Playing with NeMo for Building an Automatic Speech Recogniser for Italian

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

This paper presents work in progress for the creation of a Large Vocabulary Automatic Speech Recogniser for Italian using NVIDIA NeMo. Thanks to this package, we were able to build a reliable recogniser for adults' speech by fine tuning the English model provided by NVIDIA and rescoring it with powerful neural language models, obtaining very good performances. The lack of a standard, reliable and publicy available baseline for Italian motivated this work.

1 Introduction

The advent of the "Deep Learning Revolution" introduced astonishing changes also in the field of speech processing allowing for the development of brand new tools and devices able to recognise and synthesise speech exhibiting performances never seen before. It is sufficient to think to the new virtual assistants that populates our houses and mobile phones for getting an immediate idea about the improvements in this research field.

Most big IT companies developed, in the past 3/4 years, solutions well integrated with various devices that include high performance tools for speech processing. However, these solutions very often are not released freely, sometimes they require registrations and fees and, in the best situations, codes are free, but the models for a specific language are not available. A notable exception regards NVIDIA NeMo[1], a conversational AI toolkit built for researchers working on Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Text-To-Speech synthesis (TTS). The primary objective of NeMo is to help researchers from industry and academia to reuse prior work, namely code and pretrained models for various languages, and make it easier to create new conversational AI models, maybe adapting tools and models to specific languages or particular domains.

This paper reports an attempt to build a high performance Large Vocabulary ASR system for Italian adults' speech by exploiting all the features available in NeMo and most of the largest Italian spoken corpora available to the community.

Section 2 describes the various speech datasets used for developing the model, followed by Section 3 that describes the state of the art; in Section 4 we will describe the NeMo ASR model used in the experiments and Section 5 will discuss the experiments and the obtained results. Section 6 draws some provisional conclusions about our work.

2 Italian Spoken Corpora for ASR

This section describes the datasets we used for the creation of the Italian ASR model. We have to say that, of course, these are not the only spoken corpora available, but they are the biggest corpora commonly used for setting up an ASR system for Italian. They are typically very big, already organised and structured exactly for training ASR systems or specifically designed to maximise their impact and usefulness for ASR. We have also to say that, as far as we know, this is the first attempt to use all of them for ASR training in a single project.

2.1 Mozilla Common Voice (v7.0)

Common Voice (Ardila et al., 2020) is a crowdsourcing project started by Mozilla to create a free database for setting up speech recognition software. The project is supported by volunteers who record sample sentences with a microphone and review recordings of other users. The transcribed

[1] Other exceptions providing also multilingual models including Italian are Facebook Wav2Vec and SpeechBrain.

utterances will be collected in a voice database available under the public domain license CC0. This license ensures that developers can use the database for voice-to-text applications without restrictions or costs.

With regard to the Italian subcorpus, they currently[2] released version 7.0 (MCV7), containing 6,407 speakers for a total of 160,570 utterances with the correct transcriptions. In the standard splitting provided with the dataset the training set contains 131,041 utterances corresponding to 189.50 hours of speech, the validation set 14,764 utterances for 24.41 hours and the test set 14,765 utterances corresponding to 25.74 hours.

These splitting are very important for our experiments, as discussed in Section 5.

2.2 Multilingual LibriSpeech

Multilingual LibriSpeech[3] (MLS) dataset (Pratap et al., 2020) is a large multilingual corpus suitable for speech research. The dataset is derived from read audiobooks from LibriVox and consists of 8 languages - English, German, Dutch, Spanish, French, Italian, Portuguese and Polish. The Italian section contains 42,935 utterances for a total of 160.06 hours of transcribed speech.

2.3 VoxForge

VoxForge[4] is an open speech dataset that was set up to collect transcribed speech for use with Free and Open Source Speech Recognition Engines. The Italian portion of VoxForge contains 10,633 utterances totalling 20.16 hours of transcribed speech.

2.4 APASCI

APASCI (Angelini et al., 1994) is an Italian speech database recorded in an insulated room with a Sennheiser MKH 416 T microphone. The speech material, consisting of 2,170 utterances with a wide phonetic/diphonic coverage and totalling 2.91 hours of speech, was read by 100 Italian speakers (50 male and 50 female). The database includes the transcription of each utterance both at phonemic and at orthographic levels. This database in the past allowed to design, train and evaluate continuous speech recognition systems (speaker independent, speaker adaptive, speaker dependent, multispeakers). It was also designed for research on acoustic modelling as well as on acoustic parameters for speech recognition and for research on speaker recognition.

