Investigation of insertion-based speech recognition method

[1,a]ORKEN MAMYRBAYEV, [2,c]DINA ORALBEKOVA, [3,e]MOHAMED OTHMAN, [1]TOLGANAY TURDALYKYZY, [1,d]BAGASHAR ZHUMAZHANOV, [1,f]KURALAI MUKHSINA

[1]Institute of information and computational technologies Almaty, KAZAKHSTAN
[3]Department of "Cybersecurity, information processing and storage" Satbayev University Almaty, KAZAKHSTAN
[5]Department of Communication Technology and Network, Universiti Putra Malaysia, Kuala Lumpur, MALAYSIA
[a]0000-0001-8318-3794, [c]0000-0003-4975-6493, [d]0000-0002-5035-9076, [e]0000-0002-5124-5759, [f]0000-0002-8627-1949

Abstract: End-to-end models have come to the field of speech recognition, replacing traditional and hybrid ones. The basic principle of operation of modern end-to-end models is the generation of the output sequence from left to right, applying an autoregressive function during decoding. Until this time, it has not been proven that this decoding method is the best in text-to-speech technology. In addition, end-to-end models only consider previous information to predict the next output. This approach does not address the issue of speech conversion when the previous information was slurred. Thus, we began to apply the insertion method, which uses non-autoregressive generation of output data in random order. In this work, the model was trained on the basis of the insertion method and connectionist temporal classification for Kazakh speech recognition. The conducted experiments showed that this model improves the quality of Kazakh speech recognition.

Keywords: automatic speech recognition, end-to-end, insertion-based, connectionist temporal classification, Transformer.

Received: June 22, 2021. Revised: May 19, 2022. Accepted: June 13, 2022. Published: July 29, 2022.

1. Introduction

Text to speech technology has already become an integral part in robotic and automated systems and is widely used in call centers to reduce the load on operators, as well as to quickly switch the client to the right specialist and in other banking services as an intelligent voice assistant, etc. Emergence of end-to-end (E2E) models [1], namely models based on Connectionist Temporal Classification (CTC) [2] and encoder-decoder with attention [3] and online models for streaming speech recognition [4, 5], greatly simplified the scheme for constructing speech recognition systems. The E2E model replaces the classical system's multi-stage pipeline with a single deep network, which reduces training time and decoding time, and it allows for joint optimization with post-processing such as natural language understanding. A lot of studies has been done to improve the performance of E2E models, namely CTC [6,7], attention [8,9]. The combined use of E2E models has been proven to outperform not only traditional automatic speech recognition (ASR) models, but also E2E models that have been implemented separately. In addition, the appearance of the Transformer and its use with CTC [10,11] brought the ASR system to a new level.

E2E models work on a sequence-to-sequence (seq2seq) basis, autoregressively generating utterances from left to right, because speech is sequential. Speech-to-text or sentence decoding takes place on the basis of previous utterances, which does not take into account future utterances. And it has not been proven that this decoding method is the best in text-to-speech technology. Of course, we can predict the next word based on previous utterances, but this method will not work if some of the previous words were slurred or too quiet. In this case, future information can be used to determine what was said before. You need to understand that when decoding, you need to take into account not only the previous context, but also the future one, which plays an important role in converting audio to text.

Machine translation technologies actively use the non-autoregressive transformer model [12]. There have been attempts to use this model for speech recognition [13]. The essence of the method is to predict masked tokens that were randomly masked during decoding. In this process, the model is trained to predict masked labels using the attention mechanism. And uses a non-autoregressive method when decoding sequences by evaluating hidden labels. This approach has achieved high efficiency in comparison with models with an autoregressive orientation. To predict from sequence masks, it is necessary to have an idea of the length of the label output sequence, which is a difficult process. To solve this problem, an insertion model is used that generates an output sequence in an arbitrary order without adding auxiliary elements to determine the length of output labels [14]. This process reduces the number of steps and decoding time.

Our main work is related to the development, implementation and promotion of the Kazakh language in information systems technologies. For machine translation from other languages into Kazakh and morphological analysis of the Kazakh language, some works have been devoted [15, 16]. In addition, there were attempts to create an identification and verification system based on Kazakh speech [17]. In our previous works, we implemented CTC-based models with an attention model for recognizing agglutinative languages like Kazakh and Azerbaijani, as part of transfer learning [18], and we also implemented an RNN-T model for streaming Kazakh speech recognition in real time [19]. However, these models require improvements to accurately convert Kazakh speech to text. In this paper, we consider the implementation of an insertion-based model for end-to-end recognition of Kazakh speech and compare the results with our previous works. In addition, a review of other work on the implementation of

insertion-based models for end-to-end speech recognition is provided.

In this paper, we propose insertion-based recognition of the Kazakh language in order to improve recognition accuracy and reduce errors in word and character recognition.

