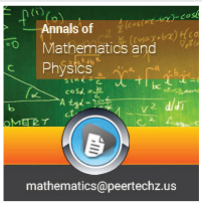
Peertechz



MATHEMATICS AND PHYSICS GROUP

ANNALS OF Mathematics and Physics OPEN ACCESS JOURNAL

ISSN: 2689-7636 DOI: https://dx.doi.org/10.17352/amp

Research Article

Development online models for automatic speech recognition systems with a low data level

Mamyrbayev OZh[1]*, Oralbekova DO[1,2]*, Alimhan K[1,3], Othman M[4] and Zhumazhanov B[1]

[1]Institute of Information and Computational Technologies, Almaty, Kazakhstan

[2]Satbayev University, Almaty, Kazakhstan

[3]LN Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

[4]Universiti Putra Malaysia, Kuala Lumpur, Malaysia

Received: 01 August, 2022
Accepted: 22 August, 2022
Published: 23 August, 2022

***Corresponding authors:** Mamyrbayev OZh, Institute of Information and Computational Technologies, Almaty, Kazakhstan, Tel: +7 777 366 2727; E-mail: morkenj@mail.ru

ORCID: https://orcid.org/0000-0001-8318-3794

Oralbekova DO, Institute of Information and Computational Technologies, Almaty, Kazakhstan, Satbayev University, Almaty, Kazakhstan, Tel: +7 771 131 0188; E-mail: dinaoral@mail.ru

ORCID: https://orcid.org/0000-0003-4975-6493

Keywords: Automatic speech recognition; End-to-end; RNN-T; Neural transducer; Monotonic chunkwise attention

https://www.peertechzpublications.com

Check for updates

Abstract

Speech recognition is a rapidly growing field in machine learning. Conventional automatic speech recognition systems were built based on independent components, that is an acoustic model, a language model and a vocabulary, which were tuned and trained separately. The acoustic model is used to predict the context-dependent states of phonemes, and the language model and lexicon determine the most possible sequences of spoken phrases. The development of deep learning technologies has contributed to the improvement of other scientific areas, which includes speech recognition. Today, the most popular speech recognition systems are systems based on an end-to-end (E2E) structure, which trains the components of a traditional model simultaneously without isolating individual elements, representing the system as a single neural network. The E2E structure represents the system as one whole element, in contrast to the traditional one, which has several independent elements. The E2E system provides a direct mapping of acoustic signals in a sequence of labels without intermediate states, without the need for post-processing at the output, which makes it easy to implement. Today, the popular models are those that directly output the sequence of words based on the input sound in real-time, which are online end-to-end models. This article provides a detailed overview of popular online-based models for E2E systems such as RNN-T, Neural Transducer (NT) and Monotonic Chunkwise Attention (MoChA). It should be emphasized that online models for Kazakh speech recognition have not been developed at the moment. For low-resource languages, like the Kazakh language, the above models have not been studied. Thus, systems based on these models have been trained to recognize Kazakh speech. The results obtained showed that all three models work well for recognizing Kazakh speech without the use of external additions.

Introduction

Speech technologies provide more natural user interaction with computing and telecommunication systems compared to a standard graphical interface. Initially, automatic speech recognition systems were developed for people with physical disabilities that face the difficulty of typing by hand. However, at present, these systems are especially used in the everyday life of an ordinary person who wants to make his life easier and eliminate the routine work of typing dictation or writing a letter just by voicing commands.

Today the most popular speech recognition systems are based on end-to-end (E2E) structure models [1]. However, the conditions for the implementation of such systems are not so easy to satisfy for low-resource languages, since data in the amount of thousands or more hours are required to obtain a high-quality speech recognition system. However, the data

Citation: Mamyrbayev OZh, Oralbekova DO, Alimhan K, Othman M, Zhumazhanov B (2022) Development online models for automatic speech recognition systems with a low data level. Ann Math Phys 5(2): 107-111. DOI: https://dx.doi.org/10.17352/amp.000049

question can be solved using the method of data augmentation and other techniques [2-4]. Because the Kazakh language has very little data in electronic form, a tremendous amount of work needs to be done to collect and develop speech and text data. An increase in data during the implementation of end-to-end systems favorably affects the quality of recognition not only of traditional models but also of online end-to-end models. Thus, this topic is very relevant for other languages with limited training data.

