




Investigation of Transfer Learning for End-to-End Russian Speech Recognition

Irina Kipyatkova^(✉) 

St. Petersburg Federal Research Center of the Russian
Academy of Sciences (SPC RAS), St. Petersburg, Russia
kipyatkova@iiias.spb.su

Abstract. End-to-end speech recognition systems reduce the speech decoding time and required amount of memory comparing to standard systems. However they need much more data for training, which complicates creation of such systems for low-resourced languages. One way to improve performance of end-to-end low-resourced speech recognition system is model's pre-training by transfer learning, that is training the model on the non-target data and then transferring the trained parameters to the target model. The aim of the current research was to investigate application of transfer learning to the training of the end-to-end Russian speech recognition system in low-resourced conditions. We used several speech corpora of different languages for pre-training. Then end-to-end model was fine-tuned on a small Russian speech corpus of 60 h. We conducted experiments on application of transfer learning in different parts of the model (feature extraction block, encoder, and attention mechanism) as well as on freezing of the lower layers. We have achieved 24.53% relative word error rate reduction comparing to the baseline system trained without transfer learning.

Keywords: End-to-end speech recognition · Transfer learning · Encoder-decoder · Russian speech

1 Introduction

In recent years, developing of end-to-end systems became the main trend in researches on automatic speech recognition (ASR). End-to-end ASR systems transform an input speech signal to a sequence of letters using single deep neural network (DNN). This results in reducing the processing time and required amount of memory comparing to standard ASR systems consisting of independent components. However, training the end-to-end system requires much more data than standard system. This drawback makes it difficult to create an end-to-end ASR system for low-resourced languages. One way to overcome this drawback is to use transfer learning methods.

Transfer learning consists in transferring the knowledge obtained on one or several initial tasks to be used for improving the training on target task. There are several ways of application of transfer learning. The most common of them are [1, 2]: (1) instances-based (instances of non-target domain are added to target train dataset with appropriate

weight); (2) feature-based, which can be asymmetric (original features are transformed to match the target features) and symmetric (source and target features are transformed into a new feature representation); (3) mapping-based (instances from target and non-target domains are mapped into a new data space with better similarity); (4) network-based (the pre-trained network including its structure and parameters is transferred to target domain with its subsequent fine-tuning on the target data); (5) adversarial-based (the adversarial technology is used to find transferable features that are both suitable for two domains).

Transfer learning is very effective at training DNNs. In the view of speech recognition, the idea of transfer learning is based on the fact that features learned by lower layers of DNN do not depend on language while language specific features are learned by higher layers [3]. At creating the end-to-end ASR for under-resourced language, transfer learning is mostly performed by pre-training the model on data of non-target language and then fine-tuning the model on data of the target language. The parameters of low layers of DNN can be frozen that means that they are not updated during fine-tuning. The transfer learning scheme is presented on Fig. 1.

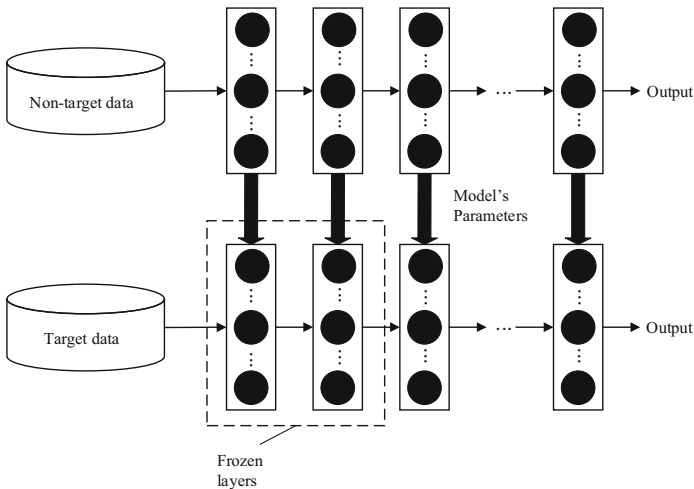


Fig. 1. The scheme of network-based transfer learning method.

