recognition. The experiments were carried out using the Kazakh speech corpus with a volume of 400 hours of mixed speech, and the result showed that the system can achieve high results using RNN-based language models. Decoding based on these models does not increase computational costs, and due to this, the decoding speed does not slow down, moreover, when implementing the insertion model, the number of decoding iterations is reduced. Thus, the best CER reached 10.21%, which is a competitive result today. The proposed method is quite flexible and does not require conditional independence of variables. In addition, it can be concluded that this model can also be used to recognize other languages that are part of the group of related languages, such as Turkic.

In future work, we plan to study other insertion-based models for recognizing agglutinative languages.

Acknowledgment

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP08855743).

References

[1] 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.

[2] Graves A., Fernandez S., Gomez F., and Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In ICML, Pittsburgh, USA, 2006.

[3] W. Chan, N. Jaitly, Q. Le and O. Vinyals (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

[4] Jaitly, Navdeep & Le, Quoc & Vinyals, Oriol & Sutskeyver, Ilya & Bengio, Samy. (2015) An Online Sequence-to-Sequence Model Using Partial Conditioning, 2015.

[5] Chung-Cheng Chiu and Colin Raffel, "Monotonic chunkwise attention," in Proceedings of ICLR, 2018.

[6] Deng, Keqi & Cao, Songjun & Zhang, Yike & Ma, Long & Cheng, Gaofeng & Xu, Ji & Zhang, Pengyuan. (2022). Improving CTC-based speech recognition via knowledge transferring from pre-trained language models. https://doi.org/10.48550/arXiv.2203.03582

[7] J. Heymann, K. C. Sim and B. Li, "Improving CTC Using Stimulated Learning for Sequence Modeling," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5701-5705, doi: 10.1109/ICASSP.2019.8682700.

[8] Zeyer, Albert, Kazuki Irie, Ralf Schlüter and Hermann Ney. "Improved training of end-to-end attention models for speech recognition." ArXiv abs/1805.03294 (2018): n. pag.

[9] L. Lu, X. Zhang and S. Renais, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5060-5064, doi: 10.1109/ICASSP.2016.7472641.

[10] Huang Z, Wang P, Wang J, Miao H, Xu J, Zhang P. Improving Transformer Based End-to-End Code-Switching Speech Recognition Using Language Identification. Applied Sciences. 2021; 11(19):9106. https://doi.org/10.3390/app11199106

[11] Miao, Haoran & Cheng, Gaofeng & Gao, Changfeng & Zhang, Pengyuan & Yan, Yonghong. (2020). Transformer-Based Online CTC/Attention End-To-End Speech Recognition Architecture. 6084-6088. 10.1109/ICASSP40776.2020.9053165.

[12] J. Gu, J. Bradbury, C. Xiong, V. O. Li, and R. Socher, "Non-autoregressive neural machine translation," arXiv preprint arXiv:1711.02281, 2017.

[13] N. Chen, S. Watanabe, J. Villalba, and N. Dehak, "Listen and fill in the missing letters: Non-autoregressive transformer for speech recognition." arXiv preprint arXiv:1911.04908, 2020.

[14] Yuya Fujita, Shinji Watanabe, Motoi Omachi, Xuankai Chan. Insertion-Based Modeling for End-to-End Automatic Speech Recognition. INTERSPEECH 2020. https://doi.org/10.48550/arXiv.2005.13211.

[15] Rakhimova, D., Sagat, K., Zhakypbaeva, K., Zhunussova, A. (2021). Development and Study of a Post-editing Model for Russian-Kazakh and English-Kazakh Translation Based on Machine Learning. In: Wojtkiewicz, K., Treur, J., Pimenidis, E., Maleszka, M. (eds) Advances in Computational Collective Intelligence. ICCCI 2021. Communications in Computer and Information Science, vol 1463. Springer, Cham. https://doi.org/10.1007/978-3-030-88113-9_42

[16] Rakhimova, D., Turarbek, A., Kopbosyn, L. (2021). Hybrid Approach for the Semantic Analysis of Texts in the Kazakh Language. In: Hong, TP., Wojtkiewicz, K., Chawuthai, R., Sitek, P. (eds) Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2021. Communications in Computer and Information Science, vol 1371. Springer, Singapore. https://doi.org/10.1007/978-981-16-1685-3_12

[17] 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.

[18] 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

[19] O. Mamyrbayev, D. Oralbekova, A. Kydyrbekova, T. Turdalykyzy and A. Bekarystankyzy, (2021) "End-to-End Model Based on RNN-T for Kazakh Speech Recognition," 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), 2021, pp. 163-167, doi: 10.1109/ICCCI51764.2021.9486811.

[20] Chan, W., Kitaev, N., Guu, K., Stern, M., & Uszkoreit, J. (2019). KERMIT: Generative Insertion-Based Modeling for Sequences. ArXiv, abs/1906.01604.

[21] Jiatao Gu, Qi Liu, and Kyunghyun Cho. 2019. Insertion-based Decoding with Automatically Inferred Generation Order. Transactions of the Association for Computational Linguistics, 7:661–676.

[22] Stern, Mitchell et al. "Insertion Transformer: Flexible Sequence Generation via Insertion Operations." ICML (2019).

[23] Zhang, Chuan-Fei & Liu, Yan & Zhang, Tian-Hao & Chen, Song-Lu & Chen, Feng & xu, Yin. Non-autoregressive Transformer with Unified Bidirectional Decoder for Automatic Speech Recognition. Computation and Language, 2021. https://doi.org/10.48550/arXiv.2109.06684.

[24] Fan, R., Chu, W., Chang, P., Xiao, J., & Alwan, A. (2021). An Improved Single Step Non-autoregressive Transformer for Automatic Speech Recognition. Interspeech.

[25] Uday Kamath, John Liu, Jimmy Whitaker - Deep Learning for NLP and Speech Recognition (2019, Springer).