technology. HMM is mainly used for dynamic time warping at the frame level and Gaussian mixture models (GMM) are used to represent signal distributions over a fixed small period of time, which usually corresponds to a pronunciation unit. For a long time, the HMM-GMM model was the general framework for speech recognition. Recently, deep learning has brought significant improvements in many studies, and in the development of speech recognition. The active use of artificial neural networks on each element of the scenario of the classical speech recognition system increases the efficiency of its work, which is reflected in many research papers. With the development of deep learning technologies, deep neural networks (DNNs) have begun to be used in speech recognition for acoustic modeling. The role of the DNN is to calculate the posterior probability of the HMM state, which can be converted to probabilities, replacing the usual GMM observation probability. Thus, the HMM-GMM model turns into HMM-DNN, which achieves better results, and becomes a popular automatic speech recognition model. Scientific papers have shown that deep neural networks have been applied to obtain an efficient acoustic model, and other research papers have built language models and vocabulary, using recurrent neural networks and networks with long and short-term memory (LSTM), respectively. Also convolutional neural networks (CNN) have been applied to

extract features from the speech signal. Thus, many published results show that the proposed approach demonstrates the best performance among all modern speech recognition systems, which is the basis for the application of various architectures of artificial neural networks on all modules of speech recognition systems.

Recently, an end-to-end method of speech recognition using machine learning methods has become widespread. In such systems, the model is implemented using only one neural network. The end-to-end implementation of the model often represents the best performance in terms of speed and accuracy of speech recognition. Foreign research works prove that the progress of the obtained results of end-to-end systems depends on an increase in the volume of training data for network training. Currently, popular applications like Voice to Text Messenger, Google Listen, Attend, Spell, Baidu Deep Speech and others work on the basis of an end-to-end approach. The basic principle of work is that modern end-to-end models are trained on the basis of big data. From the above it is possible to detect the main problem, it concerns the recognition of languages with limited training data, such as Kazakh, Kyrgyz, Turkish, etc. For such low-resource languages, large training data corpora do not exist. It should also be noted that models and methods of end-to-end architecture for low-resource languages have not been developed and studied. Currently, there are no effective algorithms and software tools for end-to-end recognition for the Kazakh language.

Foreign scientists such as Dario Amodei, Chao Weng, William Chan (USA), Xinhui Hu, Chiori Hori (Japan), Jinchuan Tian, Eric Chang, Jianlai Zhou, Jianwei Sun (China), Jan Chorowski (Poland) and other researchers have achieved high results in the improvement of end-to-end systems for recognition of popular languages, such as English and Chinese, and scientists from the near CIS, namely from the St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences Karpov A.A., Kipyatkova I.S. and their colleagues have been researching the field of speech recognition for many years and have achieved good results in the development and improvement of end-to-end Russian speech recognition systems.

It is worth noting that there are developments of domestic scientists on Kazakh speech recognition systems based on deep neural networks. In the works of scientists of the L.N. Gumilyov Eurasian National University - Sharipbai A.A., Esenbaev Zh.A., Al-Farabi Kazakh National University - Tukeev U.A., Rakhimova D.R., Nazarbayev University - Khasanov E., as well as in the scientific works of researchers from the Institute of Information and Computational Technologies - Amirgaliyev E.N., Mussabayev R.R. et al. were reflected in the development of a system for automatic recognition of Kazakh speech based on traditional, hybrid HMM-DNN models and the CTC model with Transformer. The results of these works have reached a good level of correct speech recognition, but so far these systems require improvements in reducing word recognition errors to the human level and in increasing the amount of data for training system. In addition, an end-to-end system for recognizing Kazakh speech based on the Encoder-Decoder model using the attention mechanism, which shows promising results, has not been developed. The model with the attention mechanism can work well both with and

without language models, while demonstrating a low speech recognition error. This approach is a promising direction that can be used to develop speech recognition systems with a limited training dataset.

Based on the foregoing, we can conclude that at the moment the need for effective methods, algorithms and software to improve the accuracy of Kazakh speech recognition using end-to-end models is especially *relevant*.

Purpose of the study. Research and development of a model, architecture and algorithm to improve the accuracy of continuous Kazakh speech recognition based on an end-to-end approach.

Research objectives. To achieve the goals of the study, the following tasks are solved:

1) Analysis of methods and models of speech recognition based on an end-to-end approach.

2) Development of a speech corpus for end-to-end recognition of Kazakh speech.

3) Development of an effective end-to-end model and algorithm based on the Encoder-Decoder using the attention mechanism to create a Kazakh speech recognition system.

4) Development of an end-to-end architecture and software for Kazakh speech recognition using models and methods obtained in the course of research based on an end-to-end approach.

The object of research. Modern technologies and systems of automatic speech recognition.

