DOI: 10.17469/O2111AISV000024

ELEONORA RODEGHER, ALESSIO BRUTTI, DANIELE FALAVIGNA

Automatic speech recognition for low-resource languages: exploring transfer learning from related languages

Automatic speech recognition systems typically require huge amounts of data, which are not always available for languages lacking digital resources. In this work we address the problem of automatic speech recognition for low-resourced languages by apply transfer learning approaches to the large-scale pre-trained model wav2vec2-xls-r. In particular, we present a sequential adaptation strategy where, using a common tokenizer, the large-scale model is firstly fine-tuned on a genealogically related well-resourced language ("pivot") and then further adapted on the target low-resourced one. Experimental results in terms of word error rate, show that this approach considerably outperforms models directly fine-tuned on the target language, in particular when very few data are available. Experiments involve Italian, Spanish, Portuguese and German as well-resourced languages and two minority languages: Galician and the variety of Romansh called Vallader.

Keywords: Automatic Speech Recognition, Transfer Learning, Large-Scale pre-trained model, low resourced languages.

1. *Introduction*

Automatic speech recognition (ASR) systems typically require huge amounts of data, which are not always available for languages lacking digital resources. Recently, new approaches have been investigated to overcome the scarcity of labelled data (Besancier, Barnard, Karpov & Schultz, 2014). One method is to adapt models pre-trained on vast quantities of data from well-resourced languages to poorly resourced ones. Differently from what commonly done in literature, i.e. training multilingual models or fine-tuning a large-scale pre-trained model on the target language, in this work we present a transfer learning strategy which fine-tunes the multilingual pre-trained model wav2vec2-xls-r (Babu, Wang, Tjandra, Lakhotia, Xu, Goyal, Singh, Platen, Saraf, Pino, Baevski, Conneau, & Auli, 2021) on different low-resourced languages using a pivot well-resourced one. In particular, we considered four well-resourced languages: Italian, Portuguese, Spanish and German; and two minority languages: Galician and Romansh Vallader.

2. *State of the art on ASR systems for low resource languages*

Recent works on ASR systems for low-resource languages are illustrated in this section. We refer the reader to (Meng, Yolwas, 2022) for a more complete survey. The work in (Docío-Fernandez & Garcia-Mateo, 2002) was one of the first investigations on the use of multiple languages to develop ASR systems for low-resourced languages. The paper proposes a bilingual ASR system for Spanish and Galician, having a scarce amount of data for Galician, training a multilingual model on a global set of phonemes. A more recent study (Yi, Wang, Cheng, Zhou, & Xu, 2021), used limited speech data from real spoken scenarios in six different languages to adapt the wav2vec2.0 model pretrained on English (Baevski, Zhou, Mohamed, & Auli, 2020) The authors evaluated different speech units and showed that using sub-words or characters led to slightly better results. More analogous to our approach, (Bermuth, Poeppel, & Reif, 2021) demonstrated the benefit of transfer learning where language similarities occur. A QuartzNet architecture, pre-trained for English is adapted to German and other languages with a similar alphabet. Similarly, in (Jacobs, Kamper, 2021) transfer learning on genealogically related languages is investigated by training multilingual word embedding models on South African low-resourced languages. The results proved that training a multilingual model using closely related languages improves the model's performance. Finally, a transliteration transfer learning approach was tested in (Khare, Mittal, Diwan, Sarawagi, Jyothi, & Bharadwaj, 2021), converting the transcriptions of a high-resource language using the graphemes of low-resource ones. The languages involved were English and six low-resourced, e.g. Hindi, Telugu, Gujarati, Bengali, Korean, and Amharic. The model wav2vec2 was pre-trained on these transliterations and fine-tuned on limited speech data of the other languages.

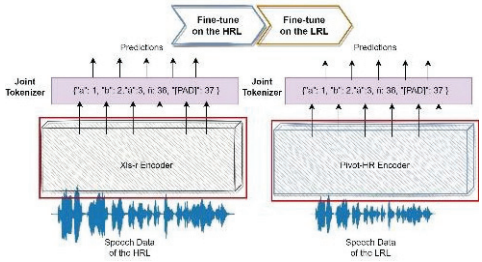
Although our approach relies on a pre-trained large-scale model as wav2vec2.0-xls-r and on similarities between languages, as many of the works mentioned above, we propose a novel two step strategy which uses the well-resource language for a first adaptation of the pre-trained model.

