Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition

Geoffroy Vanderreydt, François Remy, Kris Demuynck

IDLab Ugent-imec

geoffroy.vanderreydt@ugent.be, francois.remy@ugent.be, kris.demuynck@ugent.be

Abstract

In this article, we propose a simple yet effective approach to train an end-to-end speech recognition system on languages with limited resources by leveraging a large pre-trained wav2vec2.0 model fine-tuned on a multi-lingual speech translation task. We show that the weights of this model form an excellent initialization for Connectionist Temporal Classification (CTC) speech recognition, a different but closely related task. We explore the benefits of this initialization for various languages, both in-domain and out-of-domain for the speech translation task. Our experiments on the CommonVoice dataset confirm that our approach performs significantly better in-domain, and is often better out-of-domain too. This method is particularly relevant for Automatic Speech Recognition (ASR) with limited data and/or compute budget during training.

Index Terms: Speech Recognition, ASR, CTC, XLS-R.

1. Introduction

Speech recognition of low-resource languages has always been a challenge. But recently, it was shown by Conneau et al. [1] that leveraging cross-lingual transfer from high-resource languages (by building better audio speech representations) enables to produce better speech recognition models for low-resource languages. Their work (XLSR-53) was further extended by Babu et al. [2] to produce XLS-R, a large-scale model for cross-lingual speech representation learning based on wav2vec2.0 [3]. Compared to the original XLSR-53, the new XLS-R is pre-trained with more diverse data: from 58 to 128 languages. For speech recognition, XLS-R improves over the best known prior work on 4 popular tasks, lowering error rate by 14 to 34% relatively. This builds on a longer tradition of attempting to transfer knowledge from high-resource languages to low-resource ones [4, 5, 6].

Unsupervised speech representations also enabled progress in the Speech Translation task [7], where spoken language is translated directly into text in another language, without using an intermediate transcription in the original language [8]. In that setting, a sequence to sequence model is initialized with a pre-trained multi-lingual speech encoder (XLS-R [2]) and a pre-trained multi-lingual text decoder (mBART [9]), which are both fine-tuned end-to-end, and achieve state-of-the-art results.

In this paper, we investigate our intuition that fine-tuning a XLS-R model to do multi-lingual Speech Translation produces even better speech representations, and we thus probed how effectively these learned representations transfer to the task of mono-lingual Automatic Speech Recognition (ASR).

Between January 24th to February 7th 2022, HuggingFace [10] organized the Robust Speech Recognition challenge [11], where participants were invited to fine-tune XLS-R models to build Connectionist Temporal Classification (CTC) [12] speech recognition systems. Participants were provided with one V100s GPU with 32Gb of memory. Within the time and hardware constraints, we investigated how to best develop a Dutch speech recognition model, and won the competition with a significant margin. In this article, we further explore the theoretical aspect of our Speech-Translation-based CTC training strategy and its benefits across multiple languages. We expect this strategy to become the baseline for future competitions organized in similar settings.

This paper is organized as follows. In section 2 we explain the intuition behind our proposed approach and why it could help in a low resource setup. Next we present the experimental setup in which we show our method's superiority over the standard XLS-R. Section 4 is devoted to the experiments and results. Finally, conclusions are presented in Section 5.

2. Methodology

Transcribing speech from a set of languages into text from another language is a challenging task requiring speech understanding at least as good as for mono-lingual ASR [8]. Therefore, intuitively, we expect speech translation models to yield even better speech representations. The goal of this work is three-fold; we explore how much a mono-lingual ASR system can benefit from these potentially superior speech representations, we investigate if they generalize well to languages unseen by the speech translation system, and we look into the feasibility of this approach given the computational resource constraints imposed by HuggingFace's challenge.

These speech representations, if they prove useful, should enable training a CTC decoder with very few labelled data, even for languages which were not covered by the Speech Translation task. To verify our hypothesis, we compare the word error rates (WER) of CTC decoders initialized with XLS-R models which either have or have not undergone further training in the speech translation task.

For the remaining of this article, we will refer to the fine-tuned encoder of the speech translation system [7] as *XLS-R-st*, in opposition to *XLS-R base*, the model which only underwent wav2vec2.0 masked-language-modelling pre-training [2].

