Optimized European Portuguese Speech-To-Text using Deep Learning

Eduardo Medeiros[1]
efarofia@uevora.pt
Leonel Corado[1]
leonel.corado@uevora.pt
Luís Rato[1,2]
lmr@uevora.pt
Paulo Quaresma[1,2]
pq@uevora.pt
Pedro Salgueiro[1,2]
pds@uevora.pt

[1] Escola de Ciências e Tecnologia
Universidade de Évora
Évora, Portugal

[2] Centro ALGORITMI, Vista Lab,
Universidade of Évora, Portugal;

Abstract

We have developed an ASR system for European Portuguese implementing the QuartzNet [3] architecture with the NeMo [4] framework. Two approaches were used in this work: from scratch and using transfer learning. The experiments were data-driven focused instead of algorithm fine-tuning. Experiments confirm that models developed using transfer learning have shown better results (**WER=0.0513**) than developing models from scratch (**WER=0.1945**).

1 Introduction

Automatic Speech Recognition, commonly known as speech-to-text, is the process to transform speech into the respective sequence of words. Dictation has shown to be faster than typing but some work still needs to be done in order to improve automatic speech recognition systems, especially regarding its efficiency in less favourable conditions, *e.g.* noisy environments.

Deep learning is a subset of Machine Learning which makes use of Deep Neural Networks, a special type of Artificial Neural Networks (ANNs) that use a large number of layers. This large amount of layers allows features to be extracted from the raw input data without any pre-processing needed [2].

This work's goal is to develop an automatic speech recognition model using deep learning for the European Portuguese language.

2 Datasets

The performance of ASR systems doesn't rely solely on the type of algorithms used, it also depends on the quality and quantity of data available. Data to build an ASR system may be acquired through private entities, open access datasets, crowdsourcing, or creating the dataset from scratch *e.g.* audiobooks and respective transcriptions. Although scarce, there are some sources of audio and the respective transcriptions available for Portuguese. Nevertheless, the quantity and quality of publicly available datasets are not usually good enough to create high performance ASR systems. These sources also tend to lack in audio or transcriptions quality, data quantity and structure standardisation. The following sections describe the data sources used in this project.

2.1 LibriSpeech

LibriSpeech is a dataset from the OpenSLR repository, composed of 1000 hours of English speech audio recordings with a sampling rate of 16 kHz and was created with the purpose of being used to build and test ASR systems [5].

For this study, a 100 hours clean sub-sample of the Librispeech dataset was used as the control dataset. This control was meant to verify the performance of the transfer learning process, *i.e.*, to verify if the models pre-trained in English were properly learning the Portuguese data. The 100 hours set contains transcripts with a total of 990101 words, from which 33798 are unique words.

2.2 Multilingual LibriSpeech

Multilingual LibriSpeech (MLS) is an extension of LibriSpeech which increases the amount of English speech to 44500 hours of audio recordings and adds 6000 hours of audio recordings from seven other languages, including Portuguese among them [6].

MLS provides the dataset split into three sets – train, development and test – in which there is no speaker overlap. For the last two sets, it is also guaranteed that the gender and duration of the speaker are balanced. Audios are also ensured to unambiguously contain only one speaker. The MLS contains a large number of hours ($\simeq$168 hours of Portuguese audio recordings) when compared with the majority of other available ones (shown in hours) as stated by Lima *et al.* [1]. These correspond to a total of 1321326 words, of which 77292 are unique, and an average of 33.68 words per transcription.

2.3 SpeechDat

The SpeechDat European Project (1996-98) was developed with the goal of providing speech resources to stimulate research and development of automated services such as speech recognisers[1].

The Portuguese database of the SpeechDat project was collected by INESCTEL and had a good spread geographically.

The recorded audios were encoded using A-LAW (an algorithm used for encoding audio signals, in particular, voice encoding) with a sampling rate of 8 kHz 8-bit, which was accompanied by an ASCII SAM label file containing audio metadata.

Each label file contained an assessment code regarding the quality of the respective audio. The possible values for these codes are **OK**, **NOISE**, **GARBAGE**, **OTHER**, **NO_PTO**.

The 186 total hours of audio recordings present in the SpeechDat dataset represent a lexicon of approximately 15000 different words. The duration of audio recordings of each audio quality label (assessment code) was estimated since the project's documentation didn't provide the precise size (duration) of the Portuguese database.

SpeechDat dataset is slightly larger than the Multilingual LibriSpeech by $\simeq$18 hours. This reinforces the scarceness statement of speech data concerning the Portuguese language.

2.4 Data pre-processing

The MLS dataset wasn't subject to any pre-processing, whereas the SpeechDat was restructured and cleaned, *i.e.* only labelled OK and NOISE data samples were used. These audios, originally encoded in A-LAW with a sampling rate of 8 kHz, were converted to WAV with a sampling rate of 16 kHz. The transcriptions of these audios were also pre-processed such that markers of noise and word truncation were removed. The cleaned data was split into 75% train, 15% validation and 15% test.

After the data pre-processing, JSON manifest files (used as NeMo's input) were created for each of the datasets' subsets, each entry being the absolute path to the audio file, duration and respective transcription.

3 Model Architecture

State-of-the-art deep learning frameworks tend to explore hardware features to their full potential, such as multi-CPU, multi-GPU, and multi-node with high-speed interconnects, as offered by the **NVIDIA NeMo** [4] framework, to improve train and inference times.