3. *Proposed Approach*

A common approach to facing the sparseness of data is to use transfer learning. This technique consists in reusing a model trained on one task for another similar or related task. In ASR domains, transfer learning techniques can be employed for transferring the knowledge from a model trained on speech from high- resource languages to low-resourced ones, by fine-tuning all or parts of the model on speech from the low-resourced ones (Kahre *et al.*, 2021). This permits to reuse the speech representations shared among languages when fine-tuning on the target language. The pre-trained model wav2vec2-xls-r is the large-scale multi-lingual version of wav2vec2.0. The model was pre-trained on half a million hours of speech from 128 languages, using the self-supervised learning approach. For ASR tasks, a linear layer is added on top of the pre-trained model to predict the output tokens using the connectionist temporal classification (CTC) loss (Graves, Fernandez, Gomez, & Schmidhuber, 2006).

Being able to provide representations for multilingual speech inputs wav2vec2-xls-r is a promising candidate for applying transfer learning for a specific low-resourced language. Nevertheless, when the amount of training data is extremely low, a direct fine tuning of the pre-trained multi-lingual model is not sufficiently effective. In this work we propose a 2-step approach: first the pre-trained model is fine-tuned on a language genealogically similar for which enough training material is available (we refer to this as the pivot language); then the fine-tuned model is further adapted on the data of the low-resourced target language. Figure 1 shows a graphical example of the proposed approach applied to Galician with Italian as pivot language.

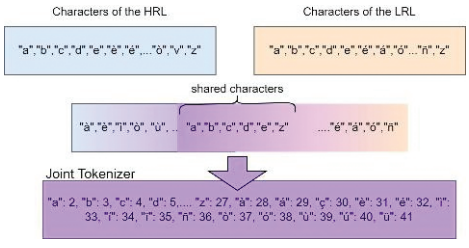
Figure 1 - *Example of the proposed approach applied to Galician with Italian as pivot language. The wav2vec2.0-xls-r model is firstly adapted on the Italian data. Then, the fined tuned model is further trained using the Galician data*



Clearly, the choice of the pivot language is fundamental as it should share acoustic and graphemic similarities with the target language.

A relevant aspect in the implementation of the proposed approach is the definition of the tokenizer to be used on top of the wav2vec2.0 encoder. Given that we use graphemes as tokens of our speech recognition system, we utilize a joint tokenizer obtained by merging the graphemes of the pivot and target languages. This implementation choice allows preserving the knowledge related to the common graphemes and use the small training set of the low-resourced language to learn the specific graphemes. Figure 2 shows an example of joint tokenizer.

Figure 2 - *Example of the creation of the joint tokenizer*



4. *Dataset*

The dataset adopted in our analysis was Version 8 of the Mozilla Common Voice by (Ardila, Branson, Davis, Kohler, Meyer, Henretty, Morais, Saunders, Tyers, & Weber, 2020). The open crowd-sourced multi-language dataset, as of August 2022, counts 96 languages. Users can contribute to enriching the corpus of a specific language by recording their voice or evaluating the transcriptions of already recorded speech data. The audio files are in MPEG-3 format with a 48kHz sampling rate (Ardila *et al.*, 2020). The recordings are primarily read sentences from Wikipedia. The data are complemented with a rich set of information regarding the amount of validated audio files, the overall duration, and the average length of an audio file. Each corpus is split into the train, test, and validation sets. The TSV file contains information about the speakers, such as their age, sex, and accent, the path to the audio files, and the transcriptions of the read sentences. Speech data and transcriptions have been evaluated and only clips marked as valid are included in the official training, development, and testing sets for each language.

In our study, for each corpus the train set was used to fine-tune the models, while we made use of the validation set to test the model in training. The test set was passed to the fine-tuned models for inferences, as these data were never passed to the model in training. As mentioned above, we extracted 4 well-resourced languages: Italian, Spanish, Portuguese and German and 2 very low-resourced languages: Galician and Romansh Vallader. More details about the data of the latter two languages are reported later.

5. *Experimental analysis*

The experiments considered different amounts of data to better assess on the impact of the pivot language. Performance was evaluated in terms of Character and Word Error Rates.

The code is available here: https://github.com/elerdg/ASR-for-low-resource-languages. We point the reader to that repository for details about the hyperparameters and the data pre-processing. As baselines, we created monolingual models for Galician, and Romansh Vallader, considering for each language a tokenizer containing all the characters plus special tokens for CTC decoding. Note that specific training strategies, as for example data augmentation, can be used to improve the performance and the generalization capabilities of the model. Since we are interested in investigating the impact of using a pivot language, we do not apply any of these methods.