The aim of this research was to explore transfer learning method for training the end-to-end Russian speech recognition system in low-resourced conditions. We have tried several languages for pre-training the model and we have investigated influence of low layer freezing on speech recognition results. The rest of the paper is organized as follows. In Sect. 2 we give a brief survey of the researches in which transfer learning was used for the training of the ASR system, in Sect. 3 we describe our end-to-end Russian speech recognition system with transfer learning, the experimental results are given in Sect. 4, in Sect. 5 we make a conclusion to our work.

2 Related Work

There are many scientific researches on application of transfer learning for training the ASR systems. One of the earliest methods of transfer learning in ASR is a tandem approach [4]. In the tandem approach, at first DNN with bottleneck is trained and then the parameters of bottleneck output are used in standard Hidden Markov model (HMM) based system or hybrid DNN/HMM system [5].

Recently, transfer learning is mostly used for training the hybrid HMM/DNN and end-to-end systems. For example, transfer learning was applied to train the acoustic models for two Tibetan dialects with usage of Mandarin as non-target language in [6]. In [7], German ASR system based on convolutional neural network was trained using transfer learning from the model of English speech trained on Librispeech corpus, with the lower layers of the network being frozen. The influence of freezing of low layer parameters on results of end-to-end speech recognition was researched in [8]. The authors performed experiments on German and Swiss German, with English being used for pre-training. The experiments have shown that freezing of the low layers results in increasing of speech recognition accuracy and reduction of training time. Significant improvement of the accuracy was achieved when the first layer was frozen. The freezing of the higher layers did not lead to recognition accuracy increasing.

The research on training of the hybrid DNN/HMM children speech recognition system using transfer learning was presented in [9]. The adult speech database was used for pre-training. The authors obtained 16.5% relative reduction of WER comparing to the baseline system with Speaker Adaptive Training technique.

In the paper [10], feature transfer learning was performed. At first, the encoder's lower layers predicting spectral features on the raw waveform were trained. Then trained parameters were transferred to the attention-based encoder-decoder model.

In [11], the transfer learning method called teacher-student was used for initialization of parameters of online speech recognition system by parameters obtained at training of the large offline end-to-end system. The teacher-student learning is an approach to transfer the knowledge from a large deep ("teacher") network to shallower model [12]. The student neural network is trained to minimize difference between its own output distributions and teacher network's distributions.

It also should be noted that transfer learning can also be realized as multi-task learning in a multilingual system. Such approach was realized, for example in [13, 14]. In [15] a technique for DNN-based acoustic model adaptation to specific domain in multilingual system was proposed. It performs adaptation of low-resourced language system trained for one source domain into a target domain using adaptation data of high-resourced language.

In the current paper we consider the training of monolingual ASR system in low-resourced condition.

3 End-to-End Speech Recognition Model with Transfer Learning

3.1 Architecture of the End-to-End Speech Recognition Model

We used joint CTC-attention based encoder-decoder model similar to the model proposed in [16]. Our model was described in detail in [17]. Encoder was Bidirectional Long

Short-Term Memory (BLSTM) network contained five layers with 512 cells in each with highway connections [18]. Decoder was Long Short-Term Memory (LSTM) network contained two layers with 512 cells in each. Location-aware [19] attention mechanism was used in decoder. Before the encoder, there was a feature extraction block that was VGG [20] model with residual connection (ResNet). At the training stage, the CTC weight was equal to 0.3. Filter banks features were used as input.

At the decoding stage, we additionally used LSTM-based language model (LM), which was trained on text corpus of about 350M words. The text corpus was collected from online Russian newspapers. LSTM contained one layer with 512 cells. The vocabulary consisted of 150K most frequent word-forms from the training text corpus.

For training and testing the end-to-end Russian speech recognition model we used ESPnet toolkit [21] with a PyTorch as a back-end part.

3.2 Application of Transfer Learning at Model’s Training

Transfer learning was carried out by pre-training the model on non-target speech data, transferring the trained parameters of neural network to the target model and the following training the model on Russian speech data.

The first step was to choose speech corpora for pre-training. The main criteria for selection the speech corpora were the following: (1) speech data duration of more than 100 hours; (2) sentence-level segmentation; (3) availability of transcripts. We chose five speech corpora of non-target languages which are presented in Table 1. Among these corpora there is a corpus of Ukrainian speech which does not meet the requirement of duration. However, we decided to use this corpus as well because Ukrainian language is related to Russian, so we hypothesized that pre-training on these speech data may be useful.

Table 1. Characteristics of speech corpora used for pre-training.

