НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2022 Том 22 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2022 Vol. 22 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2022-22-6-1143-1149

Improving out of vocabulary words recognition accuracy for an end-to-end Russian speech recognition system

Andrei Yu. Andrusenko[1]✉, Aleksei N. Romanenko[2]

[1,2] STC-Innovations Ltd., Saint Petersburg, 194044, Russian Federation
[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] andrusenkoau@itmo.ru✉, https://orcid.org/0000-0002-8697-832X
[2] AlexeySk8@gmai.com, https://orcid.org/0000-0001-6267-018X

Abstract

Automatic Speech Recognition (ASR) systems are experiencing an active introduction into our daily lives, simplifying the way we interact with electronic devices. The advent of end-to-end approaches has only accelerated this process. However, the constant evolution and a high degree of inflection of the Russian language lead to the problem of recognizing new words outside the vocabulary (Out Of Vocabulary, OOV) because they did not take part in the training process of the ASR system. In such a case, the ASR model tends to predict the most similar word from the training data which leads to a recognition error. This is especially true for ASR models that use decoding based on a Weighted Finite State Transducer (WFST), since they are obviously limited by the list of vocabulary words that may appear as a result of recognition. In this paper, this problem is investigated on the basis of an open data set of the Russian language (common voice) and an integrated ASR system using the WFST decoder. A method for retraining an integral ASR system based on the discriminative loss function MMI (maximum mutual information) and a method for decoding the integral model using a TG graph are proposed. Discriminative learning allows smoothing the probability distribution of acoustic class prediction, thus adding more variability in the recognition results. Decoding using the TG graph, in turn, is not limited to recognizing only vocabulary words and allows the use of a language model trained on a large amount of external text data. An eight-hour subset from the common voice base is used as a test set. The total number of OOV words in this test sample is 18.1 %. The results show that the use of the proposed methods allows to reduce the word recognition error (Word Error Rate, WER) by 3 % in absolute value relative to the standard method of decoding integral models (beam search), while maintaining the ability to recognize OOV words at a comparable level. The use of the proposed methods should improve the overall quality of recognition of ASR systems and make such systems more resistant to the recognition of new words that were not involved in the learning process.

Keywords

automatic speech recognition, end-to-end ASR, discriminative training, OOV words, weighted finite state transducer

For citation: Andrusenko A.Yu., Romanenko A.N. Improving out of vocabulary words recognition accuracy for an end-to-end Russian speech recognition system. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 6, pp. 1143–1149. doi: 10.17586/2226-1494-2022-22-6-1143-1149

УДК 004.93

Повышение точности распознавания внесловарных слов для интегральной системы автоматического распознавания русской речи

Андрей Юрьевич Андрусенко[1]✉, Алексей Николаевич Романенко[2]

[1,2] ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация
[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] andrusenkoau@itmo.ru✉, https://orcid.org/0000-0002-8697-832X
[2] AlexeySk8@gmai.com, https://orcid.org/0000-0001-6267-018X

Аннотация

Предмет исследования. Системы автоматического распознавания речи (Automatic Speech Recognition, ASR) активно внедряются в нашу повседневную жизнь, тем самым упрощая способ взаимодействия с электронными

