Incorporation of Iterative Self-supervised Pre-training in the Creation of the ASR System for the Tatar Language

Aidar Khusainov(✉), Dzhavdet Suleymanov, and Ilnur Muhametzyanov

Tatarstan Academy of Sciences, Kazan, Russia
http://antat.ru/ips

Abstract. In this paper, we study the iterative self-supervised pretraining procedure for the Tatar language speech recognition system. The complete recipe includes the use of base pre-trained model (the multilingual XLSR model or the Librispeech (English) Wav2Vec 2.0 Base model), the next step was a "source" self-supervised pre-training on collected Tatar unlabeled data (mostly broadcast audio), then the resulting model was used for additional "target" self-supervised pretraining on the annotated corpus (target domain, without using labels), and the final step was to fine-tune the model on the annotated corpus with labels. To conduct the experiments we prepared a 328-h unlabeled and a 129-h annotated audio corpora. Experiments on three datasets (two proprietary and publicly available Common Voice as the third one) showed that the first "source" pretraining step allows ASR models to show on average 24.3% lower WER, and both source and target pretraining - 33.3% lower WER than a simple finetunes base model. The resulting accuracy for the Common Voice (read speech) test dataset is WER 5.37%, on the private TatarCorpus (read clean speech) is 4.65%, and for the spontaneous speech dataset collected from the TV shows is 22.6%, all of the results are the best-published results on these datasets. Additionally, we show that using a multilingual base model can be beneficial for the case of fine-tuning (10.5% less WER for this case), but applying self-supervised pretraining steps eliminates this difference.

Keywords: Iterative pretraining · Self-supervised learning · Speech recognition · The Tatar language

1 Introduction

Recent results in many domains like NLP and Computer Vision benefited from the use of self-supervised pretraining method, which can be described as a process of learning robust universal representations based on unlabeled datasets. 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 min of records) of labeled data [5]. The idea of the technology is to use a large amount of unlabeled data to construct an acoustic representation of the speech

K. Ekštein et al. (Eds.): TSD 2021, LNAI 12848, pp. 481–488, 2021.
https://doi.org/10.1007/978-3-030-83527-9_41

signal samples. Wav2Vec2 model solves a problem that does not require manual annotation of the corpus. It uses the CPC (Contrastive Predictive Coding) criterion, and the model needs to distinguish the true speech representation from distractors that are uniformly sampled from other masked time steps of the same utterance [6,9,14]. In [10], it is shown that features, revealed by the model in the process of solving this problem, demonstrate robustness to changes in the domain and the language. An illustration of the model from the original article wav2vec2 [5] is shown in Fig. 1.

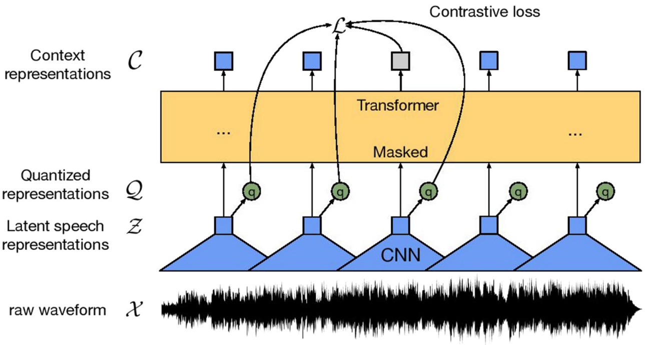
Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
Masked
...
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Fig. 1. An illustration of the work of the wav2vec2 model, which learns the contextual representation of audio fragments based on unlabeled data [5]

And if a few years ago the vast majority of recognition systems were based on the "classical" ASR systems that consist of separate acoustic models, a pronunciation model, and a language model, recently end-to-end systems (E2E) have come to the fore. E2E ASR systems allow 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 to use a model that has been trained for high-resources language with a lot of labeled data. The possible benefits of using the wav2vec2 E2E approach are as follows: systems are becoming more robust to various background noises, dialects, pronunciation features; moreover, for low-resource languages, it's much easier to find a significant amount of unlabeled data.

In this paper, we describe the results of experiments on the creation of Tatar speech recognition systems. We compare different training scenarios, the full scenario consists of 4 training steps:

1. Base self-supervised pretraining (BaseSS).
2. Source self-supervised pretraining (SourceSS).
3. Target self-supervised pretraining (TargetSS).
4. Target fine-tuning (TargetFT).

All scenarios are shown in Fig. 2. In the following sections, we give a training procedure description, provide details of data collection, and present the comparative analysis of the experiments' results.

