Kalpa Publications in Computing

Volume XXX, 2022, Pages 47–50

Proceedings of V XoveTIC Conference. XoveTIC 2022

Design and Evaluation of a Cross-Lingual ML-based Automatic Speech Recognition System Fine-tuned for the Galician Language

Iván Froiz-Míguez[1,2], Óscar Blanco-Novoa[1,2], Paula Fraga-Lamas[1,2], Diego Fustes[2], Carlos Dafonte[2], Javier Pereira[2], and Tiago M. Fernández-Caramés[1,2, *, †]

[1] Department of Computer Engineering, Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain

[2] Centro de investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain

Abstract

In recent years Machine Learning (ML) strategies have proven to be useful to automate numerous classification and pattern detection tasks in diverse fields thanks to the increase of computational power in hardware. One of such fields is the Automatic Speech Recognition (ASR), which can use ML architectures to transcribe human speech into readable text. The Word Error Rate (WER) obtained with ML strategies can become relatively low while providing quick responses, reaching accuracy levels that approach human transcription accuracy. However, one of the main drawbacks in traditional architectures is the high demand of transcribed data to obtain a low WER in training. This kind of data is particularly hard to be achieved due to the high dependency on human processing. Luckily, a new framework proposed in 2020 (wav2vec2), considerably reduces the need for audio labelling thanks to the use of a Convolutional Neural Network (CNN) with self-supervised training on cross-lingual unlabelled audios of multiple languages and the ability to fine-tune the obtained results with labelled audios of a specific language. Thus, the framework can obtain results that outperform previous architectures by using much smaller audio datasets with transcriptions. This paper presents an ASR system based on wav2vec 2.0 that is fine-tuned for Galician, a language which currently only has small audio datasets available. Such a system is evaluated with a spontaneous speech dataset of approximately 1 hour from the Galicia Parliament, showing a relatively low WER (18.61%).

1 Introduction

An Automatic Speech Recognition (ASR) system is a set of mechanisms based on Machine Learning (ML) and signal processing employed for transcribing speech audio into text. There are several cases in which this technology is applied, such as to provide accessibility to multimedia content (e.g., automatic generation of subtitles or voice-based interfaces), for the transcription of legal and medical procedures, or to convert speech information into metadata.

Neural networks, and in particular Convolutional Neural Networks (CNNs), are a subset of ML that rely on training data to learn and to improve their accuracy over time through iterations. Typically, the models used in ASR require thousands of hours of labelled data for

[*] Correspondence: ivan.froiz@udc.es (I.F.-M.); paula.fraga@udc.es (P.F.-L.); Tel.: +34-981167000

[†] This work has been funded by the CITIC Research Center (by grant ED431G 2019/01)

the training of the neural network. However, getting thousands of hours of language-specific transcription is usually very difficult in order to obtain a proper Word Error Rate (WER).

Luckily, a framework for ASR based on self-supervised learning called wav2vec2 was presented in 2020 [1]. The advantage of such a framework over other CNN-based is that it employs first a self-supervised training stage with large unlabelled multi-language audios and then performs a much shorter supervised training phase with labelled audios of a specific language.

This is especially useful for low-resource languages like Galician, which only has approximately 20 hours of speech with transcription available on the most popular repositories [2, 3]. However, thanks to multi-language self-supervised training, it is possible to achieve high levels of accuracy. Moreover, it is possible to improve WER by applying a Language Model (LM) for a specific language or by adjusting the decoder for better word detection.

2 Design and Implementation

Figure 1 illustrates the designed ASR system, which consists of a pipeline of processes that receives raw speech audio and generates a transcription at the output. The pipeline can be divided into three stages: the first is the audio signal pre-processing; the second is the input of the resulted signal over the generated ML model; finally, the third is post-processing, in which the output generated in the previous stage is refined.

Therefore, the first step consists in processing the raw waveform so that it can be presented as an input to the model. This part applies several signal processing techniques and filters to suppress noise or unwanted sounds in the signal. Next, the signal is normalized as an input of the neural network. Finally, in the case of working with large audios, to get a successful audio processing by the model, it is necessary to chunk the audio input in shorter samples using a stride value to avoid cutting words. These audio segments are then processed by the trained Wav2Vec2 model to generate the representation of speech as output. Such a training is divided in two phases, named in Figure 1 as Pre-train and Fine-tuned phases (further details on how both stages work can be found in [1]).

The model output is then decoded by using a CTC decoder (Wav2Vec2CTCTokenizer), which is based on a greedy decoder (it is possible to use other types of algorithms for decoding, like the beam search [4]). Moreover, on top of the acoustic model, the use of a LM for the desired language can also reduce WER values. This post-processing task results in the transcription of the audio together with the probabilities and timestamps of each character/word.