устройствами. Развитие интегральных (end-to-end) подходов только ускоряет данный процесс. Тем не менее постоянная эволюция и большая степень флективности русского языка приводят к проблеме распознавания новых вне словарных (Out Of Vocabulary, OOV) слов, которые не принимали участие в процессе обучения ASR-системы при ее создании. В связи с этим ASR-модель может прогнозировать наиболее похожее слово из обучающих данных, что влечет к ошибке распознавания. Особенно это касается ASR-моделей, использующих декодирование на основе взвешенного конечного автомата (Weighted Finite State Transducer, WFST), так как они заведомо ограничены списком словарных слов, которые могут появиться в результате распознавания. Выполнено исследование проблемы на основе открытой базы русского языка (common voice) и интегральной ASR-системы, использующей WFST-декодер. **Метод.** Предложен метод дообучения интегральной ASR-системы на основе дискриминативной функции потерь MMI (Maximum Mutual Information) и метода декодирования интегральной модели с помощью TG графа. Дискриминативное обучение позволило сгладить распределение вероятностей предсказания акустических классов, добавив таким образом большую вариативность в результате распознавания. Так как декодирование с помощью TG графа не имеет ограничений на распознавание только словарных слов, оно позволило использовать языковую модель, обученную на большом количестве внешних текстовых данных. **Основные результаты.** В качестве тестового множества использована восьмичасовая подвыборка из базы common voice. Общее число OOV-слов в тестовой выборке составило 18,1 %. Полученные результаты показали, что использование предложенных методов сократило пословную ошибку распознавания на 3 % в абсолютном значении относительно стандартного метода декодирования интегральных моделей. При этом сохранилась возможность распознавания OOV-слов на сравнимом уровне. **Практическая значимость.** Использование предложенных методов может улучшить общее качество распознавания ASR-систем и сделать их более устойчивыми к распознаванию новых слов, которые не участвовали в процессе обучения модели.

Ключевые слова

автоматическое распознавание речи, интегральные системы, дискриминативное обучение, OOV-слова, взвешенный конечный автомат

Ссылка для цитирования: Андрусенко А.Ю., Романенко А.Н. Повышение точности распознавания внесловарных слов для интегральной системы автоматического распознавания русской речи // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 6. С. 1143–1149 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-6-1143-1149

Introduction

The active development of artificial deep neural networks has led to a significant breakthrough in Automatic Speech Recognition (ASR) tasks. For a long time, the conventional hybrid approach to building ASR systems [1] was dominant in this area. However, it is a complex and lengthy process of obtaining separate modules that are eventually combined into a single system. This encouraged the development of end-to-end learning methods for ASR models [2] the task of which is to combine all modules into one deep neural network. Unlike the conventional hybrid approach, the end-to-end method learns to generate text from an input audio signal directly without using an intermediate signal representation. This approach dramatically simplifies building and training an ASR system, showing high recognition accuracy comparable with conventional systems. In many ASR tasks, it is already the best solution [3, 4] which only increases the scientific community's attention to the development of this technology.

However, the constant evolution (emergence of new words) and a high degree of inflection (a large number of different spellings of the same word) of the Russian language leads to the degradation of the recognition accuracy of ASR systems since such "new" words may not be present in the training data. These words are called Out Of Vocabulary (OOV) words. The words in the training data are called IV (in vocabulary) words. Conventional methods, by definition, are not capable of recognizing new words since their decoder is pre-limited by the list of words (lexicon) presented in training data. To tackle this problem, such OOV words can be added to the lexicon graph [5], but it is necessary to know the list of OOV words in advance. To recognize words previously unknown to the model, there is a method of adding the unk subgraph [6]; it can recognize new words in cases where the ASR model could not find the corresponding IV word. This is achieved by adding a phoneme subgraph with arbitrary transitions within itself which allows generating new sequences of phonemes that are then converted into words. Such unk subgraph is integrated into the lexicon graph which has only known paths based on IV words. However, the accuracy of this method is far from perfect. A comparison of OOV word recognition techniques for classical ASR systems is considered in [7].

End-to-end ASR models in the recognition process produce separate target acoustic units; it can be letters (graphemes) or pieces of words (subwords) consisting of several letters. Such a method is theoretically capable of recognizing OOV words, but in practice, this can lead to degradation of the overall recognition accuracy. The model may be wrong in IV word recognition because it does not have a lexicon constraint during the decoding process, as it does in the conventional hybrid model. With a limited number of training data (less than 100–150 hours of transcribed data), this problem only gets worse. To improve this situation, BPE (Byte pair encoding) augmentation [8] of acoustic classes can be used generating new word divisions into acoustic units during each training epoch. A comparison of conventional hybrid and end-to-end ASR systems for the Russian language can be found in [9].