5.1 Experiments on Galician

Our first experiment involves the Galician, a minority language spoken in the Spanish region of Galicia situated in the northwest of the Iberic peninsula. This language is part of the romance language family and has historical roots with Portuguese (Ramallo, 2015).

Version 8 of Mozilla Common Voice contains 10.17 validated hours of Galician, equivalent to 7618 sentences, from 486 different speakers. The average duration of an audio file is 4.808 seconds. In the train set there are 3062 files which are equal to 4 hours of speech, in the test and the validation set there are 2258 and 2240 respectively, both correspond to approximately 3 hours of total duration.

5.1.1 Tokenizer

The vocabulary of the language contains: 26 characters of the alphabet, 7 letters marked by diacritics “á”, “é”, “í”, “ó”, “ú”, “ï”, “ü” and the apostrophe. Moreover, not knowing which standard was adopted in the transcriptions of the read sentences, we included the grapheme “ç”.

5.1.2 Results

Figure 3 - *CER on the Galician test set when fine-tuning using the Galician data only (orange line) and when using Italian, Spanish, Portuguese and German as pivot, considering different amount of data for fine- tuning*

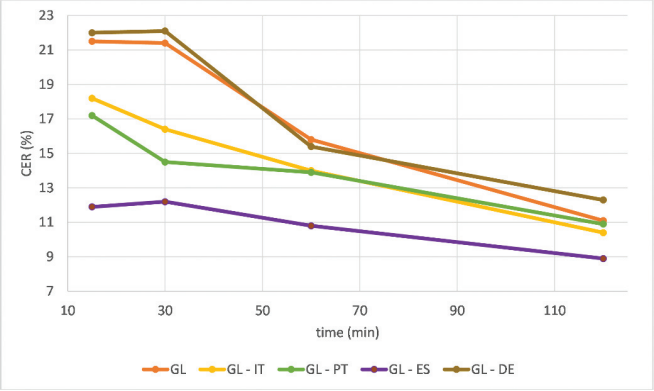
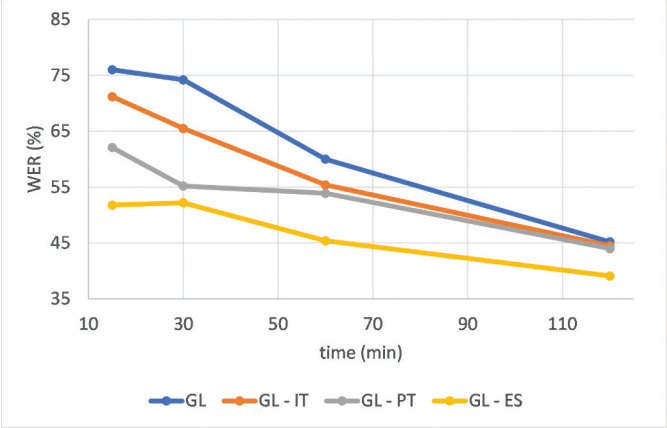


Figure 3 reports the CER obtained on Galician considering different amounts of material for training and different pivot languages. The orange line is our baseline consisting of the pure adaptation of wav2vec2.0-xls-r on the Galician data. Note that using Spanish (purple line), as expected, leads to the best performance, since it shares a large number of graphemes and have high overlap of phonemes, with more than 10 percent point reduction with respect to the baseline. Italian and Portuguese also give a clear improvement but rather inferior than Spanish. Finally, German, being a language from a different family with different phonology and orthography, brought no advantage at all, providing very similar performance as the monolingual model obtained using only Galician data. Note that the number of tokens is equivalent in all the experiments. Moreover, the performance of the Spanish system on Galician is above the baseline (around 30 % CER). Clearly, as the amount of training data increases, the gap between the different systems is reduced as the model tends to forget what learned in the first adaptation step. Note,

however, that Spanish is remarkably better than the other languages. Fig. 4 reports the results using WER as metric. Note that we are not using any language model.

Figure 4 - *Performance in terms of WER for Galician. No language model is used*



5.2 Experiments on Vallader

Spoken in the southeastern part of Switzerland, Romansh is a romance language belonging to the Rhaeto-Romance branch. In 1938 it was recognized as one of the four national languages of the country, together with German, Italian, and French. The origin and its classification inside the romance family is a debated theme among linguists.

The dataset available for the Romansh Vallader variety in Common Voice is extremely small since it contains only 2.34 hours of labelled validated data. The total number of the audio files is 1459 divided into the train, test, and validation sets. The average duration per audio file is 5.791 seconds. The number of speakers in this corpus is 50.

