Exploring Capabilities of Monolingual Audio Transformers using Large Datasets in Automatic Speech Recognition of Czech

Jan Lehečka, Jan Švec, Aleš Pražák, Josef V. Psutka

Department of Cybernetics, University of West Bohemia Pilsen, Czech Republic

{lehecka,jlehecka,honzas,aprazak,psutka_j}@kky.zcu.cz

Abstract

In this paper, we present our progress in pretraining Czech monolingual audio transformers from a large dataset containing more than 80 thousand hours of unlabeled speech, and subsequently fine-tuning the model on automatic speech recognition tasks using a combination of in-domain data and almost 6 thousand hours of out-of-domain transcribed speech. We are presenting a large palette of experiments with various fine-tuning setups evaluated on two public datasets (CommonVoice and VoxPopuli) and one extremely challenging dataset from the MALACH project. Our results show that monolingual Wav2Vec 2.0 models are robust ASR systems, which can take advantage of large labeled and unlabeled datasets and successfully compete with state-of-the-art LVCSR systems. Moreover, Wav2Vec models proved to be good zero-shot learners when no training data are available for the target ASR task.

Index Terms: speech recognition, audio transformers, Wav2Vec

1. Introduction

Self-supervised models recently became a very popular alternative to large vocabulary continuous speech recognition (LVCSR) systems in automatic speech recognition (ASR) tasks. They can learn contextualized speech representations from large-scale unlabeled audio datasets (pretraining self-supervised phase), and consequently employ the knowledge in the ASR training from labeled data (fine-tuning supervised phase). One of the most studied self-supervised ASR model architectures is Wav2Vec 2.0 [1]. It is a deep neural network pretrained to reconstruct the input signal. The input raw audio signal is processed by a multi-layer convolutional neural network into a sequence of latent-speech representations which are fed into a multi-layer Transformer [2]. The output of the Transformer is a sequence of frame-level contextualized speech representations which are then processed by the connectionist temporal classification (CTC) layer [3, 4] decoding the most probable sequence of graphemes.

In contrast with standard LVCSR systems, these end-to-end approaches alleviate the need for word pronunciation modeling and do not require any alignment of data.

The main focus of this paper is to explore the capabilities of monolingual Wav2Vec-based models, their ability for zero-shot transfer learning, performance dependency on the training metaparameters, scaling for large ASR datasets and additional language models (LMs). We also compare the end-to-end models with state-of-the-art hybrid DNN-HMM LVCSR systems.

2. Related work

Monolingual Wav2Vec models for languages other than English are very rare [5]. For languages like Czech, there are none. However, there are many multilingual pretrained models of sizes from large [6] to extremely large [7]. These models include also Czech. The common practice is to adopt a multilingual pretrained model and fine-tune it on the labeled ASR data from the target language. Since we had access to large datasets and were not satisfied with results from multilingual models, we decided to pretrain our own monolingual model from scratch and release it to the public. We are not aware of any similar model for Czech mentioned in the literature.

3. Pretraining

Self-supervised audio transformers are known to scale well with the size of pretraining data, even with extremely huge datasets [7]. Hence, we tried to gather as much public and in-house unlabeled audio data as possible. Together, we were able to collect more than 80 thousand hours of Czech speech. We are not aware of any similar collection of Czech speech data at this scale mentioned in the literature so far. The collection includes recordings from radio (22k hours), VoxPopuli dataset [8] (18.7k hours), TV shows (15k hours), shadow speakers (12k hours), sports (5k hours), telephone data (2k hours), and a smaller amount of data from several other domains. We used also raw audio files from all speech recognition datasets (see Sec. 4).

Since the feature extraction of the input signal is limited by the memory of GPUs in use, we sliced all records not to exceed 30 s, which we found to be a reasonable input size for batching.

We used the same pretraining steps and hyperparameter values as for the base Wav2Vec 2.0 model in [1]. We pretrained the model for 400 thousand steps with a batch size of about 1.6 hours, corresponding to more than 11 epochs over the dataset. We released our pretrained model under the nickname CITRUS (abbreviation for Czech language TRansformer from Unlabeled Speech) for public non-commercial use.