The **QuartzNet** [3] architecture has the goal to achieve state-of-the-art results while using smaller models. QuartzNet architecture is based on Jasper's B$\times$R block architecture. It consists of one convolutional

[1] https://cordis.europa.eu/project/id/LE24001

pre-processing block, B×R blocks, three convolutional post-processing blocks, and uses CTC as the loss function. The R sub-blocks perform similar operations, the only difference being the replacement of the 1D-convolution with a 1D time-channel separable convolution. We used the default QuartzNet-15x5 model [3], totalling 18.9 M parameters. Each model consists of an encoder and a decoder. The encoder holds the blocks and the weights of the model, while the decoder translates the audio interpretation into letters from a given alphabet.

4 Experiments

Regarding the infrastructure, the entire pipeline was run in the Vision Supercomputer[2], which is an HPC cluster made of 2 compute nodes (NVIDIA DGX A100), interconnected for parallel processing.

Metrics used to evaluate the performance of ASR models measure the distance between the transcription generated by the model and the real one. The Word Error Rate (WER) metric, defined by Equation 1, was used to evaluate the developed models. This metric determines the distance between transcripts by evaluating substitutions, insertions, and deletions made on single words or word sequences so that the transcripts match.

$$WER = \frac{Substitutions + Insertions + Deletions}{Number\ Of\ Spoken\ Words} \quad (1)$$

WER results usually range from 0 to 1 (or 0% to 100%), but these results can be higher than the upper bound (above 100%) if the number of additional insertions, substitutions or deletions done by the model on its predictions is very large.

4.1 Train from scratch

In contrast to the SpeechDat dataset, the MLS dataset contains a highly unbalanced mix of European (≃1 hour of audio recordings) and Brazilian Portuguese (≃167 hours of audio recordings).

The Portuguese variants were divided into different subsets of the MLS dataset for these experiments. Both variants of Portuguese were used separately and together in order to carry out different sets of experiments to measure how data quantity affects the performance of models developed from scratch. Combining both subsets in the train and validation sets, and testing with the Brazilian Portuguese subset, we achieved the best result of **WER=0.8065**.

Models were also trained from scratch using the SpeechDat dataset with and without data pre-processing. Regarding these experiments, the models were developed with SpeechDat's train and validation subsets. Pre-processing data increased performance by 0.1090 achieving the best result of **WER=0.1945**. It's also worth noting that pre-trained English models performed the worst (**WER=0.9862**) when testing with SpeechDat and MLS datasets.

4.2 Transfer Learning

To generate good results, deep learning models require large amounts of data. From Section 4.1 results it can be concluded that creating deep neural models from scratch would require more data than we have access to. Experiments using both the SpeechDat and MLS datasets have been developed using the transfer learning (TL) technique. TL is a process in which different sets of data are used to complement each other, *i.e.* data from a given set is assumed to have enough of the pretended characteristics that denote the goal of the given task, which will help to generalise the data of a second set [2].

A variety of training, validation and testing sets, were used in these experiments, from using combinations of the MLS subsets to using different mixes of subsets from both SpeechDat and MLS. NeMo provides a model for the English language pre-trained with ≃3300 hours of audio recordings. This model's parameters were adapted from their initial state by excluding the decoder, which meant changing the English alphabet to the Portuguese alphabet, while keeping the encoder for reuse.

The usage of transfer learning shows to improve the model's performance when compared with models developed from scratch. Table 1 shows the best performance of models developed with the MLS subsets, and with SpeechDat subsets before and after data pre-processing.

[2] https://vision.uevora.pt/

Train/Validation	Transfer	Test	WER
Pre-trained ENG	PT + BR	BR	0.5025
	SpeechDat	SpeechDat	0.1603
	SpeechDat	SpeechDat	**0.0557**

Table 1: Best performance of models developed using transfer learning

MIX ID	Test set			
	SpeechDat	ENG	MLS	SpeechDat + MLS
0.00	**0.0513**	1.0014	0.7682	0.1912
0.25	0.1321	0.9969	0.5665	0.2155
0.50	0.0581	1.0013	0.3918	0.1221
0.75	0.0667	0.9943	0.3574	0.1225
1.00	0.0735	0.9949	0.3306	0.1231

Table 2: Average WER of the mix models over each test set

In the mix experiments, the proportion of each dataset (MLS and SpeechDat) is defined by a linear variation between 0% and 100% with a step of 25% for each mix so that the total number of audios is equal in all training and validation mixes.

After a general analysis of the results, it can be observed that the larger the amount of data from a certain source used in the train or transfer process, the better the performance over the test set from the same source. The only outliers are the results obtained on the second mix (MIX ID 0.25) as can be observed in Table 2

The best average performance of **WER=0.0513** was achieved when using the mix with data only from the SpeechDat dataset.

Data quantity is undoubtedly essential for building deep neural models, either from scratch or using transfer learning, but data quality in terms of selecting instances that contribute for model performance, or removing instances that do not, is also a critical step to successfully create these models.

5 Conclusions

In this paper we presented two methods to develop Deep Learning ASR models using European Portuguese datasets. From the proposed methods, the results yielded training from scratch, **WER=0.1945**, were inferior to those obtained with the transfer learning approach, **WER=0.0513**. Transfer learning shows a clear improvement in the performance achieving close to state-of-the-art results [3], which allows concluding that the amount and quality of data used to train the model impact its performance.

References

[1] Thales Aguiar de Lima and Márjory Da Costa-Abreu. A survey on automatic speech recognition systems for portuguese language and its variations. *Computer Speech and Language*, 62, 7 2020. ISSN 10958363. doi: 10.1016/J.CSL.2019.101055.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

[3] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. 2019. URL https://github.com/NVIDIA/NeMo.

[4] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. 2015. URL http://www.gutenberg.org.

[6] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research a preprint, 2020. URL https://www.openslr.org/94/.