# Self-Supervised Training for the Tatar Speech Recognition System

Aidar Khusainov [1], Dzhavdet Suleymanov [1] and Ilnur Mukhametzyanov [1]

[1] *Tatarstan Academy of Sciences, Kazan, Russia*

**Abstract**

In this paper, we show the results of experiments on the creation of the Tatar speech recognition system based on self-supervised learning. We constructed a 328-hour unlabeled and a 129-hour annotated corpora. As the base model, we used the multilingual XLSR model, pre-trained it on collected unlabeled data, and then fine-tuned it on an annotated corpus. The resulting accuracy on the Common Voice (read speech) test dataset is WER 6.48%, on the Tatar Corpus (read clean speech) is 6.46%, and for the spontaneous speech dataset collected from the TV shows is 28.73%, all of the results are the best-published results on these datasets.

**Keywords**

Iterative self-training, speech recognition, the Tatar language

## 1. Introduction

Research in the field of automatic speech recognition is actively developing in several main areas: systems are becoming more robust to various background noises, dialects, features of pronunciation, and approaches are being developed for working with low-resource languages. And if a few years ago the absolute majority of recognition systems were based on the "classical" approach of dividing into acoustic models, a pronunciation model, and a language model, recently end-to-end systems (E2E) have proved the ability to show better recognition quality.

The use of E2E recognition systems allows obtaining a better result, however, they require a large amount of training data, which is not available for low-resource languages. One way to overcome the lack of training data is to pre-train the system on data for related languages or use a model that has been trained for another language with a lot of annotated data, for example, for English.

In addition, following the Computer Vision and Natural Language Processing areas models pre-trained on a large amount of unlabeled data are more actively used in the field of speech technologies. In the field of speech analysis, this approach was implemented within the wav2vec2 model, which made it possible to obtain high-quality results for the English language with a minimum amount (from 10 minutes of records) of labeled data [1]. The essence of the technology is to use a large amount of unlabeled data to construct an acoustic representation of the speech signal samples.

In this paper, we describe the results of experiments on the construction of a Tatar speech recognition system based on the technology of preliminary self-learning on unlabeled data. We created an unlabeled audio corpus for the Tatar language with a total volume of 340 hours and an annotated speech corpus with a volume of 128 hours. An experiment was carried out in which the multilingual model XLSR [2], trained on 56 thousand hours of speech in 53 languages, was sequentially pretrained using unlabeled Tatar data and fine-tuned on the annotated Tatar corpus. The testing of the proposed speech recognition system confirmed the high quality of work for different types of speech (read and spontaneous) and noise conditions. The stage of self-learning on unlabeled data made it possible to reduce the value of recognition errors WER (word error rate) for the Common Voice test subcorpus from 26.76 [3] to 6.48;

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL: khusainov.aidar@gmail.com (A. 1); dvdt.slt@gmail.com (A. 2); iIlnur.mukhametzyanov@gmail.com (A. 3)
ORCID: 0000-0002-7763-1420 (A. 1)

the previous best-published result for the Tatar Corpus read speech corpus of 12.89 WER [4] has been reduced to 6.46.

## 2. Approach to train the ASR system

This article uses an approach with iterative retraining of the basic multilingual model XLSR [2]. At the first stage, a self-supervised learning approach is applied, which consists of solving a task that does not require manual annotation of the corpus. In this case, the CPC (Contrastive Predictive Coding) criterion is used, and the model solves the classification problem by determining the following fragment from other fragments [5, 6, 7]. In [8], it is shown that the characteristics of audio, revealed by the model in the process of solving the task, demonstrate robustness to changes in the domain and even the language of speech. An illustration of the model from the original article wav2vec2 [1] is shown in Fig. 1.