4. Fine-tuning

We prepared all training and development ASR data consistently for all datasets. We sliced long training audio signals on speech pauses not to exceed the length of 30 s, longer utterances were discarded. We removed non-speech events and punctuation from the transcripts and mapped texts into lowercase. We carefully analyzed all transcripts and fixed any data-specific deviation. If not stated otherwise, we fine-tuned all models with the same setting as the base model in [1] using Fairseq tool.
4.1. Speech recognition datasets

We were experimenting with three Czech speech recognition datasets. Basic statistics are shown in Tab. 1.

Table 1: Speech recognition datasets. We are showing the number of hours, the number of words in transcripts (in thousands), and the average length of records (in seconds).

|                  | CommonVoice | VoxPopuli | MALACH |
|------------------|-------------|-----------|--------|
|                  | train dev   | train dev | train dev |
| # hours          | 32.2 8.1    | 8.1 52.3  | 3.0 3.1 |
| # words          | 183 46      | 45 404    | 23 23   |
| avg-len          | 4.1 4.6     | 4.5 10.2  | 10.0 9.9|

The CommonVoice dataset is a Czech portion of crowd-sourced project Mozilla Common Voice [9]. We used corpus version 7.0 containing 49 hours of validated speech. We decided to keep also sentences reported as offensive and difficult pronunciation in our training data. All other reported sentences (e.g., incomplete transcription, different language etc.) were ignored.

The VoxPopuli dataset [8] is a large-scale multilingual speech corpus collected from 2009-2020 European Parliament event recordings. The Czech portion contains 18.7 thousand unlabeled hours and 62 hours with transcription. We ignored all train/dev records without the raw transcription, decreasing the amount of transcribed data to 58.4 hours.

The MALACH data⁴ is a subset of the USC Shoah Foundation Visual History Archive of digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers, and witnesses of the Nazi Holocaust. We used the Czech portion [10] which is known to be a very difficult and challenging speech recognition task due to the strong emotional and heavily accented speech of Holocaust survivors [11, 12].

To scale the fine-tuning up, we used also a large additional in-house dataset, denoted as Extra, containing almost 6 thousand hours of transcribed Czech speech from various domains. The dataset includes radio shows records (3.9 thousand hours), TV sports recordings (645 hours), telephone recordings (440 hours), and several other domains. The transcripts contain 43 million words.

5

5. Decoding

We studied two different decoding setups: (1) connectionist temporal classification (CTC) [3], which is the training loss we used during fine-tuning of the models, and (2) CTC beam search decoder with an LM. CTC is an alignment-free method for grouping audio frames belonging to the same output symbol in order to convert a sequence of audio frames into a much shorter sequence of characters. Thus, Wav2Vec with the CTC decoder usually improves the speech recognition accuracy by bringing useful language information into the decoding process and penalizing improbable outputs. On the other hand, it is a step back from the idea of an end-to-end recognizer introducing once again problems known from LVCSR systems, such as out-of-vocabulary words and the low-relevance or low-quality LMs.

For our experiments, we prepared 3 different word-based n-gram LMs. (1) LM-ASRSpec is a model trained specifically for each speech recognition dataset only from in-domain training transcripts. (2) LM-ASRAII was trained from transcripts from all speech recognition training data described in Sec. 4.1, including the Extra dataset. It is a general domain-independent LM trained from ASR transcripts with 44 million words. (3) LM-C5 is an LM trained from Czech Colossal Clean Crawled Corpus (C5) [13] which is a huge collection of cleaned and deduplicated web pages from Common Crawl project⁵. Since this text corpus contains almost 13 billion words (93 GB of cleaned text), we pruned all unigrams with counts lower than 10 and higher-order n-grams with counts lower than 100.

We used implementation from Transformers [14] for CTC decoding and pyctcdecode⁶ decoder for CTC beam search decoder with n-gram LM. To train LMs, we used KenLM [15] and to work with models of practical sizes, we limited the maximum order of models to 4-grams. We trained all LMs in lowercase. The sizes of LM vocabularies were between 17 and 22 thousand (LM-ASRSpec), 329 thousand (LM-ASRAII), and 4.8 million words (LM-C5).