There is a method for combining an encoder trained in end-to-end mode with a WFST (Weighted Finite State Transducer) decoder used for conventional systems [10]. This approach allows using external text data to train the language model and make fewer mistakes in recognizing IV words due to the presence of a lexicon. This combines the

strengths of end-to-end systems (robust encoders trained in end-to-end mode) and a strong language model integrated into the WFST graph from the classical system. However, the lexicon still limits this approach making it impossible to recognize OOV words.

This paper proposes a WFST graph method to decode an end-to-end model unrestricted by a predetermined lexicon L. Instead of TLG, the TG graph is used in which the language model is based on graphemes. We also use the discriminative MMI (Maximum Mutual Information) loss function for finetuning the pre-trained end-to-end model similarly from [11] which further increases the accuracy of OOV word recognition by smoothing the probabilities of acoustic unit prediction. The final model reduces the overall WER (Word Error Rate) by 3 % in absolute value relative to the baseline end-to-end decoding algorithm while retaining the ability to recognize OOV words on the same level.

Decoding method

The conventional hybrid ASR system is divided into two main components: acoustic and language models. The classifier of target acoustic units (phonemes, graphemes, subwords), based on a deep neural network, usually acts as an acoustic model. Its task is to predict the probability distribution of acoustic units on each short segment (25 ms long) of the input signal. The language model is integrated into the decoding module and generates the most probable word sequences based on the predictions received from the acoustic model. The lexicon sets the rules for dividing words into acoustic units mapping each word from the training data to the corresponding sequence of acoustic units. For example, if graphemes are used as acoustic units, then the lexicon will display the rules for splitting each word into letters. The final decoding graph for the conventional hybrid ASR system is a WFST that sequentially transforms the outputs of the acoustic model into the final recognition result in the text. Such a graph is a composition of four separate graphs:

— H — topology graph for acoustic model outputs; converts acoustic model outputs (tying states) to context-sensitive acoustic classes (such as triphons).
— C — transition graph from context-dependent acoustic classes to context-independent ones (for example, from triphons to phonemes).
— L — lexicon.
— G — word language model.

As a result, we get the HCLG graph which is used in the process of decoding the conventional hybrid system.

For end-to-end ASR systems, it is possible to use such a WFST graph for decoding [10] since its encoder also plays the role of an acoustic model. To do this, it is necessary to replace the HC subgraph of the conventional model with T, the graph of the CTC (Connectionist Temporal Classification) topology. As a result, a TLG graph is obtained which can decode the results of recognition of the end-to-end model. This method uses both approaches strengths: an encoder based on the latest architectures, trained in end-to-end mode, and a WFST graph with a powerful language model. However, due to the L graph presence, this method cannot recognize OOV words.

To solve this problem, we propose to remove the L graph from the composition of the decoding graph. As a result, we get a TG graph, where G is already a grapheme language model instead of a word one. Using positional graphemes as acoustic units makes it possible to directly restore the sequence of words from the results of such "non-word" recognition. This approach retains all the advantages of the hybrid method of decoding the end-to-end model using the WFST graph and makes it possible to recognize new OOV words by getting rid of the lexicon restriction.

Discriminative finetuning

To improve the recognition accuracy of end-to-end systems, one can use discriminative training methods based on MMI objective function [11]. In contrast to the MMI finetuning for the conventional hybrid system, in the end-to-end approach, the loss function is calculated immediately over the entire input signal, leading to extensive memory consumption. The compact CTC topology can be used to combat this, it described in detail in [12]. Additional finetuning of the pre-trained end-to-end model for a certain number of epochs with the MMI loss function can have a beneficial effect on the accuracy of the final system. Our experiments show that this approach improves the overall recognition accuracy and leads to a noticeable improvement in OOV word recognition. This effect may be because the model trained only on the CTC loss function initially has excessively high probabilities for predictions, which was already noted earlier in [13]. Additional finetuning with MMI allows smoothing out this "peaky" behavior and increases the probability of recognition of alternative acoustic units improving the recognition accuracy of OOV words. An example of comparing the probability distribution by acoustic units for a model trained only on CTC loss and then finetuned with MMI loss is shown in Fig. 1.