To differentiate between direct transfer of knowledge and cross-lingual generalization, we repeated this experiment for a varied set of languages, some of which are covered by the speech translation task, and some of which are not. This enables to quantify how much of the learned representations transfer to different languages. Verifying their generalization power on several languages is even more necessary given that speech translation requires more specialized data than speech recognition, hence models fine-tuned on speech translation will continue training on a smaller subset of languages than speech recognition models, due to high resource requirements. While these languages can be selected to be as representative as possible, assessing whether that is sufficient to compensate for the

This is the accepted version of : Vanderreydt, G., REMY, F., Demuynck, K. (2022) Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition. Proc. Interspeech 2022, 3053-3057, doi: 10.21437/Interspeech.2022-10744

benefits of being trained on a more exhaustive set of languages remains a relevant question.

To compare the models potential, we focused on building CTC speech recognizers, as those have been well-studied in the context of wav2vec2.0 models. We chose to grade the models mainly based on the WER rather than the loss because it is the most challenging metric for speech recognition [13]. While it is not directly comparable between languages (some languages have inflections while others do not, and some languages have more recent writing systems with fewer phoneme-to-grapheme oddities), the WER can be compared for each language individually, given the models were trained and evaluated on the same dataset splits. Our main focus in this experiment is the relative improvement of the *XLS-R-st* over the *XLS-R base* models.

3. Experimental setup

3.1. Datasets

As finetuning data we consider the crowd-sourced Mozilla CommonVoice [14] corpus (release 8). We distinguish two main language categories in our experiments: in-domain (*cat-in*) and out-of-domain (*cat-out*) languages to the speech translation (ST) model. The ST model translates speech from 22 languages to text in 16 languages (Table 1). We further subdivide the in-domain category into two subcategories depending on whether the language belongs to the target languages (*cat-in-in*) of the ST model or not (*cat-in-out*). Given that the list of 22 in-domain languages is diverse at first sight but fails to include languages from every linguistic branch, we decided to investigate whether our approach performs differently on European out-of-domain languages (*cat-out-eur*) compared to languages which are geographically and culturally unrelated to all in-domain languages. We chose two African languages for this (*cat-out-afr*). For all languages, we train the end-to-end models on the union of the CommonVoice 8 train and validation sets. Table 2 gives the number of hours and number of speakers per language and split.

Table 1: *Languages of the speech translation model [7].*

Domain	Languages
Input & target	English, Catalan, Chinese, Persian, Estonian, Mongolian, Turkish, Arabic, Swedish, Latvian, Slovenian, Tamil, Japanese, Indonesian, Welsh
Input only	Dutch, Italian, Russian, Portuguese, French, Spanish

3.2. The models

XLS-R base is available in 3 flavours with different number of parameters; a small (300M), a medium (1B) and a large (2B) version. From the XLS-R paper [2], there is a large performance gap between the small and medium variants and a more limited one between medium and large. Therefore, in this work, we solely consider the medium and large versions of *XLS-R base*. *XLS-R-st* is in essence a copy of the largest variant of *XLS-R base* fine-tuned for speech translation. It therefore contains about 2B trainable parameters.

However, with the given hardware constraints (32Gb V100s GPU), it is not feasible to train such large 2B models entirely. Indeed, none of the HuggingFace competition participants were able to train these large models on the provided hardware. We circumvent that problem by freezing the parameters of half of the transformer layers.

The whole idea of cross-lingual pre-training is that different languages can share similar pronunciations and therefore the speech representation learning can benefit from multiple languages at the same time. Fine-tuning with CTC is generally intended to map those speech representations to a monolingual token vocabulary, leaving the robust feature encoder untouched. Intuitively, the transformer layers closer to the feature encoder work at a lower abstraction level (speech representations, phonemes) than those towards the output end (just before the decoder if there is one). In order to reduce the amount of trainable parameters, we extended the freezing to include the first half of the transformer layers of 2B models, i.e. those layers closest from the feature encoder. This means that the 1B and 2B variants of the models can be fine-tuned based on a similar amount of trainable parameters.

This simple trick worked for *XLS-R-st*. However, the large (2B) *XLS-R base* is not able to carry out the ASR task properly with half of its transformer layers frozen (Section 4.2). Hence, with the given hardware limitations, only the medium *XLS-R base* 1B model is a feasible option. In this work, we show the superiority of *XLS-R-st* as initialization over *XLS-R base* under the data and computational budget of the competition.