Language	Speech corpus	Duration	Description
English	LibriSpeech (clean) [22]	360 h	Read audiobooks
Italian	TEDx [23]	107 h	Recordings of TEDx talks
Catalan	ParlamentParla v1.0 [24]	320 h	Recordings of the Catalan Parliament plenary sessions
German	M-AILABS German Corpus [25]	237 h 22 m	Read audiobooks
Ukrainian	M-AILABS Ukrainian Corpus [26]	87 h 08 m	Read audiobooks

The weights obtained from the model trained on non-target data were used for initialization of weights of the feature extraction block, encoder, and attention mechanism. Then, we conducted experiments on freezing parameters of low layers at transfer

learning. Architecture of our end-to-end model with transfer learning is presented on Fig. 2.

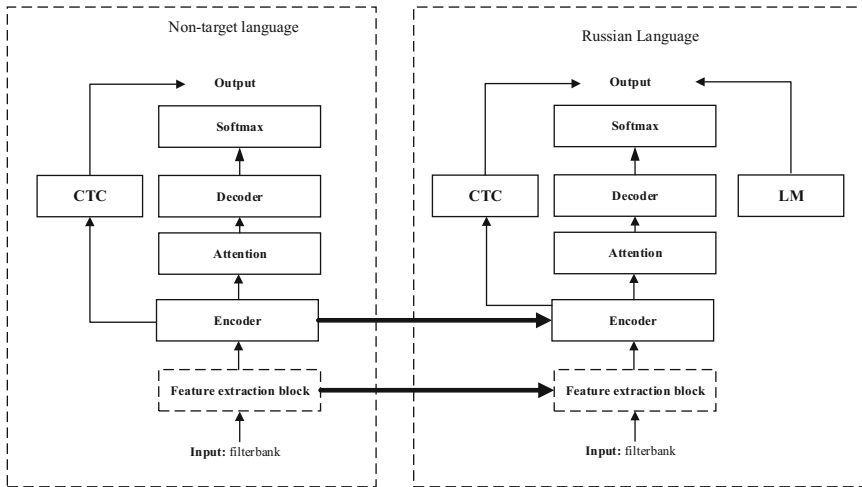


Fig. 2. Architecture of end-to-end speech recognition model with transfer learning.

The end-to-end model was trained on Russian speech data composed from the speech corpus collected at SPC RAS [17] as well as free speech corpora Voxforge [26] and M-AILABS [25]. The corpus collected at SPC RAS consists of the recordings of phonetically rich and meaningful phrases and texts, also it includes commands for the MIDAS information kiosk [27] and 7–digits telephone numbers. As a result we had 60.6 h of speech data. This speech dataset was splitted into validation and trains parts with sizes of 5% and 95%.

4 Experiments

Experiments on continuous Russian speech recognition were performed on our test speech corpus consisting of 500 phrases pronounced by 5 speakers. The phrases were taken from online newspaper which was not used for LM training. During experiments we used beam search pruning method similar to the approach proposed in [28] and substituted softmax with gumbel-softmax [29]. The decoding setup is described in [30].

We have obtained $CER = 14.9\%$ and $WER = 37.1\%$ without usage of transfer learning. The results obtained after application of transfer learning using different languages for pre-training are presented in Table 2. In the table, “Init.” means that pretrained parameters were used only for initialization of the parameters of the model’s block without their freezing.

Table 2. Experimental results on Russian speech recognition using different non-target languages for transfer learning (%).