Data set

An open database of the Russian language common voice [14] is used to train and test ASR models for the proposed method. This database consists of voice recordings from users' personal devices (smartphones, PCs, etc.). The sampling rate of audio files is adjusted to 16000 Hz. The entire data available for download from the original source was divided into a train (104 hours and 38 minutes) and a test set (8 hours and 16 minutes). The number of OOV words for the selected test set was 18.1 %. It is important to note that this is a percentage of the total words spoken on the test, not a percentage of unique OOV words.

As an acoustic feature, 80-dimensional log-Mel filterbank coefficients are used, to be calculated with a window of 25 ms and a step of 10 ms. In order to reduce the effect of model overfitting, SpeecAugment [15] is used. The number of frequency masks is 2 with a mask range of 30 bins. The number of time masks is 2 with a length of 40 ms.

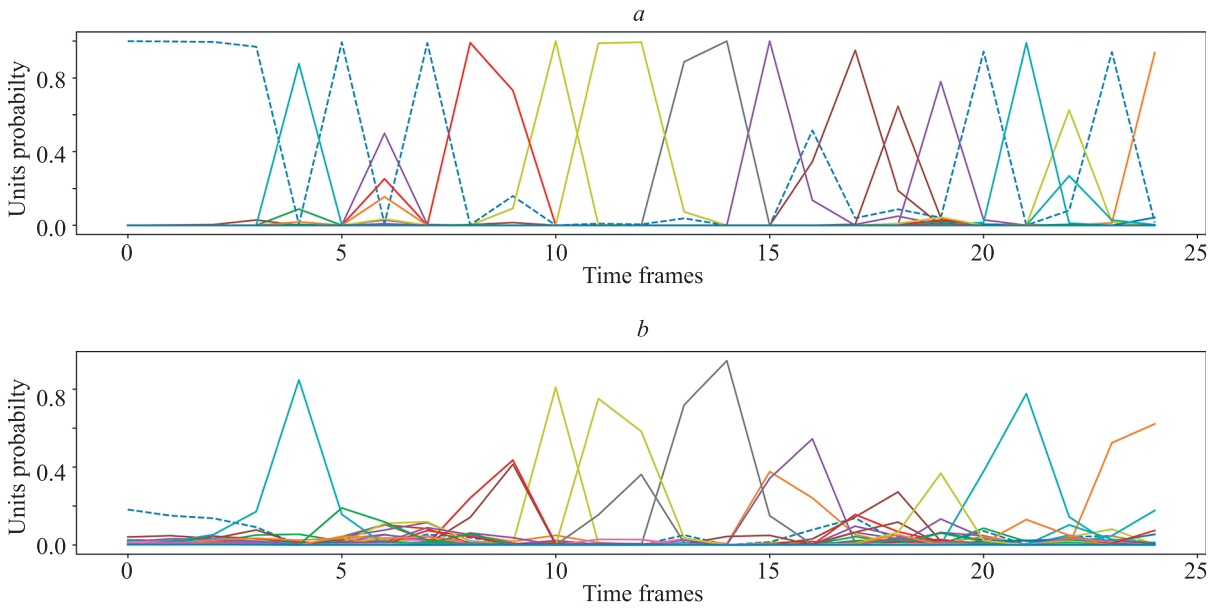

Fig. 1. Comparison of the acoustic units distribution probabilities of the end-to-end model for two training options: only with the CTC loss function (*a*) and the use of additional training with 10 MMI epoches (*b*). The dotted line indicates the probability of a blank symbol. Continuous lines correspond to the remaining acoustic units

ASR system

The acoustic model (encoder) of our end-to-end ASR system is Conformer [16] — a deep neural network architecture based on the attention mechanism and local convolution blocks. This architecture works effectively with a long-term and local context within utterances (for example, with whole words and individual graphemes). In this work, we use a 12-layer Conformer with the size of the attention mechanism embedding of 320. The number of attention heads is 8. The size of fully connected layers is 1024, and the size of the conformer convolution block is 5.