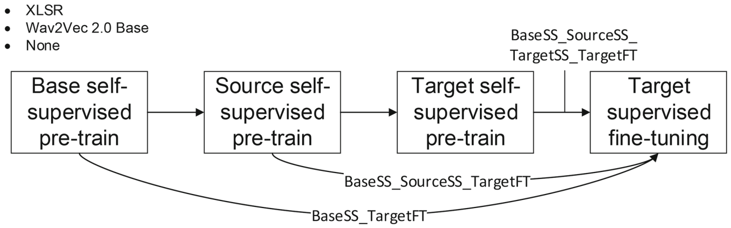
XLSR
Wav2Vec 2.0 Base
None
Base self-supervised pre-train
Source self-supervised pre-train
Target self-supervised pre-train
Target supervised fine-tuning
BaseSS_SourceSS_TargetSS_TargetFT
BaseSS_SourceSS_TargetFT
BaseSS_TargetFT

Fig. 2. Model training options

2 System Description

This article uses an approach with iterative self-supervised pretraining steps on audio data that is increasingly closer to the target domain. We implement 4 main training stages: base self-supervised pretraining (BaseSS), source self-supervised pretraining (SourceSS), target self-supervised pretraining (TargetSS), and target supervised fine-tuning (TargetFT), and analyze the effect of each pretraining step on the resulting recognition quality of ASR systems. The first stage is the BaseSS pretraining step. This step is the initial training where a (very) large dataset is used. The resulting model learned acoustic representation for a wide variety of noise conditions and speakers' variability. For our experiments we have chosen three possible alternatives to use as the base pre-trained model:

1. No pre-trained model.
2. Base Wav2Vec 2.0 Librispeech model (language: English, total duration: 1000 h).
3. Multilingual XLSR model (53 languages, total duration: 56k h).

For the second training step, we use source datasets consisting of heterogeneous Tatar audio data. This data allows the model to start learning language-specific acoustic features with a diverse set of speakers, noise conditions, etc. Data for the SourceSS stage were collected from TV shows, radio transmissions, audiobooks, and YouTube videos. More on data collection procedure can be found in the Data Collection section.

The TargetSS stage performs additional self-supervised training with the target Tatar datasets that have annotations, but they are not used here. We haven't set any hard restrictions on the style of speech for Target datasets due to the small number of available annotated Tatar speech corpora. Therefore, we use all of the existing data including both close-distance microphones read speech and broadcast spontaneous speech.

And at the last stage, the Tatar annotated speech corpus is used to fine-tune the model obtained at the previous stages. Additional training is based on the CTC (Connectionist Temporal Classification) algorithm [6,7]. A randomly initialized layer with a dimension equal to the number of elements in the dictionary is added to the model. For the case of the Tatar language, the dictionary consists of 39 elements: 38 letters and an additional character '—' as a words' separator.

3 Data Collection

The multistage approach that we chose for the training of ASR systems dictates the training data requirements. We need an unlabeled dataset for self-supervised pretraining steps and annotated dataset for supervised fine-tuning. To the moment there are two available datasets for the Tatar language: one from the CommonVoice project [1], and another from the TatarCorpus dataset [11]. Both datasets contain read speech with good SNR, all audios are manually annotated. To collect unlabeled datasets we obtained audios from several sources: a private dataset of audiobooks from Tatar book publishing company, records of TV and radio broadcasting, YouTube videos.

The resulting unlabeled corpus consists of 4 subcorpora:

1. Subcorpus of audiobooks: read speech recorded in studio conditions, 520 files with a total duration of 114 h.
2. Subcorpus of television broadcasting: spontaneous speech, variety of external noises and background music, 62 files - 733 h.
3. Subcorpus of two radio stations' recordings: read and spontaneous speech, background music, 398 files - 215 h.
4. Subcorpus of scientific video lectures from the YouTube platform: mostly read speech, good recording quality, 100 files - 87 h.

We carried out some basic preprocessing of the obtained video and audio files, which included audio track extraction from video files and audio file conversion to 16 bits per sample, 16 kHz mono PCM format. Taking into account the specifics of the initial data (long audiobooks, 12-h fragments of TV snippets, 40-min YouTube clips), the next task was to divide audio files into shorter fragments containing speech. The goal was to convert all data into 5–30 s fragments, where each fragment contains the speech of only one speaker. To solve this problem, we used the Silero-VAD tool [4]. Selective analysis of resulting 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. But the duration of split fragments varied markedly. Based on the recommendations of the developers of the wav2vec2 model [3], short (less than 4.5 s) and long (longer than 30 s) audio files were filtered. The summary statistics on the number of files and their duration for each subcorpus are presented in Table 1.

The annotated corpus of Tatar speech, which was used for target self-supervised pretraining and target fine-tuning steps, consists of 3 parts:

1. Tatar speech corpus "TatarCorpus" [11]: close-microphone recordings, read speech - 99 h and 9 min, 500 speakers.
2. Subcorpus of television broadcasting: crowdsource annotation using the web-service [12] - 1 h and 33 min.
3. The Tatar part of the CommonVoice corpus [1] - 28 h and 47 min, 15 speakers.