The research work is presented in the following sequence: section 2 provides an overview of the work on the topic under study, section 3 describes the insertion model. Further, in section 4, the main settings of the model, information about the training data, the obtained experimental data, and analysis of the results are given. The final section contains conclusions and plans for future works.

2. Related Works

The insertion model was first mentioned in [12] (2017), where it was applied in machine translation. A model was presented that, instead of generating an output word based on a previously generated output, creates its output in parallel, which reduces the delay during output by an order of magnitude. By using the input tokens as a latent variable and fine-tuning the gradient, the model performed better than the Transformer autoregressive network. However, it should be noted that the model cannot be trained when only positional embeddings are used as input to the decoder.

The researchers of [14] (2020) propose for the first time to apply insertion-based models to ASR problems, which were originally proposed for machine translation. Insertion-based models solve mask prediction problems and can generate an arbitrary output sequence generation order. In addition, a new formulation of joint training of models based on insertion and CTC was demonstrated. This formulation strengthens CTC by making it dependent on token generation based on insertion without autoregression. The experiments were carried out on three public tests and achieved innovative performance compared to the conventional Transformer.

Chan, W. *et al.*, 2019) [20] (2019) presented the KERMIT model, which works on the principle of insertion in sequence modeling. KERMIT models the joint distribution and its decomposition with a single neural network without relying on a predefined factorization of the data distribution. The model supports both sequential fully autoregressive decoding and parallel partial autoregressive decoding. This approach matches or exceeds the performance of specialized modern systems for a wide range of tasks without the need for problem-oriented architectural adaptation of the model.

Jiatao Gu *et al.* [21] (2019) proposed a new InDIGO decoding algorithm that supports flexible generation of sequences in random order using insert operations. By extending the Transformer, the authors effectively implemented the proposed approach by letting it be trained on a predetermined adaptive order obtained from the beam search. Experiments were conducted to solve problems such as word order recovery, machine translation, image caption, and code generation. As a result, the proposed algorithm showed competitive performance compared to conventional left-to-right generation.

Mitchell Stern *et al.* [22] (2019) presented Insertion Transformer, an iterative, partially autoregressive model for sequence generation based on insertion operations. Unlike typical autoregressive models that rely on a fixed, often left-to-right, output ordering, the implemented approach allows for arbitrary ordering, allowing tokens to be inserted anywhere in the sequence during decoding. Such flexibility has a number of advantages, such as teaching the model to follow not only certain orders, like generating from left to right or traversing a binary tree, but it can also be trained to maximize entropy over all possible insertions for reliability. In addition, the proposed model can support both autoregressive generation and partially autoregressive generation. Insertion Transformer has been found to outperform many prior non-autoregressive transformation approaches at comparable or higher levels of concurrency and successfully restore the performance of the original transformer by requiring only a logarithmic number of iterations during decoding.

Chuan-Fei Zhang *et al.* [23] (2021) presented a new non-autoregressive transformer with a unified bidirectional decoder (NAT-UBD) that can use left-to-right and right-to-left contexts simultaneously. However, direct use of bidirectional contexts will cause information leakage, which means that the output of the decoder can be affected by character information from the input at the same position. To avoid information leakage, a new attention mask has been proposed and normal queries have been modified, like keys and value matrices for NAT-UBD. Experimental results confirm that NAT-UBD can achieve a Symbol Error Rate (CER) of 5.0%/5.5% on Aishell1 development/test sets, outperforming all previous NAR converter models.

Ruchao Fan *et al.* [24] (2021) proposed several methods for improving the accuracy of an end-to-end CTC alignment-based single step non-autoregressive transformer (CASS NAT) followed by performance analysis. First of all, blocks of internal attention, supplemented by convolution, are applied to the encoder and decoder modules. Secondly, the authors proposed to expand the trigger mask (acoustic boundary) for each marker in order to increase the reliability of the CTC alignment. In addition, repeated loss functions are used to improve the gradient updating of the parameters of the bottom layer. Without the use of an external language model, WER using the three methods is 3.1% and 7.2% on pure and other Librispeech test sets, and CER is 5.4% on the Aishell1 test set.

3. Prepare Your Paper Before Styling

The most probable word sequence W∗ can be estimated by maximizing $P(W|X)$ for all possible word sequences $V*$ (1). This process can be represented by the following expression [25]:

$$W^* = \underset{W \in V^*}{argmax}\, P(W|X) \quad (1)$$

Acoustic characteristics are a sequence of feature vectors X, and a sequence of words is defined as W. Therefore, the main work of the end-to-end APP is to create a model that can accurately calculate the posterior distribution $P(W|X)$ depending on the end-to-end model for all possible word sequences V^*.

3.1 Insertion-based Model for ASR

To implement the insertion model, word positions and insertion order must be determined, and thus an additional Order parameter is added to take into account the possible permutations of the ordering, which depends on the number of

words spoken. Thus, Equation (1) is transformed as follows Equation (2):

$$P(W|X) = \sum_{Order} p(W, Order|X) = \sum_{Order} p(W_{Order}|X) p(Order|X) \quad (2)$$

Insertion-based decoding uses a Transformer network with the concept of position order (Fig. 1).