3.3. Fine-tuning

We fine-tuned the models for each language with about 100h training data, eventually re-sampling multiple epochs if the dataset was smaller than that. We used a batch size of 32 sentences of approximately 5-6 seconds long and a linearly-decreasing learning rate with a 5e-5 maximum and a warm-up stage of 200 batches. We froze the parameters of the feature encoder and eventually those of the first half (24 layers) of transformer layers in order to fit the training in a 32Gb V100s GPU. We used Huggingface's *run_speech_recognition_ctc.py* script for training and CTC decoding, available on Huggingface's Github page [15].

We ran a few prototype experiments on French, Dutch and German to help us pick good hyper-parameters for the experiments reported in this article. It should be noted that the results were similar in each of the prototype experiments, only the final WER differed. We did not cherry-pick favorable hyper-parameters for our proposed approach, we picked those who performed the best overall across these three languages.

4. Experiments and Results

4.1. *XLS-R base* vs. *XLS-R-st* across languages

Figure 1 and 2 depict the evolution of the WER as a function of the amount of train data for in-domain languages (*cat-in*). We focus on the first 100h as we are interested in knowledge transfer for low-resource languages.

From Figure 1, we observe that our proposed method based on *XLS-R-st*, is consistently better than *XLS-R base* for fully in-domain languages (*cat-in-in*) over the first 100h train data. In German, *XLS-R base* starts performing decently (WER below 20%) after seeing about 50 hours of speech data. *XLS-R-st* needs only about 15 hours of speech data to achieve the same accuracy. We observe a similar trend for Catalan. As expected, fine-tuning *XLS-R* on the ST task for a limited number of languages yields better speech representations for that set of languages. Hence, *XLS-R-st* is a better ASR initialization for languages belonging to both input and target in Table 1.

Table 2: *Dataset specifications in Mozilla CommonVoice 8 for the languages used in this article.*

Category	Lang	Lang ISO	train+val hours	train+val speakers	test hours	test speakers
cat-in-in	**German**	**de**	690	9016	27	4658
	Catalan	**ca**	770	3118	27	3150
cat-in-out	**Dutch**	**nl**	47	191	13	1193
	French	**fr**	642	9182	26	4622
cat-out-eur	**Polish**	**pl**	31	597	11	2110
	Basque	**eu**	25	218	10	701
cat-out-afr	**Swahili**	**sw**	44	44	13	154
	Luganda	**lg**	111	148	21	296

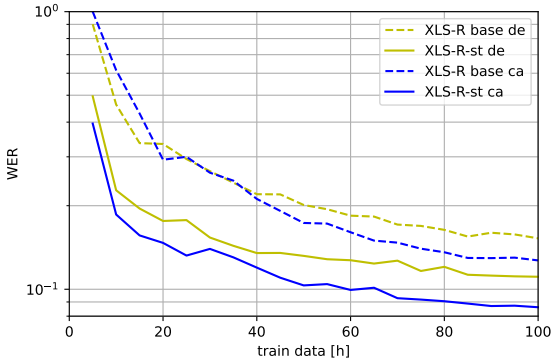

Figure 1: *WER of ASR systems initialized with and XLS-R base and XLS-R-st for German (de) and Catalan (ca):* cat-in-in.

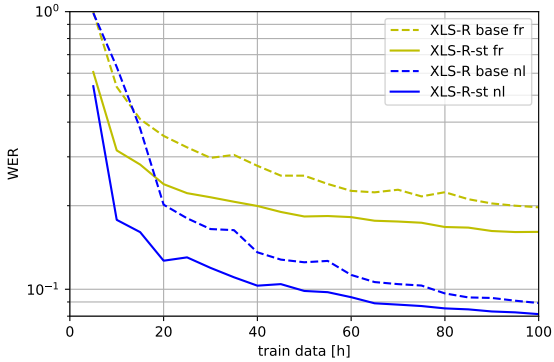

Figure 2: *WER of ASR systems initialized with and XLS-R base and XLS-R-st for Dutch (nl) and French (fr):* cat-in-out.