The Joint CTC-Attention mode (the weight of the CTC loss function is 0.3) is used for training the end-to-end model. An attention decoder part based on the Transformer architecture [17] consists of 6 layers with an attention mechanism size of 320. The number of attention heads is 8, and the size of fully connected layers is 1024. This block is also used for decoding in the beam search mode. The weight of the encoder block is 0.3. Adam algorithm is used as an optimizer with 16000 warm-up steps to a peak learning rate of 0.002, followed by a quadratic decrease. The total number of training epochs is 100.

For decoding in conventional hybrid mode, a WFST graph is used based on the Kaldi toolkit [18]. For language modeling, a 3-gram word language model is used in the case of the TLG.fst graph and a 7-gram grapheme language model for the TG.fst graph.

The word error rate (WER) metric is used to evaluate the overall accuracy of speech recognition. OOV word recognition accuracy is measured using Character Error Rate (OOV-CER) and Word Error Rate (OOV-WER) for specific OOV words. As an implementation of such a metric, the texterrors[1] are used. It is worth noting that WER is the highest priority indicator in comparing the accuracy of model recognition. OOV-CER and OOV-WER are secondary metrics that show only OOV words recognition accuracy.

All experiments were carried out in the ESPnet speech recognition toolkit [19].

Experiment results

As baseline values for comparison with subsequent modifications, the speech recognition accuracy of the end-to-end model was investigated using a standard beam search decoding algorithm. The influence of the choice of acoustic units for ASR system modeling was also investigated. The results of these experiments are presented in Table 1. Subwords were obtained using sentencepice tokenizer with 250 BPE units. Additionally, the effect of BPE augmentation with bpe_alpha 0.1 (dp0.1) on the recognition accuracy of OOV words was tested. It can be concluded that the graphemes units work the best for this task. The BPE model shows a slightly better WER, but noticeably loses in OOV word recognition. BPE dropout helps improve OOV recognition but messes up WER a lot.

The following experiments use the WFST graph to decode the predictions of the end-to-end model encoder. Table 2 shows that the transition from TLG.fst to TG.fst can significantly reduce the OOV-CER/WER and the

Table 1. Decoding the end-to-end model using the beam search algorithm for different acoustic units, %

Acoustic units	WER	OOV-CER	OOV-WER
Graphemes	30.5	28.8	58.1
BPE 250	30.4	32.8	62.2
BPE 250 + dp0.1	37.5	32.6	60.0

[1] Available at: https://github.com/RuABraun/texterrors (accessed: 15.11.2022).

Table 2. Results for the WFST decoding method of the end-to-end model using TLG.fst and TG.fst, %

Training criterion	Graph	WER	OOV-CER	OOV-WER
CTC-Att	TLG.fst	34.3	51.9	100.0
CTC-Att	TG.fst	29.7	33.6	62.4
CTC-Att+MMI	TLG.fst	33.5	50.9	100.0
CTC-Att+MMI	TG.fst	27.5	31.0	60.4

overall WER. At the same time, additional finetuning for ten epochs using the MMI loss function provides a further improvement in both WER and OOV metrics.

Table 3 compares the baseline beam search with our best result from Table 2. The method of adding the unk subgraph was also tested. It can be seen that baseline beam search is the best for recognition of OOV words but shows that WER is 3 % worse than TG.fst + MMI which still works well with OOV words and is inferior in OOV metrics quite a bit. Model with the unk subgraph is also capable of recognizing OOV words but lags far behind in WER and OOV metrics.

Additional analysis of recognition results indicates that most of the recognized OOV words in about 96.3 % differ from IV words by no more than three letters. It means that the proposed method works well with the inflection of the language (small changes in the ending, word prefix, etc.). Recognition of a new OOV word that is not similar to an existing IV word is rare. As a measure of the proximity of two words, the Levenshtein distance is used. It is defined as the minimum number of single-character operations (insertions, deletions, and replacements) required to convert one character sequence to another. A complete distribution of the number of correctly recognized OOV words depending on the Levenshtein distance to the nearest IV word from the training set can be seen in Fig. 2.