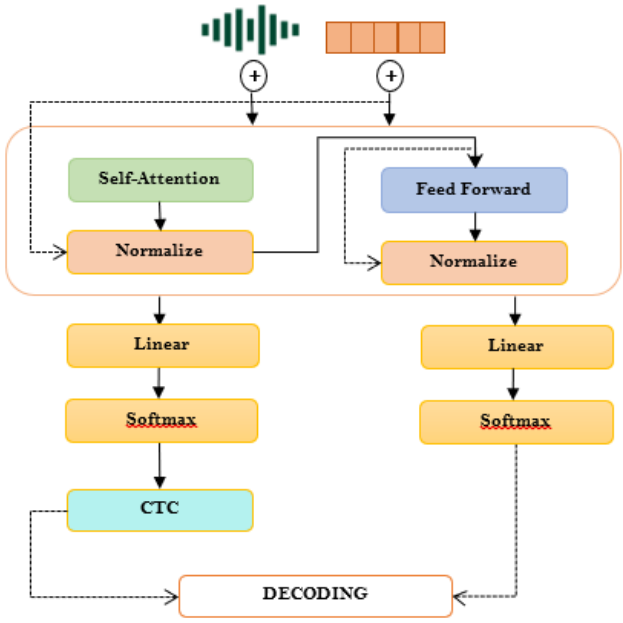

Fig. 1. Insertion-based model

This model is trained in a completely end-to-end way, retaining all the benefits of decoding. This approach allows you to ignore the need to determine the length of the sequences in advance.

As shown in Equation (2), word position prediction is performed by insertion into having words. Since the word position update does not change with respect to precomputed representations, the word position information can be reused in the next decoding step.

During sequence decoding, we model the conditional probability of a sequence of words by matching the distribution of words and their positions in the correct order.

4. Experimental Setups

4.1 Speech Corpus

To train the model based on the insertion, a speech corpus was chosen, which contains more than 400 hours of speech, collected in the laboratory of "Computer Engineering of Intelligent Systems" of the Institute of Information and Computational Technologies of the Ministry of Education and Science of the Republic of Kazakhstan (https://iict.kz/laboratory-of-computer-engineering-of-intelligent-systems/). This corpus consists of records of native speakers of the Kazakh language of different sexes and ages; telephone conversations with transcriptions; some recordings were taken from news sites and art audiobooks.

In the corpus, sound files are divided into training and test parts, these are approximately 90% and 10%, respectively.

With the help of this corpus, one can experiment and determine the effects of different datasets for evaluating the performance of the model, and checking the customizable characteristics of the model, and, secondly, studying the effect of the database size on the recognition and decoding speed for the Kazakh language.

All audio materials were in .wav format. The PCM method was used to convert the data into digital form. Discrete frequency 44.1 kHz, bit depth 16 bits.

4.2. Configuring Basic Settings

The same number of blocks for the encoder and decoder was chosen and their number was 8.

To improve the decoding result and ensure fast convergence, CTC was added to the insertion model. The CTC consists of a directional six-layer BLSTM with 256 cells in each layer. A dropout rate of 0.1 was set. To optimize the model, we apply the Adam algorithm. The decoding weight for CTC is 0.35. The beam research width at the decoding stage is 10. The experiment ended in almost 110 epochs. The number of epochs of 110 was enough to fully train the network. The language model was applied to align and search for a beam to decode the insert model.

4.3 Conducting an Experiment and Obtaining Results

To date, there are a limited number of studies on Kazakh speech recognition based on end-to-end models. In addition, you can find some research papers that were implemented based on traditional and hybrid models, but used a smaller amount of training data.

In our work [18] for end-to-end speech recognition in the Kazakh and Azerbaijani languages, an approach based on transfer learning was proposed. Joint CTC-attention models were implemented. Two language corpora were trained simultaneously. The volume of the corpus was 400 hours of Kazakh speech.

Another researsh [19] implemented a popular RNN-T based model for Kazakh speech recognition. A language model was applied, which positively influenced the output indicators. The model was trained on a 300-hour corpus.

The results of experiments on speech recognition using the above works and our model are shown in Table 1.

TABLE I. CER FOR DIFFERENT SPEECH RECOGNITION SYSTEMS

Model	*CER (%)*	*Data volume, h*
End-to-end with transferring (Kazakh + Azerbaijan language) without LM	14.23	400
LSTM and BLSTM based RNN-T with LM	10.60	300
Insertion-based model	10,21	400

The results obtained using the Insertion-based model improved their performance due to decoding with the STS model and the language model by 4%, but their absence could reduce the efficiency of the insertion model. In addition, when implementing this model, the number of steps was reduced, due to this, the learning and decoding speeds decreased compared to other models.

5. Conclusion

The work is devoted to the study of the influence of joint end-to-end models of CTC and insertion for Kazakh speech

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).