Figure 2 depicts similar curves for languages belonging to *cat-in-out*: Dutch and French. Here as well, the ASR system benefits from the higher quality speech representations resulting from the fine-tuned model (*XLS-R-st*). The similar performance improvement for both *cat-in-out* and *cat-in-in* languages confirms our straightforward intuition that most of the speech understanding in the ST system is done by the encoder.

We observe a significant performance difference between Dutch and French for either models after an equivalent amount of seen speech data. Our best guess for why, is that the CTC decoder should be more effective on a phonetically consistent language like Dutch. On the contrary, word pronunciations in French highly deviate from how they are written. The tokenizer needs to know the language really well in order to output characters that have not been pronounced but should be present to be orthographically correct.

So far, we showed that fine-tuning XLS-R on a speech translation task yields better speech representations in the languages the model has been fine-tuned on, resulting in a higher speech recognition accuracy on those same languages. It remains to be seen whether these learned representations generalize to out-of-domain languages. Fine-tuning on a limited set of languages might lose the generalization learned from the unsupervised pre-training on a large set of languages.

Figure 3 depicts the WER curves for *cat-out-eur* languages: Polish and Basque. They show that the weights of *XLS-R-st* form an excellent initialization for CTC speech recognition on out-of-domain languages as well. With about 15h of speech data, *XLS-R-st* achieves a similar WER than the *XLS-R base* with around 55h of speech for Polish and Basque.

The ST model covers a large variety of languages in the world (Table 1), which are nevertheless dominated by European languages. Hence, we investigate the potential benefit of the approach on languages highly unrelated to those listed in Table 1. Sub-Saharan Africa is not represented in this list and thus languages from this part of the world form a good choice for this type of experiments. We opted for languages with sufficient amount of speech data. The CommonVoice Swahili and Luganda datasets fulfilled our requirements.

Figure 4 depicts the WER curves for these *cat-out-afr* languages. It shows that the proposed approach is beneficial for Luganda, though with a more limited margin than what was observed in the previous categories. On Swahili however, we note similar trends between *XLS-R base* and *XLS-R-st*. The proposed approach does not outperform the base model but does not underperform either.

Table 3 summarizes the relative WER decrease per language category when fine-tuning from *XLS-R-st* instead of *XLS-R base*. The relative improvements are given in percentage and are averaged within each category, for 25h, 50h, 75h and 100h of seen speech data. These averages confirm that our proposed method provides the most benefits when few labeled speech data is available (25h). For some languages, the WER is almost halved in such low resource conditions. The more fine-tuning data is available, the better the speech representations of the base model and the benefit of our method is reduced. Hence, our proposed task-transferred initialization approach is mainly intended for a low resource setup.

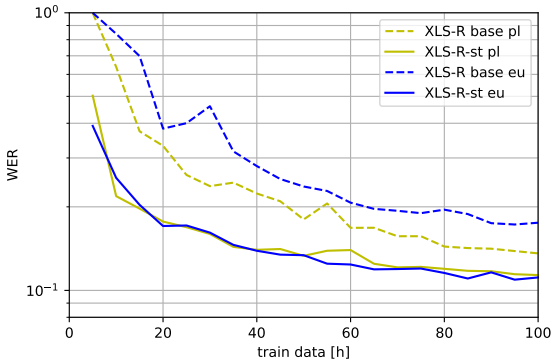

Figure 3: *WER of ASR systems initialized with and XLS-R base and XLS-R-st for Polish (pl) and Basque (eu):* cat-out-eur.

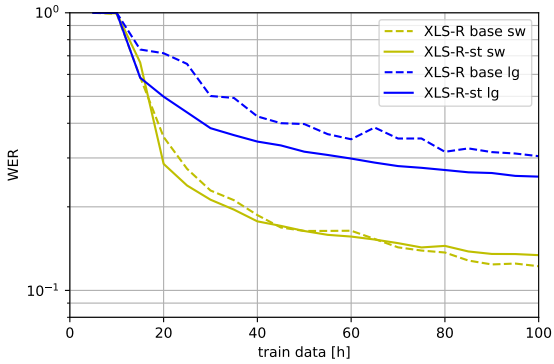

Figure 4: *WER of ASR systems initialized with and XLS-R base and XLS-R-st for Swahili (sw) and Luganda (lg):* cat-out-afr.