3 State of the Art for Italian ASR

In order to properly describe the state of the art, we should first define the typical metrics used for evaluating ASR systems. Given the system transcription for an utterance and the correct transcription extracted from the gold standard, the most important metric is certainly the Word Error Rate (WER) defined as

$$WER = \frac{(Insertions + Substitutions + Deletions)}{Gold\ Number\ of\ Words},$$

typically expressed in percentage. It compares the two transcriptions counting all the differences at word level using the edit distance between them. We can also define the Phone Error Rate (PER) and the Character Error Rate (CER) that use the same principle but applied, respectively, at phone or character level.

Examining the literature for the construction of ASR models for Italian we immediately recognise a lack of works devoted to the building of a general Large Vocabulary ASR for adults' speech. The only work we found on that was presented by Cosi and Hosom (2000), used a rather old approach to the problem (a hybrid HMM/ANN architecture) and measures the performance only on phones and not on words. Using PER instead the most common WER is a common trait of all the subsequent works we found in literature (Cosi and Pellom, 2005; Cosi, 2008; Cosi et al., 2014; Cosi, 2015) that applied a lot of different system architectures only on child speech. This large bundle of works represent the main line of research for building Italian ASR systems, but the aim of these studies is completely different from ours and, moreover, their results are not directly comparable with ours.

An exception to what we said before is represented by the work of Gretter (2014): he first built a large multilingual benchmark corpus, extracting data from the portal Euronews, consisting of about 100 hours of adults' speech for each language and, second, he developed also some ASR baselines, based on triphone Hidden Markov Models and n-gram Language models, obtaining on Italian a word recognition accuracy of 83.5% leading to a

[2] July 2021.

[3] http://www.openslr.org/94/

[4] http://www.voxforge.org/

WER=16.5%, a quite remarkable result obtained using non-neural stochastic systems.

More recent studies employing neural models were able to build other quite reliable systems. Weibin (2019) trained a system based on DeepSpeech (Hannun et al., 2014) using VoxForge, CLIPS[5], SI-CALLIOPE (Tedesco et al., 2018), LibriVox Audiobooks[6] and Mozilla Common Voice corpora for a total of 438 hours of speech, obtaining a WER=13.8% on a mixed test set. Pratap et al. (2020) made some experiments using wav2letter++[7] followed by a 5gram rescoring obtaining a test WER=28.19%. They used different test sets w.r.t. the one used in this work, thus they can only provide some general indications about WER, but they are not directly comparable to our work.

4 NVIDIA NeMo ASR

Traditional speech recognition takes a generative approach, modelling the recognition process of speech sounds acoustics (O) as $\overline{W} = \text{argmax}_W P(O|W)P(W)$ where W is a possible transcription as sequence of words. The actors of the game include a language model $P(W)$ that allows to estimate the most likely orderings of words in a given language (e.g. an n-gram model), a pronunciation model for each word in that sequence (e.g. a lexicon of phonetically transcribed words) and an acoustic model $P(O|W)$ that allows to estimate the probability of an input sequence of acoustic observations given each possible words sequence W. When we receive some spoken input, our goal would be to find the most likely sequence of text that maximises the words probability given a speech-acoustic input.

Over time, neural nets advanced to the point where each component of the traditional speech recognition model could be replaced by a neural model that had better performance and that had a greater potential for generalisation. For example, we could replace an n-gram model with a neural language model, and replace a pronunciation table with a neural pronunciation model, and so on. However, each of these neural models need to be trained individually on different tasks, and errors in any model in the pipeline could throw off the whole prediction.

Nowadays, end-to-end ASR discriminative architectures models that simply take a sequence of audio inputs and give a sequence of textual outputs, and in which all components of the architecture are trained jointly towards the same goal, largely dominate the field. The model's encoder would be akin to an acoustic model for extracting speech features, which can then be directly piped to a decoder which directly outputs text, as a sequence of characters, in a given language. If desired, we could still integrate a language model that would improve our predictions, piping it after the decoder[8].

Grasping information from NeMo github site[9], we learn that the base ASR model provided by NVIDIA is *Jasper* ("Just Another Speech Recognizer") (Li et al., 2019) a deep Time Delay Neural Network comprising of blocks of 1D-convolutional layers. The Jasper family of models are denoted as "Jasper_[BxR]" where B is the number of blocks and R is the number of convolutional sub-blocks within a block. Each sub-block contains a 1-D convolution, batch normalisation, ReLU, and dropout.