The main advantage of E2E models is not only high performance in terms of speed and accuracy of speech recognition but also their use locally on portable devices. Lately, there has been strong interest in training the E2E process for ASR that directly output a sequence of words given the input sound. There are many works related to online recognition not only of speech but also of the handwritten alphabet, which plays an important role in the development of computer vision and natural language processing using machine learning methods, namely through the end-to-end method [5,6]. Streaming Speech Recognition allows you to stream an audio stream into speech-to-text and obtain real-time stream speech recognition results as the audio is processed. To implement such a system, online attention-based models for E2E systems are used, the most popular are RNN-T, Neural Transducer (NT) and Monotonic Chunkwise Attention (MoChA).

In this article, we have carried out an overview of online models of E2E systems. And they built a system for recognizing Kazakh speech.

Materials and methods

Model Recurrent neural transducer

The Recurrent neural transducer (RNN-T) was first mentioned in [7,8] as a modification of the Connected Time Classification (CTC) model [9] for sequence marking problems where the alignment between the input sequence x and the output targets l is unknown. This is achieved in the CTC formulation by introducing an additional label, which is an empty label, which generates the probability of outputting a label corresponding to a given input frame. However, the main limitation of CTC is its assumption that the model output in a given frame is independent of previous output labels, i.e. are independent: $l_t \perp\!\!\!\perp l_j \mid x$ for $t<j$.

The structure of RNN-T consists of the following elements – an encoder [10], a joint and a prediction network; as described in [11], the RNN-T model has a similar structure as in other E2E method architectures, like an encoder with an attention mechanism, if the decoder can be represented as a connection between the network prediction element and an internetwork. The RNN is an encoder that converts the input acoustic data into a high-level intermediate representation and performs the same function as AM in a standard speech recognition system. Consequently, the RNN output is driven by the chain of previous acoustic data, as in the CTC model. The task of RNN-T is to eradicate the conditional independence assumption in the CTC by adding an RNN prediction network component that is explicitly driven by the predicted history of previous non-empty model targets [12]. For example, the prediction network takes the last non-empty label as input data to cause the output data. Eventually, the joint network, which has a feed-forward network, combines the outputs of the encoder and prediction networks to form logits. Thus, all this is conjugated by the softmax layer to obtain the distribution over the next output character/word (Figure 1).

Consideration should be given to the fact that compared to other streaming encoder-decoder architectures like the Neural Transducer, the prediction network is independent of the received encoder information. This advantage makes it possible to train the decoder as a language model on data.

Neural transducer

The Neural Transducer (NT) can generate some of the output signals as blocks of input data arrive, thus satisfying the online condition.

A speech transformer usually consists of an encoder that converts acoustic inputs to high-level representations and a decoder that produces linguistic output, which is characters or words from encoded representations. The problem is that the input and output sections are of variable (also different) lengths, and usually, no alignments are available between them. Sonneural transducers must study both the classification from acoustic performance to linguistic predictions and the agreement between them. The sensor models differ in the composition of the classifier and leveler [13].

More formally, given the input sequence $x = (x_1, .., x_T)$ of length I, and the output sequence $y = (y_1, ..., y_U)$ of length I and each yu is an I-dimensional one-time vector, the transformers simulate the conditional distribution $p(y \mid x)$. The encoder maps the input x to a high-level representation $h = (h_1, ..., h_T')$, which may be shorter than the input ($T' \leq T$) downsampled in time (Figure 2). The encoder can be built using feed-forward neural networks (DNN), recurrent neural networks (RNN), or convolutional neural networks (CNN). The decoder determines the alignment and displays from h to y [14].