Table 1. The characteristics of unlabeled speech corpus for the Tatar language

Subcorpus	Initial duration	After splitting	After filtering
Audiobooks	114 h	105 h	58 h
Television broadcasting	733 h	472 h	202 h
Radio stations' recordings	215 h	146 h	29 h
YouTube clips	87 h	81 h	39 h
Total	1 151 h	804 h	328 h

To construct a test subcorpus we chose recordings of 10 random speakers (5 male, 5 female) from the "TatarCorpus" (1 h and 37 min); for the Common Voice part, we used the original division into training and test samples, proposed by the creators of the corpus (3 h and 33 min); for the subcorpus of TV broadcasts we don't have speaker-level annotation, so the selection of 110 test fragments was carried out randomly throughout the corpus (5 min). In total it gave us 5 h 15 min test subcorpus.

As a language model for the speech recognition system, a 4-gram statistical model was built using the KenLM tool [8]. The total amount of training data was 8,760,330 sentences containing 116 million words. We downloaded and processed Tatar texts from the Internet (archives of leading news agencies, newspapers, magazines, websites of state institutions and departments, forums) and used some parts of the Tatar national corpus "Tugan Tel" [15].

4 Experiments

In total, we trained 8 different models. Taking into account the existence and type of the base model and self-supervised training steps used we will name our models in None, Base, XLSR_[SourceSS]_[TargetSS]_TargetFT format. The experiments were carried out on the fairseq platform [3]. Pretraining was carried out on 8 V100 32 GB video cards.

The recognition quality values were calculated separately for all test subcorpora. Word error rates (WER) for all built systems are presented in Table 2.

The best recognition quality on the test corpus achieved by the Base_SourceSS _TargetSS_TargetFT model: 5.67 WER even though using XLSR as the base model looked promising because of the amount of training data (56k h) and variety of languages (53, including Tatar) used during training. However, it is worth noting that on two of three test subcorpus (CommonVoice and TV) XLSR-based models show better performance than Base ones. Better quality on these subcorpora can be partially explained by the fact that CommonVoice data and Babel (telephone conversational speech) were included in the XLSR training corpus, therefore the model learned essential features right from the initial stage of training.

Table 2. Recognition quality of all trained models, WER

Model	CommonVoice	TatarCorpus	TV	Overall
None_SourceSS_TargetFT	9.54	6.98	31.42	9.30
None_SourceSS_TargetSS_TargetFT	8.17	5.99	30.78	8.04
Base_TargetFT	7.55	6.35	30.80	7.58
Base_SourceSS_TargetFT	5.80	5.08	25.08	5.98
Base_SourceSS_TargetSS_TargetFT	5.57	**4.65**	26.00	**5.67**
XLSR_TargetFT	5.73	6,52	30,03	6.39
XLSR_SourceSS_TargetFT	6.49	6.47	22.76	6.80
XLSR_SourceSS_TargetSS_TargetFT	**5.37**	5.62	**22.60**	5.77
Previous best published results	26.76 [2]	12.89 [13]	–	–

The previous best value showed by the "canonical" ASR system, built on separate acoustic models, a pronunciation model, and a language model, on the "TatarCorpus" test dataset is equal to 12.89 WER [13]. The best model proposed in this work on the same test subcorpus showed a value of 4.65 WER (Base_SourceSS_TargetSS_TargetFT). The WER values showed by the system [2] were taken as the base values for comparing the quality on the CommonVoice test dataset. The best value presented there is 26.76 WER, while our proposed system showed a value of 5.37 WER (XLSR_SourceSS_TargetSS_TargetFT).

Much higher error rates for TV test subcorpus can be explained by the complexity of spontaneous speech and partially by the fact that annotations were collected through crowdsourcing and contain mistakes. Some analysis of test TV audio fragments showed that there are several aspects that we will keep in mind in our future work:

1. Poorly distinguishable words at the beginning or end of the fragment that were not manually annotated, but were recognized by the ASR system. For instance, reference phrase 'isemendage', hypothesis 'manova isemendage' where 'manova' is an ending of a surname, where the starting part of it is not audible due to background noise);
2. Short interjections, often borrowed from the Russian language. For instance, reference phrase 'nu anda hal itep beterese', hypothesis 'anda hal itep beterese', where word 'nu' is a Russian interjection meaning 'well');
3. Other inaccuracies in annotations. For example, reference phrase 'president rostem minnehanov ta', hypothesis 'president rostam min'nehanov ta' with difference in nn' (Tatar n letter) letters; annotator made a mistake in spelling the surname in Russian and Tatar.