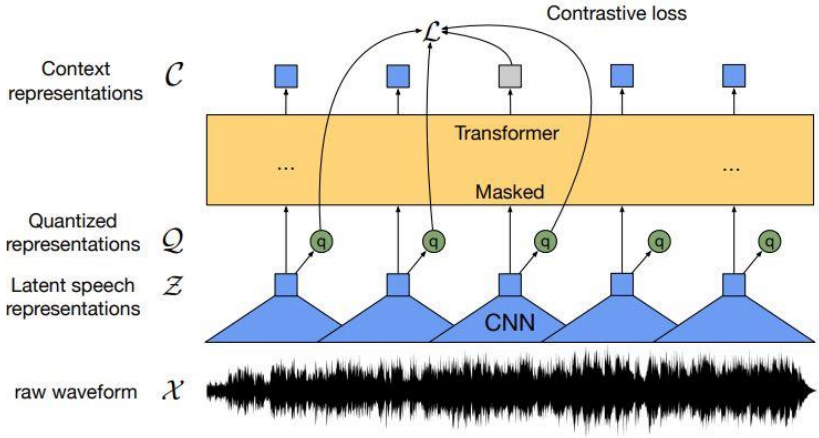


**Figure 1**: An illustration of the work of the wav2vec2 model, which learns the contextual representation of audio fragments based on unlabeled data

At the second stage, the annotated speech corpus is used to retrain the model obtained at the first stage. Additional training is based on the CTC (Connectionist Temporal Classification) algorithm [5, 9]. A randomly initialized layer with a dimension equal to the number of elements in the dictionary is added to the model of the first stage. For the case of the Tatar language, the dictionary consists of 39 elements: 38 letters and an additional words separator character '|'.

## 3. Tatar speech corpora

The created unlabeled Tatar speech corpus consists of 4 parts:

1. Audiobooks (read speech recorded in studio conditions);
2. TV broadcasting (spontaneous speech, variety of external noises, background music);

Two radio stations' recordings (read and spontaneous speech, background music);

3. Scientific video lectures from the YouTube platform (continuous speech, good recording quality).

At the preprocessing stage, an audio track was extracted from the video file, all audio files were then converted to 16 bits per sample, 16 kHz mono PCM format.

Taking into account the specifics of the initial data (long audiobooks, 12-hour fragments of TV shows, 40-minute YouTube clips), the task was to divide audio files into short fragments containing speech. In this case, the goal was to determine fragments that contain the speech of only one speaker. To solve this problem, we used the Silero-VAD tool [10]. Selective analysis of fragments showed that the model coped with filtering music content that was present in radio and TV air while retaining speech segments with background music.

Based on the recommendations of the developers of the wav2vec2 model [11], short (less than 4.5 seconds) and long (longer than 30 seconds) audio files were filtered. The summary statistics on the number of files and their duration for each subcorpus are presented in Table 1.

**Table 1**
Statistics of the unlabeled corpus of Tatar speech

| Corpus | Initial data | Speech fragments | Filtered fragments |
|---|---|---|---|
| Audiobooks | 114:28:10 (520 files) | 105 hours (36 712 files) | 58 hours (17 563 files) |
| TV | 732:59:29 (62 files) | 472 hours (263 466 files) | 202 hours (67 065 files) |
| Radio | 215:36:04 (398 files) | 146 hours (29 778 files) | 29 hours (8 941 files) |
| YouTube lectures | 87:18:42 (100 files) | 81 hours (31 437 files) | 39 hours (12 764 files) |
| Overall | 1 150:22:25 | 804 hours | 328 hours |

The Tatar annotated speech corpus, which was used for additional training of the model, consists of 3 parts:

1. Corpus of read Tatar speech “Tatar Corpus” [12];
2. Annotated fragments of TV broadcast;
3. The Tatar part of the Common Voice corpus [13].

The statistics on the duration of each of the subcorpora are presented in Table 2.

**Table 2**
Statistics of the annotated corpus of Tatar speech

| Corpus | Duration |
|---|---|
| Tatar Corpus | 99:09:59 |
| TV | 1:33:28 |
| Common Voice | 28:47:26 |
| Overall | 129:30:53 |

A test subcorpus with a total duration of 7 hours was formed from the final corpus. We randomly selected the 1 hour and 37 minutes’ recordings of 10 speakers (5 male, 5 female) from “Tatar Corpus”. For the Common Voice corpus, we adopted the division into training and test parts, proposed by the creators of the corpus. For the subcorpus of TV broadcasts, annotations on the speakers were not prepared, so the selection of test fragments was carried out randomly throughout the collection.

## 4. Language model

As a language model for the speech recognition system, a 4-gram statistical model was built based on the KenLM tool [14]. The Internet corpus of Tatar texts was collected as a text corpus for building a language model. Archives of leading news agencies, newspapers, magazines, websites of state institutions and departments, forums were downloaded and processed. The training corpus also included parts of the national corpus of the Tatar language "Tugan Tel" [15]. The total amount of training data was 8,760,330 sentences containing 116 million words.