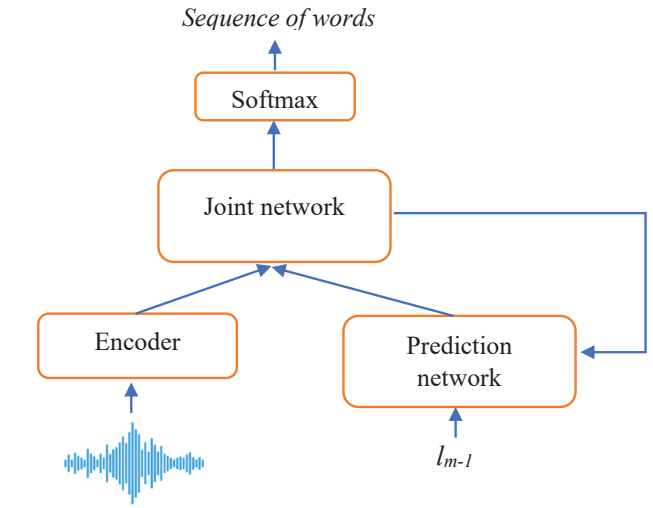


Figure 1: Structure of RNN-T.

Citation: Mamyrbayev OZh, Oralbekova DO, Alimhan K, Othman M, Zhumazhanov B (2022) Development online models for automatic speech recognition systems with a low data level. Ann Math Phys 5(2): 107-111. DOI: https://dx.doi.org/10.17352/amp.000049

In [15], in the problem of TIMIT phoneme recognition, a neural converter with a three-level unidirectional LSTM encoder and a three-level unidirectional LSTM converter achieved an accuracy of 20.8% phoneme error rate (PER), which is close to the ultra-modern for unidirectional methods. Moreover, it was also set up that with good matches, the model can reach a PER of 19.8%.

Monotonic chunkwise attention

To develop a MoChA model, it is first necessary to study well the structure of the model from sequence to sequence (seq2seq) and the most common form of soft attention used with it (described in [16, 17]).

The MoChA model computes context using two types of attention: hard monotonous attention and soft piecemeal attention (Figure 3).

MoChA allows the model to perform soft attention on small chunks of memory before where the hard monotonous attention mechanism chose to be present. It also has a learning routine that allows it to be applied directly to existing sequence-to-sequence (seq2seq) models and trained using standard backpropagation of the error. In [18], it was shown that MoChA effectively narrows the gap between monotonous and soft attention in speech recognition on the Internet and provides a relative improvement of 20% compared to monotonous attention in summarizing documents. These advantages entail only a small increment in the number of parameters and computational costs.

In [19], a MoCha-based approach was used for streaming speech recognition, resulting in a WER of 9.95%. It has been experimentally shown [20] that MoChA provides the highest performance in solving speech recognition problems on the Internet and is significantly superior to the rigid monotonic model based on attention.

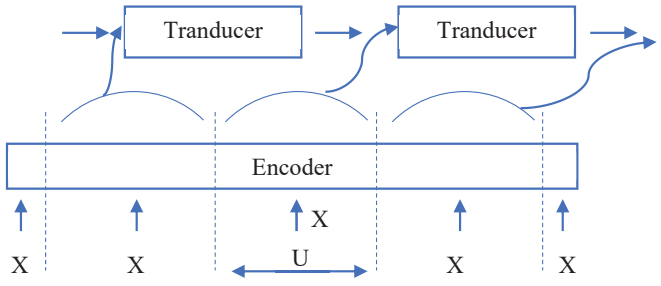


Figure 2: Architecture Neural Transducer.

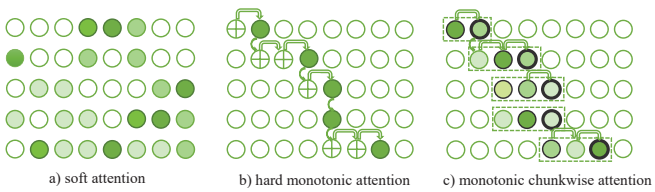


Figure 3: Model Monotonic Chunkwise Attention.

Results

Corpus for training. To train the RNN-T model, a speech corpus was chosen, which contains more than 300 hours of speech, collected in the laboratory "Computer Engineering of Intelligent Systems" of the Institute of Computer Engineering of the Ministry of Education and Science of the Republic of Kazakhstan [21]. This corpus consists of records of Kazakh speakers of different sexes and ages; telephone conversations with transcriptions; some of the recordings were taken from news sites and audiobooks of art.

Corpus data augmentation

To increase the size of the data, the Voice conversion (VC) method proposed in [2] was applied. The speed and tempo of the audio data have been changed without changing the content. Thus, the size of the body was increased to approximately 380 hours of speech.