6. Experiments

We evaluated speech recognition systems on test splits of individual datasets and compared word error rates (WER). During the evaluation, we mapped reference texts into lowercase and ignored all punctuation.

6.1. CTC beam search decoder with LM

To get a picture of how the recognition is improved when adding LM into the decoder, we prepared data-specific Wav2Vec models (denoted as W2V-ASRSpec) by fine-tuning the pretrained model only on in-domain (single-dataset) training data. Then, we evaluated the models with different LMs in the decoder. From results tabulated in Tab. 2, we can see that adding small data-specific LM improves the recognition, adding domain-independent LM from all ASR transcripts further improves the recognition, and switching to large-scale LM trained from Common Crawl could be also beneficial, especially for domains with specific words mentioned several times somewhere on the Internet (e.g., names of politicians for VoxPopuli or geographical names for MALACH).

Table 2: WER [%] for CTC beam search decoder with different LMs evaluated on three Czech datasets.

|                  | CommonVoice | VoxPopuli | MALACH |
|------------------|-------------|-----------|--------|
| W2V-ASRSpec (no LM) | 7.29 | 11.28 | 18.93 |
| + LM-ASRSpec     | 6.12 | 10.51 | 18.45 |
| + LM-ASRAII      | 5.34 | 9.78  | 16.67 |
| + LM-C5          | 5.45 | 9.62  | 15.31 |

In the following sections, we experimented with all presented LMs but observed very similar trends, so we decided to report only results with LM-C5 since it either improved the model by correctly recognizing difficult domain-specific words.

⁴https://malach.umiacs.umd.edu
⁵https://commoncrawl.org
⁶https://github.com/kensho-technologies/pyctcdecode
or – in the case of CommonVoice – insignificantly deteriorated the results when compared to LM-ASRAll.

### 6.2. Zero-shot transfer learning

In the next experiment, we investigated the Wav2Vec’s zero-shot transfer-learning ability between datasets, i.e. how well the model can transfer knowledge from one or more domains to an unobserved target domain with zero additional training.

First of all, we cross-evaluated models fine-tuned on a single domain. These models were trained only from a single dataset. Then, we tried the opposite setting, i.e. to fine-tune with all training datasets except for the one the model was evaluated on. Results with these out-of-domain models (denoted as W2V-ood) correspond to a scenario where the model is fine-tuned from a large-scale ASR dataset (6 thousand hours) and subsequently used to recognize speech from an unobserved domain with zero additional training. The zero-shot transfer results are shown in Tab. 3.

**Table 3: WER [%] for zero-shot transfer learning evaluated on three Czech datasets. We are reporting results of models fine-tuned from single-domain training data (the first 3 models, each with and without LM, and all training data except for the in-domain data (W2V-ood)).**

| Dataset            | CommonVoice | VoxPopuli | MALACH |
|--------------------|-------------|-----------|---------|
| W2V-ood            | - 15.62     | 26.57     |         |
| + LM-C5            | - 13.81     | 22.33     |         |
| W2V-VoxPopuli      | 29.94       | - 33.37   | 28.04   |
| + LM-C5            | 22.57       | -         | 28.04   |
| W2V-MALACH         | 19.21       | 16.89     |         |
| + LM-C5            | 11.63       | 13.36     |         |
| W2V-ood + ASRSpec  | 11.57       | 13.12     | 16.86   |
| + LM-C5            | 6.40        | 11.22     | 13.55   |

When comparing W2V-ood model with in-domain models (Tab. 2) for CommonVoice and VoxPopuli datasets, in-domain models are significantly better (with relative improvement about 15%), which is the expected result. However, we observed a strange WER improvement when evaluating the W2V-ood model on the MALACH dataset. After some investigation, we found out, that it was caused by a mix of formal and colloquial Czech in transcripts. Formal Czech is the grammatically correct written form, while colloquial Czech is a spoken (common) form containing for example different suffixes (e.g. “mlady” vs. “mladej”) or shortened word forms (e.g. “jsem” vs. “sem”). More details can be found in [16]. While in the training transcripts, the majority of the transcribed text is in colloquial Czech (i.e. transcribed exactly as spoken), the opposite is true for the test transcripts. So, when the in-domain model was fine-tuned from the training MALACH data, it was forced to decode rather a colloquial Czech, which, however, conflicted with the test transcripts, causing a lot of word-substitution errors. For this reason, the out-of-domain model can perform significantly better on the MALACH dataset than the in-domain model. We further explored this phenomenon in [17].