The second type of mistake can be influenced by the language model and not directly related to the training procedure of acoustic models. So we calculated WERs for the systems without the use of LM. The results are presented in Table 3.

With these "raw" acoustic WER values, we still see the same correlation: both SourceSS and TargetSS pretraining steps allow models to perform better on test datasets. The only two exceptions of this fact can be seen in comparison between Base_SourceSS_TargetFT and Base_SourceSS_TargetSS_TargetFT, XLSR_SourceSS_

Table 3. Recognition quality of all trained models without language model, WER

Model	CommonVoice	TatarCorpus	TV	Overall
None_SourceSS_TargetFT	16.81	14.50	36.53	16.61
None_SourceSS_TargetSS_TargetFT	14.06	13.12	35.58	14.22
Base_TargetFT	13.50	13.05	38.54	13.75
Base_SourceSS_TargetFT	8.83	10.08	28.17	9.47
Base_SourceSS_TargetSS_TargetFT	8.15	**9.13**	27.71	8.70
XLSR_TargetFT	11.76	12.35	32.97	12.31
XLSR_SourceSS_TargetFT	9.57	10.97	**22.91**	10.16
XLSR_SourceSS_TargetSS_TargetFT	**7.94**	9.57	24.15	**8.63**

TargetFT and XLSR_SourceSS_TargetSS_TargetFT for TV test subcorpus. For these two cases, the additional TargetSS step leads to an increase of WER for 3% and 5%, respectively. The increase in the quality of speech recognition for each type of model is presented in Table 4.

Table 4. Influence of self-supervised pre-training steps on recognition quality, % WER

Base model	SourceSS	TargetSS	Both SourceSS and TargetSS
None	N/A	−14.37%	N/A
Base	−31.16%	−8.09%	−36.73%
XLSR	−17.45%	−15.02%	−29.85%

5 Conclusion

This paper presents the results of experiments on building a Tatar speech recognition system using an iterative self-supervised pretraining procedure. We prepared 128-h annotated and 340-h unlabeled speech corpora. We propose two additional pretraining steps between the base pre-trained system and target fine-tuning. The first step that we called SourceSS uses unlabeled data from various sources (TV and radio broadcasting, YouTube clips, audiobooks) while the second TargetSS uses only an audio part from annotated target dataset. The testing of the proposed speech recognition systems confirmed good (SOTA) performance for different types of speech (read and spontaneous) and noise conditions. SourceSS step gave on average 24.3% WER improvement, TargetSS - 12.5%; both pretraining - 33.3%. These values were calculated for models that haven't used language models. As for absolute numbers, the best model in our experiments showed 5.37% WER for the Common Voice test dataset and 4.65% WER for TatarCorpus, which are 79.9% and 63.9% better than the previously published best result on these datasets.

References

1. Commonvoice (2021). https://commonvoice.mozilla.org/
2. Commonvoice tatar benchmark (2021). https://paperswithcode.com/sota/speech-recognition-on-common-voice-tatar
3. Fair-seq, wav2vec 2.0 pytorch example (2021). https://github.com/pytorch/fairseq/tree/master/examples/wav2vec
4. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier (2021). https://github.com/snakers4/silero-vad/
5. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Proceedings of NeurIPS (2020)
6. Baevski, A., Auli, M., Mohamed, A.: Effectiveness of self-supervised pre-training for speech recognition. CoRR abs/1911.03912 (2019). http://arxiv.org/abs/1911.03912
7. Graves, A., Fernandez, S., Gomez, G.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA (2006)
8. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197. Association for Computational Linguistics, Edinburgh, July 2011. https://www.aclweb.org/anthology/W11-2123
9. Kahn, J., et al.: Libri-light: a benchmark for ASR with limited or no supervision. CoRR abs/1912.07875 (2019). http://arxiv.org/abs/1912.07875
10. Kawakami, K., Wang, L., Dyer, C., Blunsom, P., van den Oord, A.: Learning robust and multilingual speech representations (2020)
11. 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, Russia, pp. 475–484 (2015)
12. Khusainov, A.: Instrument dlya rasrpredelennogo sozdaniya annotirovannyh korpusov. In: Proceedings of the 8th International Conference on Turkic Languages Processin, Ufa, Russia (2020)
13. Khusainov, A.: Recent results in speech recognition for the Tatar language. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 183–191. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_21
14. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: unsupervised pre-training for speech recognition. CoRR abs/1904.05862 (2019). http://arxiv.org/abs/1904.05862
15. Suleymanov, D., Khakimov, B., Gilmullin, R.: Korpus tatarskogo yazyka: konceptualnye i lingvisticheskiy aspekty. Vestnik TGGPU, pp. 211–216 (2011)