Table 4 shows an example of recognized OOVs and IV words closest to them.

Table 4. An example of recognized OOV words and IV words closest to them. Grapheme differences are highlighted in bold

OOV	IV	Levenshtein distance
обширны**м**	обширны**й**	1
немногим**и**	немногим	1
платформ**ой**	платформ**ы**	2
подпи**шут**	подпис**ь**	3
вознагра**жде**но	вознагра**дить**	4

Table 3. Comparison of the best results for different recognition methods of the same end-to-end ASR model, %

Method	WER	OOV-CER	OOV-WER
Baseline beam search	30.5	28.8	58.1
TG.fst + MMI	27.5	31.0	60.4
TLG.fst + unk subgraph	32.1	37.5	74.1

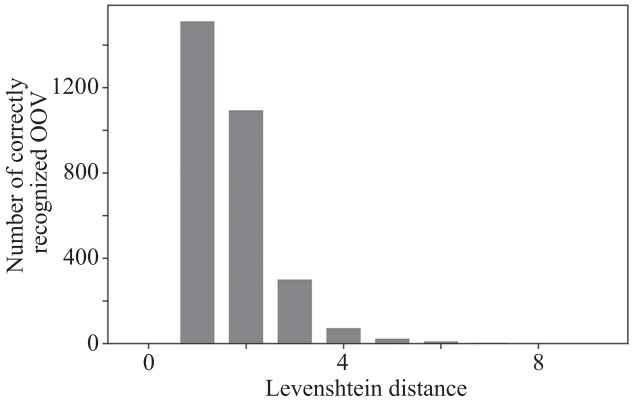

Fig. 2. Distribution of the number of correctly recognized OOV words depending on the Levenshtein distance to the nearest IV word

Conclusion

This paper considers the problem of OOV word recognition for the Russian language using the end-to-end ASR system. The possibilities of OOV recognition were explored using the beam search decoding algorithm compared to the WFST decoding graph. The proposed methods of using the TG.fst graph and discriminative MMI finetuning allowed reducing the overall WER by 3 % in absolute value compared to the beam search decoding while maintaining a high level of recognition ability for OOV words obtained as a result of the inflection of the Russian language. Our method is also significantly superior to the existing approach based on using the unk graph.

As further research, it is planned to improve the grapheme language model (G graph) by increasing the number of N-grams. It is assumed that the bigger context of the language model will help to deal better with OOV word recognition.