## 5. Experiments

The experiments were carried out via the fairseq platform [16]. The XLSR model was chosen as the architecture of the neural network for pre-training: the encoder block consisted of 24 layers with a dimension of 1024, the number of attention blocks was 16, and no dropout was used. Pre-training was carried out on 8 V100 32 GB video cards for 4 days. Fig. 2 shows train_loss and valid_loss parameters’ values during the training.

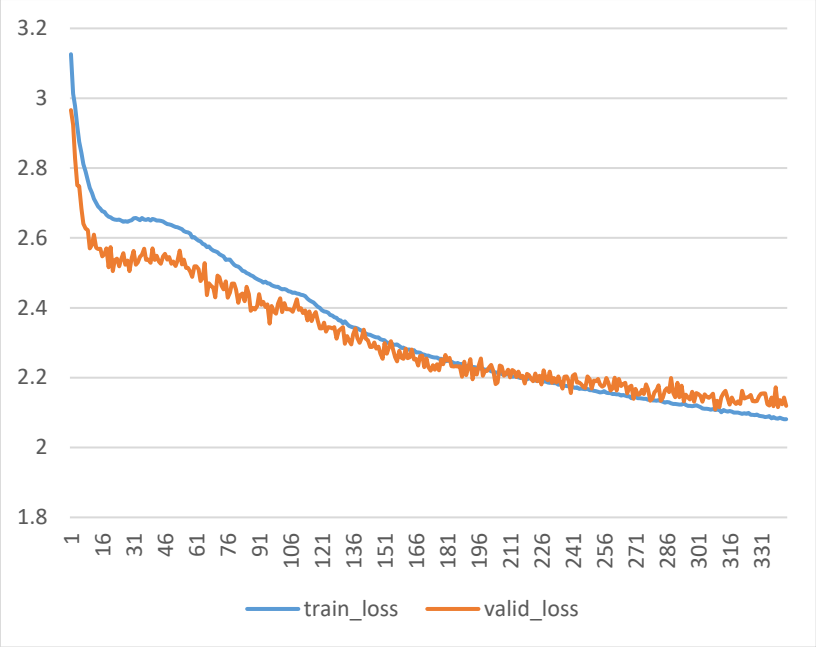


**Figure 2**: The values of the train_loss and valid_loss parameters during the pre-training of the model on unlabeled data

Additional training of the model on an annotated corpus was carried out with the learning rate parameter 3e-5; Adam [17] was used as an optimization algorithm. As in the original work of wav2vec2.0, only the last layer of the neural network was trained during the first 10 thousand iterations; further training also involved the parameters of the transformer layers. The feature extraction part remained unchanged during the entire training time.

Additional training was also carried out on 8 V100 32 GB video cards, the total calculation time was 26 hours, the calculation time to obtain the optimal model was 9 hours. Fig. 3 shows changes in the parameters train_loss and valid_loss values during training.

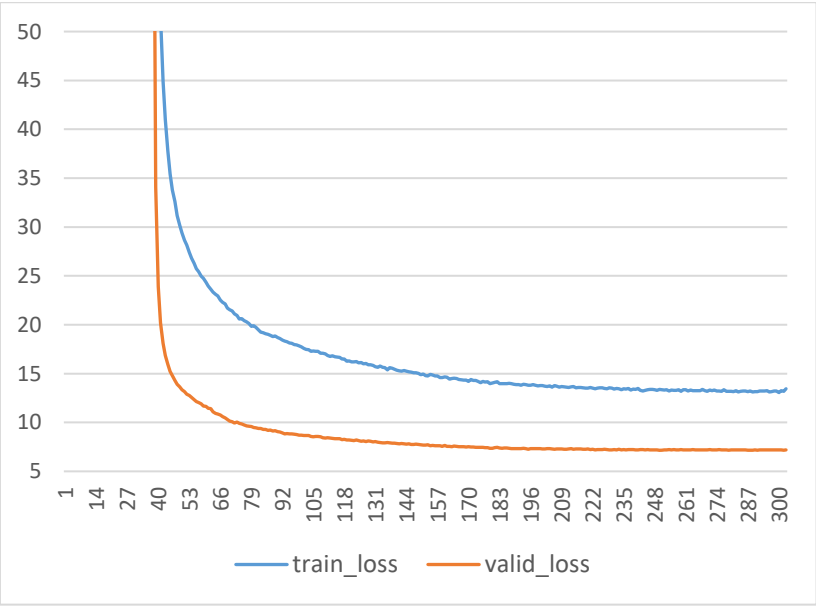


**Figure 3**: The values of the train_loss and valid_loss parameters in the process of additional training of the model on the annotated corpus

The quality of the final Tatar speech recognition system was calculated on the test corpus and equals 6.89 WER. The recognition quality values were calculated separately for the test part of Common Voice and Tatar Corpus in order to compare with other previously published models.