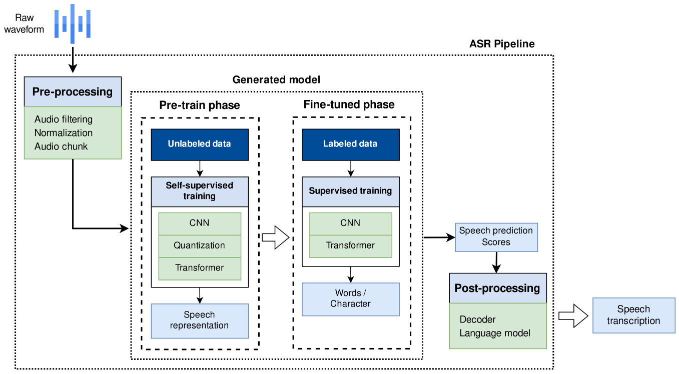

Figure 1: Architecture of the proposed ASR system.

3 Results

The proposed ASR system was implemented and fine-tuned with a corpus of approximately 20 hours (10 hours from OpenSLR and 10 from Common Voice). The total training process took approximately 34 hours using a server with 32 GB of RAM and a Nvidia Titan A100 with 80 GB of GPU memory. The training ended up with a WER of 7.15 %. The dataset was divided into 80 % for training and 20 % for evaluation, with a final training loss of 0.39 % and evaluation loss of 12.4 %. On such a model, a beam decoder with a 100-beam width was applied [4], using a Galician LM of more than 51.1 M words based on 3-gram [5] with a weight of 0.5 and an insertion factor of 0.5.

Figure 2 shows the evolution of the training loss, evaluation and WER during the fine-tuned training process. It can be observed a difference between training and evaluation loss. It is normal for the training value to be lower, but if the difference is very noticeable, it could mean a case of over-fitting [6], which is common in training especially in sets with few or unrepresentative samples. To prevent this problem, datasets are cleaned to avoid mislearning and unwanted characters with no phonetic equivalent (e.g., quotes, hyphens, slashes) are removed, accents are also removed from the fine-tuned corpus but considered later in the LM corpus.

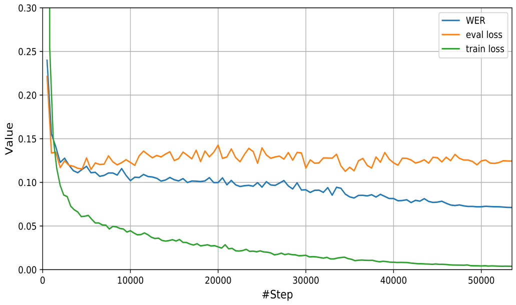

Figure 2: Evolution of train loss, evaluation loss and WER over training steps

After training, the final model was validated on a test set of approximately one hour of audio clips from the Galician Parliament [7]. Such audio was first processed with the Wav2Vec2 model, then with the model plus a beam decoder, and, finally, with the beam decoder and the LM model, obtaining a WER of 28.67%, 27.42% and 18.61%, respectively. Therefore, the combination of the WaV2Vec2 model, the beam decoder and the used LM improve WER roughly a 10% than when using the model alone.

Considering the small size of the fine-tuned corpus (approximately 20 hours) and the fact that the test corpus is spontaneous speech (not read speech, like the fine-tuned corpus), the results show 18.61% of WER over the 7.15% achieved on the training stage. These differences are usual when training and testing on different domains, so, for the test scenario, more data is required to further improve the results of ASR [8].

4 Conclusions

This article presented an ASR system based on wav2vec2 2.0 and a CNN. The system has been fine-tuned for Galician and the preliminary results look promising, achieving a relatively low WER (18.61%). Future work will include increasing the size of the fine-tuned corpus and adapting training and testing to the same content domain. Moreover, the simultaneous use of Galician and Spanish in the same transcription is a challenge that will be addressed.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," arXiv, 2020.

[2] O. Kjartansson *et al.*, "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician," in *Proc. of SLTU and CCURL*, (Marseille, France), pp. 21–27, May 2020.

[3] "Mozilla Common Voice." Available at https://commonvoice.mozilla.org/gl/datasets.

[4] E. Cohen and C. Beck, "Empirical analysis of beam search performance degradation in neural sequence models," in *International Conference on Machine Learning*, pp. 1290–1299, PMLR, 2019.

[5] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, ACM, July 2011.

[6] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics*, vol. 1168, p. 022022, feb 2019.

[7] "Mediateca do Parlamento de Galicia." Available at http://mediateca.parlamentodegalicia.gal.

[8] L. Guan-Ting *et al.*, "Analyzing the robustness of unsupervised speech recognition," arXiv, 2021.