The subject of research. Technologies, algorithms, models and methods, software for Kazakh speech recognition based on an end-to-end approach.

Research methods. Machine learning methods, technologies and methods of automatic speech recognition, probability theory and mathematical statistics, software development methods, as well as applied and computational linguistics.

The scientific novelty

1) Speech and text corpora for the Kazakh language have been developed.

2) An end-to-end model has been developed using the attention mechanism for the recognition of Kazakh speech.

3) An efficient algorithm for Kazakh speech recognition based on an end-to-end module has been developed.

4) Software has been developed that automatically converts speech to text.

Theoretical and practical significance of the work. The theoretical significance of the research work lies in the development and implementation of effective algorithms and end-to-end models for Kazakh speech recognition, as well as in the development of speech corpus for the Kazakh language.

The practical significance of the research work lies in the application of the developed algorithms and software for further use in the development of other technologies, such as speech synthesis, machine translation, voice authentication and identification, etc. The developed system of automatic recognition of Kazakh speech can be implemented in government agencies responsible for expanding the scope of national languages on the basis of information technology; in mobile

devices (increase in the number of potential buyers due to the introduction of speech technologies in the national language); in banks (call centers with support for voice functions, voice authentication); and in the sector of production of various devices with support for voice functions.

The main findings of the defense

1) To train the model, a corpus was developed in the amount of 2000 hours of speech with transcriptions. To create the corpus, various types of speech were taken into account: prepared (reading), spontaneous.

2) The attention mechanism-based model has been modified and extended to recognize similar endings in words. The architecture of this model was built using neural networks, such as LSTM and BLSTM. The results of the experiments showed that the constructed model performs well without the use of language models for the Kazakh language and surpassed not only hybrid models based on DNN-HMM, but also other end-to-end models and showed the best results in word and character recognition accuracy.

Reliability and approbation of results. Research and results related to the topic of the dissertation have been presented and discussed at various conferences and seminars based on the following publications:

1) V International Scientific and Practical Conference "Computer Science and Applied Mathematics" (Almaty, September 29 - October 1, 2020).

2) 3rd International Conference on Computer Communication and the Internet (ICCCI) (Japan, Tokyo, June 25-27, 2021).

3) VI International Scientific and Practical Conference "Computer Science and Applied Mathematics" (Almaty, September 29 - October 1, 2021).

4) International scientific conference in the field of information technology, dedicated to the 75th anniversary of Professor U.A. Tukeyev (Almaty, October 8, 2021).

5) International Scientific Conference "Satpayev Readings 2021" (Almaty, April 12, 2021)

6) 7th International conference on Computer Science and Engineering (Turkey, Istanbul, September 14-16, 2022)

Researcher's personal contribution. The researcher personally completed and solved the tasks of the dissertation work. Developed and implemented model and algorithm for Kazakh speech recognition based on end-to-end models. Developed and expanded the speech and text corpora for the Kazakh language. Performed an experimental evaluation of the developed models and algorithms.

Relationship of the dissertation topic with the plans of research projects. The research work on the dissertation was carried out within the framework of two grant funding projects: 1) “Development of technology for multilingual automatic speech recognition using deep neural networks” (2018-2020, state registration number: 0118RK00139) 2) “Development of an end-to-end automatic speech recognition systems for agglutinative languages" (2020-2022, state registration number: 0120RK00344) at the Institute of Information and Computational Technologies.

Publication of the main results of the dissertation research.

On the topic of the dissertation work, 3 copyright certificates, 1 patent for an invention were obtained and 7 articles were published, of which 3 articles were published in journals recommended by the Committee for Control in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan, 4 articles were published in publications with a non-zero impact factor, indexed by Scopus and Web of Science:

1) Mamyrbayev, O., **Oralbekova, D.,** Keylan, A. et al. A study of transformer-based end-to-end speech recognition system for Kazakh language. Sci Rep 12, 8337 (2022). https://doi.org/10.1038/s41598-022-12260-y (**Web of Science, Q1, IF=4,3**)

2) Mamyrbayev, O.Z., Oralbekova, D.O., Alimhan, K. et al. Hybrid end-to-end model for Kazakh speech recognition. International Journal of Speech Technology (2022). https://doi.org/10.1007/s10772-022-09983-8 (**Scopus, IF=1.803, percentile 93**).