Experiments

To extract features from audio data, the method of chalk-frequency cepstral coefficients was applied [22].

To the model based on RNN-T recurrent neural networks, like LSTM and BLSTM with six layers [23], for NT and MoChA BLSTM also have six layers.

The initial learning rate coefficient was set to 10e-5. Dropout was used for each output of the recurrent layer as regularization and is equal to 0.5. For our model, we used a gradient descent optimization algorithm based on Adam [24].

To measure the quality of the speech recognition system, the WER metric was used - the number of incorrectly recognized words, which is determined by Levenshtein distance [25] (Table 1).

The results showed that the MoChA model for the Kazakh language works well compared to other models.

Table 1: Results of work of online E2E models on the main and extended corpus.

Model	WER, %	Data volume, h
RNN-T	15.8	300
	14.3	380
NT	15.6	300
	14.2	380
MoChA	14.9	300
	13.7	380

Discussion

During the experiment, corpuses with two data volumes were used. The increase in data has led to improved speech recognition accuracy. The MoChA model meticulously outperformed the RNN-T and NT models by 4% and 3%, respectively. On the other hand, it took a very long time to tune the MoChA model, as it consists of many parameters. Additional components such as language model and phoneme dictionary were not applied.

Citation: Mamyrbayev OZh, Oralbekova DO, Alimhan K, Othman M, Zhumazhanov B (2022) Development online models for automatic speech recognition systems with a low data level. Ann Math Phys 5(2): 107-111. DOI: https://dx.doi.org/10.17352/amp.000049

The results obtained are approximately the same for these models, which can be said that all three approaches do a good job with speech recognition with a limited set of data.

These approaches were built based on bidirectional networks and do not have a consistent trend in recognition, and are adapted for real-time speech recognition, so there is no need to wait for audio speech recording. Given this advantage, these models can be applied and implemented in various devices and systems.

Conclusion

In this paper, the architecture of online models for automatic recognition of Kazakh continuous speech, which uses self-attention components, was considered. The considered model is easier to implement and the learning process can be reduced by parallelizing the processes. The MoChA-based model showed better results in Kazakh speech recognition in terms of WER indicators than the RNN-T and NT models. This proves that the implemented model can be leveraged in other low-resource languages. In addition, the results obtained were approximate and it allows the conclusion that all three approaches do a good job with speech recognition with a limited data set.

In this case, in the future, such models will make it possible to recognize high-quality continuous speech in real time using less computing power, while having higher performance than other APP models.

In further research, it is planned to conduct experiments with other types of E2E methods for speech recognition with limited training data.

Acknowledgement

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

Information about the authors

Mamyrbayev Orken Zhumazhanovich – Ph.D, Associate Professor, Deputy General Director in Science, Institute of Information and Computational Technologies, Almaty, Kazakhstan, morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794

Oralbekova Dina Orymbayevna – Ph.D student, specialty "Management information systems", researcher, Satbayev University, Almaty, Kazakhstan; dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493

Alimhan Keylan – Ph.D, Professor of Department of Mechanics and Mathematics, L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan; keylan@live.jp; https://orcid.org/0000-0003-0766-2229

Othman Mohamed – Ph.D, Professor of Department of Communication Technology and Network, Universiti Putra Malaysia, Kuala Lumpur, Malaysia, mothman@upm.edu.my, https://orcid.org/0000-0002-5124-5759

Zhumazhanov Bagashar – Candidate of tech. sciences, researcher, Institute of Information and Computational Technologies, Almaty, Kazakhstan; bagasharj@mail.ru.