Non-target language	Transfer learning scheme						Experimental results	
	Feature extraction			Encoder		Attention	CER, %	WER, %
	Init	1 layer frozen	2 layer frozen	Init	1 layer frozen	Init		
No transferring (baseline)	–	–	–	–	–	–	14.9	37.1
English	+	–	–	–	–	–	13.9	34.5
	+	+	–	–	–	–	13.9	35.7
	–	–	–	+	–	–	13.2	34.8
	+	–	–	+	–	–	10.5	28.0
	+	–	–	+	+	–	11.3	30.0
	+	–	–	+	–	+	10.2	28.6
Italian	+	–	–	–	–	–	14.2	35.0
	+	+	–	–	–	–	12.5	32.1
	+	+	+	–	–	–	14.7	36.6
	–	–	–	+	–	–	12.6	29.5
	+	+	–	+	–	–	12.7	31.2
	+	+	–	+	+	–	12.8	31.8
	+	+	–	+	–	+	12.4	30.3
	+	–	–	–	–	–	14.8	38.5
Catalan	+	+	–	–	–	–	14.0	36.9
	+	+	+	–	–	–	14.4	37.4
	–	–	–	+	–	–	13.5	32.9
	+	+	–	+	–	–	12.1	32.3
	+	+	–	+	+	–	12.9	32.5
	+	+	–	+	–	+	11.4	31.0
	+	–	–	–	–	–	14.8	38.2
German	+	+	–	–	–	–	15.3	36.6
	+	+	+	–	–	–	14.4	37.0
	–	–	–	+	–	–	13.6	33.0
	+	+	–	+	–	–	13.4	31.2
	+	+	–	+	+	–	13.9	35.0
	+	+	–	–	–	–	14.8	38.2

(continued)

Table 2. (continued)

Non-target language	Transfer learning scheme						Experimental results	
	Feature extraction			Encoder		Attention	CER, %	WER, %
	Init	1 layer frozen	2 layer frozen	Init	1 layer frozen	Init		
	+	+	–	+	–	+	13.4	33.6
Ukrainian	+	–	–	–	–	–	15.7	37.4
	+	+	–	–	–	–	15.9	37.9
	–	–	–	+	–	–	14.3	34.8
	+	–	–	+	–	–	12.2	31.9
	+	–	–	+	+	–	13.3	32.9
	+	–	–	+	–	+	12.1	29.9

We conducted a series of experiments on application of transfer learning in different parts of the model. Transferring neural network parameters from non-target model for initialization of feature extraction block only slightly decreased recognition error and in some cases even slightly increased it that can be connected with statistic fluctuation. Freezing of first layer of feature extraction block resulted in reduction of CER and WER when Italian, Catalan, and German languages were used. Freezing higher layers did not lead to a decrease in recognition error. Application of transfer learning for initialization of encoder’s parameters decreased recognition error in all cases. Then we conducted experiments on transfer parameters of both encoder and feature extraction block, with first layer of feature extraction block being frozen when transferring from language with which freezing gave better result than just initialization. Freezing the first layer of encoder increased recognition error, therefore we did not perform experiments on freezing the higher layers. Then transfer learning was carried out in attention mechanism. In most cases (except English) application of transfer learning for initialization of parameters in attention mechanism in addition to encoder and feature extraction block gave additional improvement of the result.

The best result (WER = 28.0) was achieved when English was used as non-target language and transfer learning was applied for initialization of parameters of both feature extraction block and encoder. This may be due to the fact that the English corpus was the largest that we used for pre-training. It should also be noted that usage of Ukrainian language gave us the result comparable to usage of other non-target languages although the size of the Ukrainian corpus was significantly smaller. This can be due to the fact that Russian and Ukrainian are related languages. Therefore we can draw a conclusion that in low-resourced condition the usage of other low-resourced language related to the target language can improve speech recognition result.

5 Conclusions and Future Work

In the paper, we have investigated the application of speech data of different non-target languages for pre-training of the end-to-end Russian speech recognition system. The best results were achieved when parameters were transferred from the model pre-trained on English speech for initialization of parameters of the feature extraction block and encoder. In this case relative reduction of WER was 24.53%. The further researches will be connected with enlarging the training data and experimenting with other architectures of neural network for Russian end-to-end speech recognition, for example, Transformer.

Acknowledgements. This research was supported by the state research № FFZF-2022-0005.