3) Mamyrbayev, O., Kydyrbekova, A., Alimhan, K., **Oralbekova, D.,** Zhumazhanov, B., Nuranbayeva, B. (2021). Development of security systems using DNN and i & x-vector classifiers. Eastern-European Journal of Enterprise Technologies, 4 (9 (112)), 32-45. doi: https://doi.org/10.15587/1729-4061.2021.239186 (**Scopus, percentile 43**);

4) Mamyrbayev, O., Alimhan, K., **Oralbekova, D.,** Bekarystankyzy A., Zhumazhanov, B. (2022). Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. Eastern-European Journal of Enterprise Technologies, 1(9(115), 84-92. https://doi.org/10.15587/1729-4061.2022.252801(**Scopus, percentile 43**);

5) Mamyrbayev O., **Oralbekova D.** Modern trends in the development of speech recognition systems // News of the National academy of sciences of the republic of Kazakhstan. – 2020. – Vol. 4, № 332. – P. 42-51 // doi.org/10.32014/2020.2518-1726.64

6) Mamyrbayev O., **Oralbekova D.**, Alimhan K., Othman M., Zhumazhanov B. Realization of online systems for automatic speech recognition// News of the National academy of sciences of the republic of Kazakhstan. – 2021. – Vol. 6, № 340. – P. 66-72 // doi.org/10.32014/2021.2518-1726.103

7) Mamyrbayev O., **Oralbekova D.**, Alimhan K., Othman M., Zhumazhanov B. Application of a hybrid end-to-end model for Kazakh speech recognition // News of the National academy of sciences of the republic of Kazakhstan. – 2022. – Vol. 1, № 341. – P. 58-68 // doi.org/10.32014/2022.2518-1726.117.

8) Copyright certificate "System for automatic recognition of Kazakh speech based on end-to-end architecture" No. 15501 dated February 25, 2021, Authors: O.Zh. Mamyrbaev, **D.O. Oralbekova**, A.S. Kydyrbekova, B.Zh. Zhumazhanov, T. Turdalykyzy.

9) Copyright certificate "Identification and authentication system through speech technologies" No. 23323 dated February 4, 2022. Authors: **Oralbekova D.O.,** Mamyrbaev O.Zh., Alimkhan K., Kydyrbekova A.S., Zhumazhanov B.Zh., Turdalykyzy T.

10) Copyright certificate "System for automatic recognition of Kazakh continuous speech based on a model with an attention mechanism" No. 24178 dated 03/05/2022. Authors: Mamyrbaev O.Zh., **Oralbekova D.O.,** Alimkhan K., Kydyrbekova A.S., Zhumazhanov B.Zh., Turdalykyzy T.

11) Patent for the invention "System and method for recognition of agglutinative continuous speech based on the end-to-end approach". No. 35886 dated 10/07/2022. Authors: Mamyrbaev O.Zh., Oralbekova D.O., Kydyrbekova A.S., Zhumazhanov B.Zh., Turdalykyzy T.

Structure and volume of dissertation work. The dissertation work consists of an introduction, 4 chapters, a conclusion, a bibliography of 111 titles and 5 appendices. The work is presented on 108 pages and contains 25 figures, 6 tables.

Brief description of the dissertation research

The **introduction** presents the relevance of the work under study, the purpose and objectives of the dissertation work, scientific novelty, theoretical and practical significance of the work and research methods.

The **first section** describes the general model of the automatic speech recognition system and provides an overview of end-to-end speech recognition systems, such as connectionist temporal classification and recurrent neural network transducer, encoder-decoder model and Transformer architecture with attention mechanism, model based on conditional random fields. An overview of related works on the considered models is given, as well as a comparative analysis of these models with traditional models is described.

The **second section** deals with the work on collecting the speech corpus for the Kazakh language, which consists of audio data with their transcriptions (textual representation of audio) and the full content of the corpus. A technique for compiling texts for transcription of telephone conversations also is given.

The **third section** describes the architecture of the encoder-decoder model with an attention mechanism. The pre-configuration of the end-to-end model for the encoder, decoder and attention mechanism is described, and evaluation metrics for speech recognition are also given. It was given the description of the data sets used in the experiments and the algorithm of the encoder-decoder with the attention mechanism. The software and hardware for the implementation of the model with attention are presented. In addition, the SmartMike Duo microphone, specially developed by Philips for speech recognition, is described, which can separate the overlap of two separate audio channels when 2 people are talking, especially when participants speak at the same time and this leads to voice overlap. An experimental verification of the proposed end-to-end model is described. To compare the results obtained, it was selected research papers related to the recognition of Kazakh speech.

The **fourth section** is devoted to the description of the system of automatic recognition of Kazakh speech. The detailed structure of the end-to-end system is considered, the work of modules for training neural networks and for validating and outputting system data is described. The interface of the program and its components are given. It describes the process of integrating an end-to-end system with a SmartMike Duo USB PSM1010 microphone, which provides a unique opportunity

for the developed system to use two separate audio channels, which provides excellent recognition accuracy and advanced speech analysis capabilities.

The **conclusion** presents the results and conclusions of the dissertation research and indicates plans for further work on the chosen direction.