The best value on the Tatar Corpus test corpus shown by the "classical" continuous speech recognition system, built on separate acoustic models, a pronunciation model, and a language model, is equal to 12.89 WER [4]. The model proposed in this work on the same test subcorpus showed a value of 6.46 WER.

The recognition quality of the system [18] was taken as the baseline for the Common Voice test dataset. The best value presented there is 26.76 WER, while our proposed system showed a value of 6.48 WER. The summary values of the test results are presented in Table 3.

**Table 3**
Results of testing the quality of the Tatar speech recognition system

| Test subcorpus | Model | Recognition error, WER |
|---|---|---|
| Common Voice test (read speech) | This work | **6.48** |
| | Previous best work [3] | 26.76 |
| TV (spontaneous speech) | This work | **28.73** |
| Tatar Corpus test (read speech) | This work | **6.46** |
| | Previous best work [4] | 12.89 |
| Overall | This work | **6.89** |

## 6. Conclusions

This paper presents the results of experiments on building a speech recognition system using a self-learning approach on unlabeled audio data. The necessary annotated and unlabeled speech corpora were formed. The results of testing the trained recognition models showed an improvement in the recognition quality of Tatar speech test databases.

## 7. References


[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, Wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Proc. NeurIPS, 2020.

[2] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, Unsupervised cross-lingual representation learning for speech recognition, 2020. URL: http://arxiv.org/abs/ 2006.13979.

[3] Wav2Vec2-Large-XLSR-53-Tatar, 2021. URL: https://huggingface.co/anton-l/wav2vec2-large-xlsr-53-tatar.

[4] A. Khusainov, Recent Results in Speech Recognition for the Tatar Language, volume 10415 of Lecture Notes in Artificial Intelligence, 2017, pp. 183-191. doi: 10.1007/978-3-319-64206-2 21.

[5] A. Baevski, M. Auli, and A. Mohamed, Effectiveness of selfsupervised pre-training for speech recognition, 2019. URL: http://arxiv.org/abs/1911.03912.

[6] S. Schneider, A. Baevski, R. Collobert, and M. Auli, Wav2vec: Unsupervised Pre-Training for Speech Recognition, in: Proc. Interspeech 2019, 2019, pp. 3465–3469.

[7] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E.Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, Libri-light: A benchmark for asr with limited or no supervision, 2019. URL: http://arxiv.org/abs/1912.07875.

[8] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord. Learning robust and multilingual speech representations, 2020. URL: https://arxiv.org/abs/2001.11128.

[9] A. Graves, S. Fernández, and F. Gomez, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: Proc. of ICML, 2006.

[10] Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. URL: https://github.com/snakers4/silero-vad.
[11] FAIR-seq, Wav2vec 2.0 PyTorch example. URL: https://github.com/pytorch/fairseq/tree/master/examples/wav2vec.
[12] Khusainov, A., Design and creation of speech corpora for the Tatar speech recognition and synthesis tasks, in: Proceedings of the 3rd International Conference on Turkic Languages Processing, Kazan, September 17-19, 2015, pp. 475-484.
[13] CommonVoice. URL: https://commonvoice.mozilla.org/.
[14] Heafield, Kenneth, KenLM: Faster and Smaller Language Model Queries, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011, vol. 7, pp. 187—197.
[15] Suleymanov D., Khakimov B., Gilmullin R., Korpus tatarskogo yazyka: konceptualnye i lingvisticheskie aspekty, Vestnik TGGPU 4 (26), 2011, pp. 211-216.
[16] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, Aairseq: A fast, extensible toolkit for sequence modeling, in: Proc. of NAACL System Demonstrations, 2019.
[17] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization, in: ICLR-2015, 2015.
[18] CommonVoice Tatar Benchmark. URL: https://paperswithcode.com/sota/speech-recognition-on-common-voice-tatar.