References

1. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27
2. Zhuang, F., et al.: A comprehensive survey on transfer learning. *Proc. IEEE* **109**(1), 43–76 (2020)
3. Wang, D., Zheng, T.F.: Transfer learning for speech and language processing. In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1225–1237 (2015)
4. Grézl, F., Karafiát, M., Kontár, S., Cernocký, J.: Probabilistic and bottle-neck features for LVCSR of meetings. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2007), pp. IV-757–IV-760 (2007)
5. Yan, Z.J., Huo, Q., Xu, J.: A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR. In: Proceedings of Interspeech-2013, pp. 104–108 (2013)
6. Yan, J., Lv, Z., Huang, S., Yu, H.: Low-resource tibetan dialect acoustic modeling based on transfer learning. In: Proceedings of SLTU. pp. 6–10 (2018)
7. Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., Stober, S.: Transfer learning for speech recognition on a budget. ArXiv preprint [arXiv:1706.00290](https://arxiv.org/abs/1706.00290) (2017). <https://arxiv.org/abs/1706.00290>
8. Eberhard, O., Zesch, T.: Effects of layer freezing on transferring a speech recognition system to under-resourced languages. In: Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), pp. 208–212 (2021)
9. Shivakumar, P.G., Georgiou, P.: Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. *Comput. Speech Lang.* **63**, 101077 (2020)
10. Tjandra, A., Sakti, S., Nakamura, S.: Attention-based wav2text with feature transfer learning In: Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 309–315 (2017)
11. Kim, S., Seltzer, M.L., Li, J., Zhao, R.: Improved training for online end-to-end speech recognition systems. In: Proceedings of Interspeech-2018, pp. 2913–2917 (2018)
12. Li, J., Zhao, R., Huang, J.-T., Gong Y.: Learning small-size DNN with output-distribution-based criteria. In: Proceedings of Interspeech-2014, pp. 1910–1914 (2014)

13. Tachbelie, M.Y., Abate, S.T., Schultz, T.: Multilingual speech recognition for GlobalPhone languages. *Speech Commun.* **140**, 71–86 (2022)
14. Qin, C.-X., Qu, D., Zhang, L.-H.: Towards end-to-end speech recognition with transfer learning. *EURASIP J. Audio, Speech, Music Process.* **2018**(1), 1–9 (2018). <https://doi.org/10.1186/s13636-018-0141-9>
15. Abad, A., Bell, P., Carmantini, A., Renais, S.: Cross lingual transfer learning for zero-resource domain adaptation. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020)*, pp. 6909–6913 (2020)
16. Kim, S., Hori, T., Watanabe, S.: Joint ctc-attention based end-to-end speech recognition using multi-task learning. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2017)*, pp. 4835–4839 (2017)
17. Kipyatkova, I., Markovnikov, N.: Experimenting with attention mechanisms in Joint CTC-attention models for Russian speech recognition. In: Karpov, A., Potapova, R. (eds.) *SPECOM 2020. LNCS (LNAI)*, vol. 12335, pp. 214–222. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60276-5_22
18. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. *arXiv preprint arXiv:1505.00387* (2015). <https://arxiv.org/abs/1505.00387>
19. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. *Adv. Neural. Inf. Process. Syst.* **28**, 577–585 (2015)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *ArXiv preprint arXiv:1409.1556* (2014). <https://arxiv.org/abs/1409.1556>
21. Watanabe, S., et al.: Espnet: End-to-end speech processing toolkit. In: *INTERSPEECH-2018*, pp. 2207–2211 (2018)
22. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2015)*, pp. 5206–5210 (2015)
23. Salesky, E., et al.: The multilingual TEDx corpus for speech recognition and translation. In: *Proceedings of Interspeech-2021*, pp. 3655–3659 (2021)
24. Külebi, B., Armentano-Oller, C., Rodríguez-Penagos, C., Villegas, M.: ParlamentParla: A speech corpus of catalan parliamentary sessions. In: *Workshop on Creating, Enriching and Using Parliamentary Corpora*, pp. 125–130 (2022)
25. The m-ailabs speech dataset. <https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/>. Accessed 30 Jun 2022
26. VoxForge. <http://www.voxforge.org/>. Accessed 30 Jun 2022
27. Karpov, A.A., Ronzhin, A.L.: Information enquiry kiosk with multimodal user interface. *Pattern Recogn. Image Anal.* **19**(3), 546–558 (2009)
28. Freitag, M., Al-Onaizan, Y.: Beam search strategies for neural machine translation. *ArXiv preprint arXiv:1702.01806* (2017). <https://arxiv.org/abs/1702.01806>
29. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. *ArXiv preprint arXiv:1611.01144* (2016). <https://arxiv.org/abs/1611.01144>
30. Markovnikov, N., Kipyatkova, I.: Investigating joint CTC-attention models for end-to-end russian speech recognition. In: Salah, A.A., Karpov, A., Potapova, R. (eds.) *SPECOM 2019. LNCS (LNAI)*, vol. 11658, pp. 337–347. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26061-3_35