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.4:332; 2020; 42 – 51 // doi.org/10.32014/2020.2518-1726.64
2. Matthew Baas, Herman Kamper. Voice Conversion Can Improve ASR in Very Low-Resource Settings. arXiv:2111.02674. 2021. [eess.AS]. (data of request: 11.11.2021).
3. Chang S, Deng Y, Zhang Y, Wang R, Qiu J, Wang W, Zhao Q, Liu D. An Advanced Echo Separation Scheme for Space-Time Waveform-Encoding SAR Based on Digital Beamforming and Blind Source Separation. Remote Sensing. 2022; 14(15):3585. https://doi.org/10.3390/rs14153585
4. Chang S, Deng Y, Zhang Y, Zhao Q, Wang R, Zhang K. An Advanced Scheme for Range Ambiguity Suppression of Spaceborne SAR Based on Blind Source Separation. IEEE Transactions on Geoscience and Remote Sensing. 2022.
5. Singh TP, Gupta S, Garg M, Gupta D, Alharbi A, Alyami H, Anand D,Ortega-Mansilla A, Goyal N. Visualization of Customized Convolutional Neural Network for Natural Language Recognition. Sensors 2022; 22:2881. https://doi.org/10.3390/s22082881
6. Popli R, Kansal I, Garg A, Goyal N, Garg K. Classification and recognition of online hand-written alphabets using Machine Learning Methods. IOP Conference Series: Materials Science and Engineering. 2021; 1757-8981. http://dx.doi.org/10.1088/1757-899X/1022/1/012111
7. Graves A. Sequence transduction with recurrent neural networks. arXiv: 2012; 1211.3711. (data of request: 02.09.2021).
8. Li J, Zhao R, Hu H, Gong Y, "Improving RNN Transducer Modeling for End-to-End Speech Recognition." 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore. 2019; 114-121.
9. Graves A, Fernandez S, Gomez F, Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In ICML, Pittsburgh, USA. 2006.
10. Prabhavalkar, Rohit, Rao, Kanishka, Sainath, Tara, Li, Bo, Johnson, Leif, Jaitly, Navdeep. A Comparison of Sequence-to-Sequence Models for Speech Recognition. 2017; 939-943. 10.21437/Interspeech. 2017-233.
11. Jaitly N, Le QV, Vinyals O, Sutskever I, Sussillo D, Bengio S, "An online sequence-to-sequence model using partial conditioning," in NIPS, 2016.
12. Chan W, Jaitly N, Le QV, Vinyals O, "Listen, attend and spell," CoRR, 2015; 1508.01211.
13. Sainath, Tara, et al. Improving the Performance of Online Neural Transducer Models. 2018; 5864-5868.
14. Jaitly, Navdeep, Le, Quoc, Vinyals, Oriol, Sutskeyver, Ilya, Bengio, Samy. An Online Sequence-to-Sequence Model Using Partial Conditioning, 2015.
15. Battenberg E, et al. "Exploring neural transducers for end-to-end speech recognition," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, 2017; 206-213.
16. Prabhavalkar, Rohit. et. al. A Comparison of Sequence-to-Sequence Models for Speech Recognition. 2017; 939-943. 10.21437/Interspeech. 2017-233
17. Chung-Cheng Chiu, Colin Raffel. "Monotonic chunkwise attention." in Proceedings of ICLR. 2018.

Citation: Mamyrbayev OZh, Oralbekova DO, Alimhan K, Othman M, Zhumazhanov B (2022) Development online models for automatic speech recognition systems with a low data level. Ann Math Phys 5(2): 107-111. DOI: https://dx.doi.org/10.17352/amp.000049

18. Chiu, Chung-Cheng, et.al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. 2018; 4774-4778. 10.1109/ICASSP.2018.8462105.

19. Kim, Chanwoo, et.al. End-to-End Training of a Large Vocabulary End-to-End Speech Recognition System. 10.1109/ASRU46091. 2020; 2019:9003976; 2019.

20. Hou J, Guo W, Song Y, et al. Segment boundary detection directed attention for online end-to-end speech recognition. J AUDIO SPEECH MUSIC PROC. 2020.

21. Orken M, Dina O, 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.

22. Mamyrbayev O, Alimhan K, Oralbekova D, Bekarystankyzy A, Zhumazhanov B. Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. Eastern-Eur. J. Enterpris. Technol. 2022; 19(115), 84–92.

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

24. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv. 2014. http://arxiv.org/abs/1412. 6980 (data of request: 01.11.2021).

25. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals // Soviet physics. Doklady. 1996; 10:707–710.

Citation: Mamyrbayev OZh, Oralbekova DO, Alimhan K, Othman M, Zhumazhanov B (2022) Development online models for automatic speech recognition systems with a low data level. Ann Math Phys 5(2): 107-111. DOI: https://dx.doi.org/10.17352/amp.000049