References

1. Hinton G., Deng L., Yu D., Dahl G.E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T.N., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, vol. 29, no. 6, pp. 82–97. https://doi.org/10.1109/MSP.2012.2205597
2. Graves A., Fernandez S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proc. of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 369–376. https://doi.org/10.1145/1143844.1143891
3. Synnaeve G., Xu Q., Kahn J., Likhomanenko T., Grave E., Pratap V., Sriram A., Liptchinsky V., Collobert R. End-to-end ASR: From supervised to semi-supervised learning with modern architectures. *arXiv*, 2019, ArXiv:1911.08460. https://doi.org/10.48550/arXiv.1911.08460
4. Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J.M., Nguyen H., Gadde R.T. Jasper: An end-to-end convolutional neural acoustic model. *Proc. of the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language (INTERSPEECH)*, 2019, pp. 71–75. https://doi.org/10.21437/Interspeech.2019-1819
5. Khokhlov Y., Tomashenko N., Medennikov I., Romanenko A. Fast and accurate OOV decoder on high-level features. *Proc. of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2884–2888. https://doi.org/10.21437/Interspeech.2017-1367
6. Alumaë A., Tilk O., Ullah A. Advanced rich transcription system for Estonian speech. *Frontiers in Artificial Intelligence and Applications*, 2018, vol. 307, pp. 1–8. https://doi.org/10.3233/978-1-61499-912-6-1
7. Braun R., Madikeri S., Motlicek P. A comparison of methods for OOV-word recognition on a new public dataset. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 5979–5983. https://doi.org/10.1109/ICASSP39728.2021.9415124
8. Laptev A., Andrusenko A., Podluzhny I., Mitrofanov A., Medennikov I., Matveev Y. Dynamic acoustic unit augmentation with BPE-dropout for low-resource end-to-end speech recognition. *Sensors (Basel)*, 2021, vol. 21, no. 9, pp. 3063. https://doi.org/10.3390/s21093063
9. Andrusenko A., Laptev A., Medennikov I. Exploration of end-to-end ASR for OpenSTT — Russian open speech-to-text dataset. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12335, pp. 35–45. https://doi.org/10.1007/978-3-030-60276-5_4
10. An K., Xiang H., Ou Z. CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency. *Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 566–570. https://doi.org/10.21437/Interspeech.2020-2732
11. Hadian H., Sameti H., Povey D., Khudanpur S. End-to-end speech recognition using lattice-free MMI. *Proc. of the 19th Annual Conference of the International Speech Communication, (INTERSPEECH)*, 2018, pp. 12–16. https://doi.org/10.21437/Interspeech.2018-1423
12. Laptev A., Majumdar S., Ginsburg B. CTC variations through new WFST topologies. *Proc. of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022, pp. 1041–1045 https://doi.org/10.21437/Interspeech.2022-10854.
13. Zeyer A., Schlüter R., Ney H. Why does CTC result in peaky behavior? *arXiv*, 2021, arXiv:2105.14849. https://doi.org/10.48550/arXiv.2105.14849.
14. Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G. Common voice: A massively-multilingual speech corpus. *Proc. of the 12th International Conference on Language Resources and Evaluation (LREC)*, 2020, pp. 4218–4222.
15. Park D., Chan W., Zhang Y., Chiu C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. *Proc. of the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language (INTERSPEECH)*, 2019, pp. 2613–2617 https://doi.org/10.21437/interspeech.2019-2680

Литература

1. Hinton G., Deng L., Yu D., Dahl G.E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T.N., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups // IEEE Signal Processing Magazine. 2012. V. 29. N 6. P. 82–97. https://doi.org/10.1109/MSP.2012.2205597
2. Graves A., Fernandez S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks // Proc. of the 23rd International Conference on Machine Learning (ICML). 2006. P. 369–376. https://doi.org/10.1145/1143844.1143891
3. Synnaeve G., Xu Q., Kahn J., Likhomanenko T., Grave E., Pratap V., Sriram A., Liptchinsky V., Collobert R. End-to-end ASR: From supervised to semi-supervised learning with modern architectures // arXiv. 2019. ArXiv:1911.08460. https://doi.org/10.48550/arXiv.1911.08460
4. Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J.M., Nguyen H., Gadde R.T. Jasper: An end-to-end convolutional neural acoustic model // Proc. of the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language (INTERSPEECH). 2019. P. 71–75. https://doi.org/10.21437/Interspeech.2019-1819
5. Khokhlov Y., Tomashenko N., Medennikov I., Romanenko A. Fast and accurate OOV decoder on high-level features // Proc. of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). 2017. P. 2884–2888. https://doi.org/10.21437/Interspeech.2017-1367
6. Alumaë A., Tilk O., Ullah A. Advanced rich transcription system for Estonian speech // Frontiers in Artificial Intelligence and Applications. 2018. V. 307. P. 1–8. https://doi.org/10.3233/978-1-61499-912-6-1
7. Braun R., Madikeri S., Motlicek P. A comparison of methods for OOV-word recognition on a new public dataset // Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2021. P. 5979–5983. https://doi.org/10.1109/ICASSP39728.2021.9415124
8. Laptev A., Andrusenko A., Podluzhny I., Mitrofanov A., Medennikov I., Matveev Y. Dynamic acoustic unit augmentation with BPE-dropout for low-resource end-to-end speech recognition // Sensors (Basel). 2021. V. 21. N 9. P. 3063. https://doi.org/10.3390/s21093063
9. Andrusenko A., Laptev A., Medennikov I. Exploration of end-to-end ASR for OpenSTT - Russian open speech-to-text dataset // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. V. 12335. P. 35–45. https://doi.org/10.1007/978-3-030-60276-5_4
10. An K., Xiang H., Ou Z. CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency // Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH). 2020. P. 566–570. https://doi.org/10.21437/Interspeech.2020-2732
11. Hadian H., Sameti H., Povey D., Khudanpur S. End-to-end speech recognition using lattice-free MMI // Proc. of the 19th Annual Conference of the International Speech Communication, (INTERSPEECH). 2018. P. 12–16. https://doi.org/10.21437/Interspeech.2018-1423
12. Laptev A., Majumdar S., Ginsburg B. CTC variations through new WFST topologies // Proc. of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH). 2022. P. 1041–1045 https://doi.org/10.21437/Interspeech.2022-10854
13. Zeyer A., Schlüter R., Ney H. Why does CTC result in peaky behavior? // arXiv. 2021. arXiv:2105.14849. https://doi.org/10.48550/arXiv.2105.14849
14. Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G. Common voice: A massively-multilingual speech corpus // Proc. of the 12th International Conference on Language Resources and Evaluation (LREC). 2020. P. 4218–4222.
15. Park D., Chan W., Zhang Y., Chiu C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition // Proc. of the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language (INTERSPEECH). 2019. P. 2613–2617 https://doi.org/10.21437/interspeech.2019-2680