Table 3: *Relative Word Error Rate (WER) improvement (in %) of XLS-R-st over XLS-R base (1B) after seeing 25, 50, 75 and 100 hours of training data.*

Lang category	25h	50h	75h	100h
cat-in-in	48	37	33	30
cat-in-out	30	25	18	14
cat-out-eur	46	35	30	26
cat-out-afr	19	2	-6	-9
average	36	25	19	15

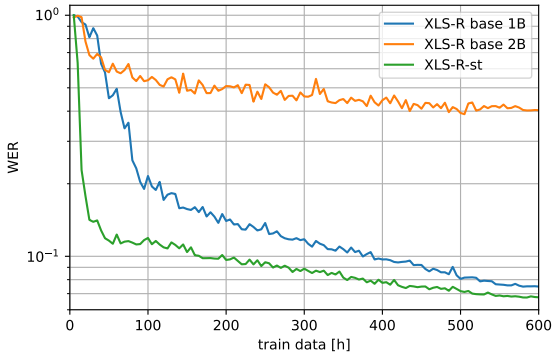

Figure 5: *WER of ASR systems intialized with XLS-R base (1B and 2B) and XLS-R-st, and fine-tuned on 600h of Catalan (ca).*

4.2. More fine-tuning data

In this section, we analyze the WER evolution of ASR systems initiated with *XLS-R base* and *XLS-R-st* when using more fine-tuning data. We increased the number of batches to 12000, corresponding to approximately 600h of speech, and the warmup to 2000 batches. We conducted the experiment on Catalan as we wanted a sufficient amount of data for that schedule (see Table 2). The rest of the parameter settings are as described in Section 3.3. Note that the difference in learning rate schedule compared to the previous experiments means the curves will not match with those in Figure 1.

Figure 5 depicts the WER evolution of *XLS-R base* (medium and large) and *XLS-R-st*. Firstly, we confirm that the ASR performance of both systems tend to converge to a similar performance with a larger fine-tuning dataset, with our proposed method remaining in the lead.

Secondly, we see that freezing half the transformer layers of *XLS-R base*'s large version does not work. With the provided hardware, this trick can only be applied to the already fine-tuned XLS-R-st model. No Robust Speech Challenge [11] participant was able to leverage that 2B model within the time and computing constraints, confirming there was no straightforward solution to make it work.

In a future work, it would be interesting to run these experiments again with 2 billion parameters unfrozen, to obtain further state of the art results. This, however, would require a compute infrastructure with more GPU memory.

5. Conclusion and final words

In this article, we proposed to leverage a large pre-trained wav2vec2.0 model fine-tuned on a multi-lingual speech translation task as initialization to train speech recognition systems for languages with limited resources. Our experiments indicate that this initialization is effective for various languages, both in-domain and out-of-domain w.r.t. the speech translation task. We believe this method is particularly relevant for ASR with limited resources. Moreover, this approach enables researchers to benefit even more from the massive efforts recently made towards creating universally applicable wav2vec2.0-based feature extractors for ASR, by unlocking the usage of speech translation models in addition to the (pre-trained) speech recognition models. From a theoretical point of view, this work also shows that, while supervised fine-tuning of unsupervised wav2vec2.0 feature extractors is effective to obtain high ASR accuracy for low-resource languages, fine-tuning them first using supervised data from multiple languages helps transfer learning even more.

6. Acknowledgements

We would like to thank HuggingFace for providing support during this competition and OVHCloud for the GPU credits. Their resources helped us build the Dutch model which won HuggingFace's Robust Speech Recognition Challenge.

7. References

[1] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[2] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[3] 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.

[4] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

[5] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at microsoft," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8604–8608.

[6] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[7] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[8] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Towards speech-to-text translation without speech recognition," 2017. [Online]. Available: https://arxiv.org/abs/1702.03856

[9] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[11] P. von Platen, "Robust speech recognition challenge," 2022. [Online]. Available: https://discuss.huggingface.co/t/open-to-the-community-robust-speech-recognition-challenge/13614

[12] 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.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[14] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[15] P. von Platen, "Connectionist temporal classification for speech recognition," 2021. [Online]. Available: https://github.com/huggingface/transformers/blob/main/examples/pytorch/speech-recognition/run_speech_recognition_ctc.py