5.2.1 Tokenizer

To construct the tokenizer, we considered the convention adopted to transcribe the audio files, which in our case is the writing system of the Vallader variety. Differently from the standard Rumantsch Grischun, Vallader transcribes three vowels with the dieresis: “ï”, “ö”, “ü”. The vocabulary of the tokenizer contains in total 39 characters, of which five vowels accented with the grave accent “à”, “è”, “ì”, “ò”, “ù”, three vowels with the circumflex accent “â”, “ê”, “ô”, and three vowels with dieresis “ï”, “ö”, “ü”. We inserted “é” due to the possible presence of loan words from Italian that may present an acute accent.

5.2.2 Results

Figure 5 - *Performance in terms of CER on Romansh Vallader using Italian as pivot*

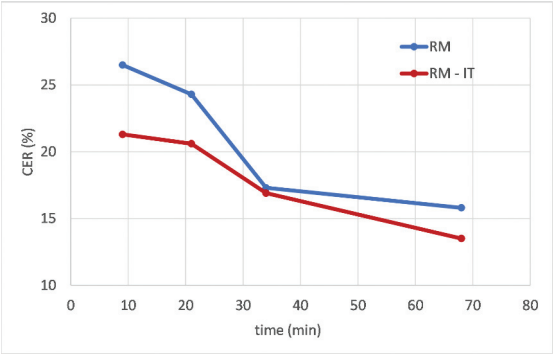


Fig. 5 shows the CER performance on Romansh Vallader with and without pivoting on the Italian language. Similar trends can be observed as for the Galician language confirming the benefit of using a well-resourced language for a first fine-tuning of the pre-trained model. Remarkably, note that the performance is similar to those of Galician using Italian when similar amounts of training data are considered.

6. *Conclusions*

In this study we presented a transfer learning approach to obtain an ASR system for very low-resourced languages using a well-resourced language as pivot. The type of language selected as the pivot was found to be influential. In the case of Galician, we observed the effect of using different romance languages. The better performances were obtained using the languages belonging to the same branch and in contact to Galician. Spanish was proven to be the best choice: the languages share a large number of graphemes and have high overlap of phonemes. The phonological differences between the two languages were demonstrated by testing the Spanish model on Galician which scored 15 % CER. Fine-tuning this model on 2 hours of the low-resource language obtains 8.9 % CER, proving the advantages of transfer learning. The experiment using German confirmed the role of the sharing between the two languages. Experiments on Romansh Vallader and Italian confirmed the trends observed for Galician

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference 2020*, Marseille, France, 4218-4222.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Platen, P.V., Saraf, Y., Pino, J.M., Baevski, A., Conneau, A., & Auli, M. (2022) XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of 23rd Interspeech Conference*, Incheon, Korea, 18-22 September 2022, 2278-2282.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (NeurIPS 2020), Vancouver, Canada, 12449-12460.

Bermuth, D., Poeppel, A., & Reif, W. (2021). Scribosermo: Fast Speech-to-Text models for German and other Languages. ArXiv 2021, ArXiv:2110.07982.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. In *Speech Communication*, 56, 85–100.

Docio-Fernandez, L., & Garcia-Mateo, C. (2002). Acoustic Modelling and Training of a Bilingual ASR System when a Minority Language is Involved. In *Proceedings of the International Conference on Language Resources and Evaluation*, Gran Canaria, Spain, May 2002, 3, 873-876.

Graves, A., Fernandez, S., Gomez, F. & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning 2006* (*ICML)*, Pittsburgh, Pennsylvania, USA, June 25-29, 369–376.

Jacobs, C., & Kamper, H. (2021). Multilingual Transfer of Acoustic Word Embeddings Improves When Training on Languages Related to the Target Zero- Resource Language. In *Proceedings of the 22nd Interspeech Conference 2021*, Brno, Czech Republic, 30 August – 3 September, 1549-1553.

Khare, S., Mittal, A., Diwan, A., Sarawagi, S., Jyothi, P., & Bharadwaj, S. (2021). Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proceedings of the 22nd Interspeech Conference 2021*, Brno, Czech Republic, 30 August – 3 September, 1529-1533.

Meng, W. & Yolwas, N. (2022). A Review of Speech Recognition in Low-resource Languages. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Learning*, Chengdu, China, 245-252

Ramallo, F. (2007). Sociolinguistics of Spanish in Galicia. In *International Journal of the Sociology of Language*, 184, 21-36.

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages. ArXiv:2012.12121.