16. Gulati A., Qin J., Chiu C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented transformer for speech recognition. *Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015
17. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *Proc. of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.
18. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. The Kaldi speech recognition toolkit. *Proc. of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.
19. Watanabe S., Hori T., Karita S., Hayashi T., Nishitoba J., Unno Y., Soplin N.E.Y., Heymann J., Wiesner M., Chen N., Renduchintala A., Ochiaiet T. ESPnet: End-to-end speech processing toolkit. *Proc. of the 19th Annual Conference of the International Speech Communication (INTERSPEECH)*, 2018, pp. 2207–2211. https://doi.org/10.21437/Interspeech.2018-1456

16. Gulati A., Qin J., Chiu C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented transformer for speech recognition // Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH). 2020. P. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015
17. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Proc. of the 31st Annual Conference on Neural Information Processing Systems (NIPS). 2017. P. 5998–6008.
18. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. The Kaldi speech recognition toolkit // Proc. of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. 2011.
19. Watanabe S., Hori T., Karita S., Hayashi T., Nishitoba J., Unno Y., Soplin N.E.Y., Heymann J., Wiesner M., Chen N., Renduchintala A., Ochiaiet T. ESPnet: End-to-end speech processing toolkit // Proc. of the 19th Annual Conference of the International Speech Communication (INTERSPEECH). 2018. P. 2207–2211. https://doi.org/10.21437/Interspeech.2018-1456

Authors

Andrei Yu. Andrusenko — PhD Student, Scientific Researcher, STC-Innovations Ltd., Saint Petersburg, 194044, Russian Federation; Software Developer, ITMO University, 197101, Russian Federation, SC 57211637170, https://orcid.org/0000-0002-8697-832X, andrusenkoau@itmo.ru

Aleksei N. Romanenko — PhD, Leading Researcher, STC-Innovations Ltd., Saint Petersburg, 194044, Russian Federation; Senior Researcher, ITMO University, 197101, Russian Federation, SC 56414341400, https://orcid.org/0000-0001-6267-018X, AlexeySk8@gmai.com

Авторы

Андрусенко Андрей Юрьевич — аспирант, научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57211637170, https://orcid.org/0000-0002-8697-832X, andrusenkoau@itmo.ru

Романенко Алексей Николаевич — кандидат технических наук, ведущий научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56414341400, https://orcid.org/0000-0001-6267-018X, AlexeySk8@gmai.com

Received 22.08.2022
Approved after reviewing 25.10.2022
Accepted 22.11.2022

Статья поступила в редакцию 22.08.2022
Одобрена после рецензирования 25.10.2022
Принята к печати 22.11.2022