**6.3. General vs. in-domain models**

In the next experiment, we compared in-domain models (W2V-ASRSpec) with a general model fine-tuned from all available training data (i.e. both in-domain and out-of-domain) at once. Moreover, we were experimenting with a 2-phase fine-tuning procedure, in which the pretrained model is first fine-tuned from multi-domain data (same as the general model) and then fine-tuned again using only the in-domain data. In this approach, the model can benefit from large-scale out-of-domain data and, at the same time, the in-domain information is accentuated.

**Table 4: WER [%] for differently fine-tuned models evaluated on three Czech datasets. We report results from models fine-tuned on all ASR data, only in-domain data, and a combination of both within a 2-phase fine-tuning procedure. We report all results without an LM and when decoded with LM-C5.**

| Model                | CommonVoice | VoxPopuli | MALACH |
|----------------------|-------------|-----------|---------|
| W2V-general          | 11.17       | 12.37     | 19.71   |
| + LM-C5              | 6.10        | 10.13     | 13.64   |
| W2V-ASRSpec          | 7.29        | 11.28     | 18.93   |
| + LM-C5              | 5.45        | 9.62      | 15.31   |
| W2V-general + ASRSpec| 6.52        | 10.07     | 18.63   |
| + LM-C5              | 4.74        | 8.80      | 15.19   |

Results are shown in Tab. 4. Our suggested 2-phase fine-tuning performed significantly better than both the general and in-domain models. The only exception is the MALACH dataset, where in-domain models with LM again learned to decode colloquial Czech contrasting with test transcripts.

### 6.4. Scaling up a batch size and updates

So far, all models were fine-tuned with the default setting [1], i.e. 80 thousand training steps with a batch size of about 27 minutes of audio corresponding to more than 5 epochs over all ASR data. Since we had large-scale ASR data, the natural question we asked was: Would it be beneficial to train the general model for a longer time? And with a larger batch?

To answer these questions, we fine-tuned the general model for a higher number of updates and/or with larger batch sizes, effectively increasing the number of training epochs over data. We fine-tuned with learning rate $8 \times 10^{-5}$ which we found to be better for large-scale datasets. Our results in Tab. 5 show that Wav2Vec speech recognizers scale very well with increasing the number of fine-tuning epochs over large datasets. For example, when we trained the model for 8x more epochs using a batch size of 108 minutes (4x larger than the default model) and 160 thousand update steps (2x more than default), the total error rate of end-to-end ASR decreased from 15.47% to 13.33%. Moreover, additional in-domain fine-tuning within the 2-phase procedure as described in Sec. 6.3 (model denoted as 40 epochs + ASRSpec) further improved the recognition. In the second phase of fine-tuning, we used the default settings since the underlying in-domain data were again small.

Also, our results show that models trained for the same number of epochs performed about the same no matter we multiplied the batch size or the number of updates. Results with the MALACH dataset with LM were again affected by the inconsistency of train and test texts leading sometimes to models preferring colloquial Czech over formal Czech and thus causing a large number of recognition errors.

### 6.5. Comparison with other models

Wav2Vec models are a very promising research area and we wanted to compare them also with existing ASR systems. As
Table 5: WER [%] for scaling the fine-tuning epochs up. We show how the model performs when increasing the batch size (BS) and/or the number of updates (UP) by multiplying the default values. For each model, we show WER without LM and with LM-C5 for individual datasets, and total WER computed by aggregating numbers of words and errors over all 3 datasets.

|                | CommonVoice | VoxPopuli | MALACH | TOTAL |
|----------------|-------------|-----------|--------|-------|
|                | no LM       | LM-C5     | no LM  | LM-C5 | no LM  | LM-C5  |
| 5 epochs (default) | 11.17       | 6.10      | 12.37  | 10.13 | 19.71  | 13.64  | 15.47  | 10.43 |
| 10 epochs (2xUP) | 9.30        | 5.04      | 11.00  | 9.42  | 18.62  | 13.33  | 14.07  | 9.79  |
| 10 epochs (2xBS) | 9.23        | 4.82      | 11.18  | 9.52  | 19.05  | 13.60  | 14.28  | 9.86  |
| 20 epochs (2xBS, 2xUP) | 8.49    | 4.59      | 10.75  | 9.30  | 17.99  | 12.98  | 13.45  | 9.44  |
| 20 epochs (4xBS) | 8.39        | 4.62      | 10.82  | 9.31  | 18.86  | 13.89  | 13.84  | 9.89  |
| 40 epochs (4xBS, 2xUP) | 7.68    | 4.29      | 10.23  | 8.81  | 18.52  | 13.73  | 13.33  | 9.61  |
| + ASRSpec       | 5.41        | 3.80      | 10.07  | 8.80  | 17.65  | 14.51  | 12.10  | 9.82  |

Results from our paper showed that the performance of the Wav2Vec-based ASR system can be significantly improved by:

- longer fine-tuning over the large-scale dataset,
- adding LM into the beam search CTC decoder; we found the Common Crawl project to be a useful corpus for large-scale language modeling,
- 2-phase fine-tuning procedure - multi-domain fine-tuning in the first phase followed by a second stage of fine-tuning using only data from the target domain.

Our best results with LM-C5 in Tab. 5 are – to our best knowledge – new state-of-the-art results in all three datasets.

8. Conclusions

In this paper, we pretrained Czech monolingual audio transformers from a large dataset containing more than 80 thousand hours of unlabeled speech and released the model to the public. We presented a large palette of experiments with various fine-tuning setups with large-scale datasets (up to 6 thousand hours) evaluated on two public datasets (CommonVoice and VoxPopuli) and one extremely challenging dataset from the MALACH project. Our results showed that monolingual Wav2Vec models are robust ASR systems, which can take advantage of large labeled and unlabeled datasets and successfully compete with state-of-the-art LVCSR systems. Moreover, our general transformer-based ASR systems proved to be good zero-shot learners when no training data are available for the target domain. Our models scored new state-of-the-art results on all three tested datasets by a large margin.

9. Acknowledgements

This research was supported by the ITI project of the Ministry of Education of the Czech Republic CZ.02.1.01/0.0/0.0/17 048/0007267 InteCom and by the Czech Science Foundation (GA CR), project No. GA22-27800S.
10. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2Vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12 449–12 460, 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[4] A. Baevski and A. rahman Mohamed, “Effectiveness of self-supervised pre-training for ASR,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7694–7698, 2020.

[5] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Midhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelle, T. Parcollet et al., ”Task agnostic and task specific self-supervised learning from speech with lebenchmark,” in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[6] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlicek, Eds. ISCA, 2021, pp. 2426–2430. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in Proceedings of the 23rd Conference on Neural Information Processing Systems, 2020, pp. 12 449–12 460. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329

[8] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, “VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[9] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[10] J. Psutka, V. Radová, P. Irung, J. Matoušek, and L. Müller, “USC-SFI MALACH interviews and transcripts Czech LCD2014S04,” 2014. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2014S04

[11] J. V. Psutka, J. Švéc, and A. Pražák, “CNN-TDNN-based architecture for speech recognition using grapheme models in bilingual Czech-Slovak task,” in Text, Speech, and Dialogue, K. Ekstein, F. Pártl, and M. Konopík, Eds. Cham: Springer International Publishing, 2021, pp. 523–533.

[12] J. V. Psutka, A. Pražák, and J. Vaněk, “Recognition of heavily accented and emotional speech of English and Czech Holocaust survivors using different DNN architectures,” in Speech and Computer, A. Karpov and R. Potapova, Eds. Cham: Springer International Publishing, 2021, pp. 553–564.

[13] J. Lehečka and J. Švéc, “Comparison of Czech transformers on text classification tasks,” in Statistical Language and Speech Processing, L. Espinosa-Anke, C. Martín-Vide, and I. Spasić, Eds. Cham: Springer International Publishing, 2021, pp. 27–37.