Most state-of-the-art ASR models are extremely large; they tend to have on the order of a few hundred million parameters. This makes them hard to deploy on a large scale given current limitations of devices on the edge. Another model is included into NeMo, *QuartzNet* (Kriman et al., 2020), a version of Jasper with separable convolutions and larger filters. It can achieve performance similar to Jasper but with an order of magnitude fewer parameters. Similarly to Jasper, the QuartzNet family of models are denoted as "QuartzNet_[BxR]", where B is the number of blocks and R is the number of convolutional sub-blocks within a block, and do not use the computationally costly recurrent layers in favour of more efficient convolutional layers. Each sub-block contains a 1-D separable convolution, batch normalisation, ReLU, and dropout (see Figure 1 for a complete diagram describing the QuartzNet internal structure). Both models described before optimise the Connectionist Temporal Classification (CTC) loss.

NVIDIA provided also a large number of pre-

[5] http://www.clips.unina.it

[6] https://librivox.org/

[7] https://github.com/flashlight/wav2letter

[8] Partially taken from, https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/intro.html

[9] https://github.com/NVIDIA/NeMo

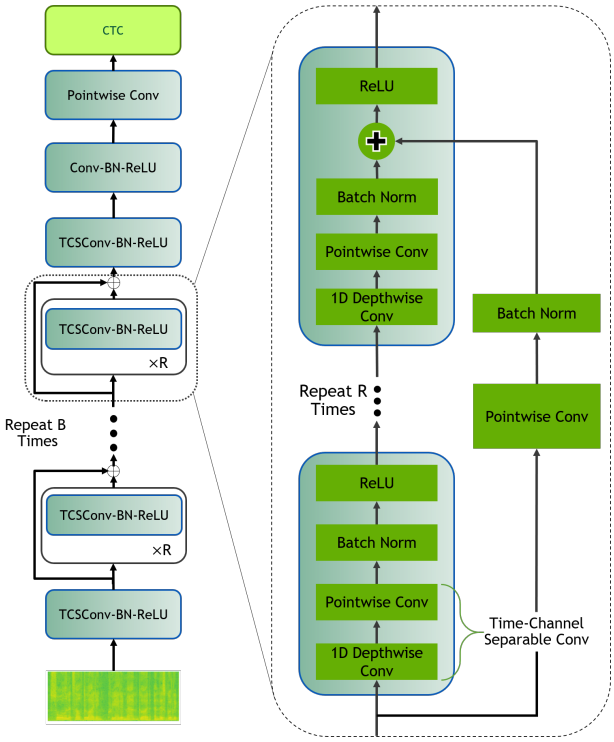

Figure 1: NVIDIA NeMo QuartzNet model.

trained models[10] for various languages. The two models for English "STT_en_Quartznet15x5" and Italian "STT_it_Quartznet15x5" (both at version 1.0.0rc1 published the 30th June 2021) are relevant for our work. The Quartznet 15x5 model family consists of 79 layers and has a total of 18.9 million parameters, with five blocks that repeat fifteen times plus four additional convolutional layers.

QuartzNet15x5 Encoder and Decoder English neural module's checkpoints from NVIDIA were trained using Multilingual LibriSpeech and Mozilla's English Common Voice 6.1 "validated" set (a huge amount of data containing more than 3,300 hours of speech) with two types of data augmentation techniques: speed perturbation and Cutout. Speed perturbation means that additional training samples were created by slowing down or speeding up the original audio data by 10%. Cutout refers to randomly masking out small rectangles out of the spectrogram input as a regularization technique. NVIDIA's Apex/Amp O1 optimization level was used for training achieveing 4.19% WER on LibriSpeech test-clean.

NeMo documentation also describes a procedure for fine-tuning the English model to adapt it to other languages, keeping the acoustic encoder frozen and fine-tuning the decoder for producing transcriptions for a different language (Huang et al., 2020). In the cited paper they also get the relevant conclusion that it is much better, in terms of performance, to fine-tune the English model than to retrain from scratch a new model for a specific language. The Italian model provided by NVIDIA has been produced following the suggested procedure, in particular by retraining the QuartzNet decoder using the training portion of MCV version 6.1. We will consider this Italian model as a baseline for our experiments.

5 Model Setup and Results

The STT_it QuartzNet model provided by NVIDIA was trained using a reduced set of data and applying an output dictionary that includes some characters that do not belong to the Italian alphabet. For these reasons we preferred to restart the fine-tuning process directly from the original STT_en_Quartznet15x5 English model.

The training set we used to fine tune the NVIDIA STT_en model to Italian is composed by joining the training portion of MCV7 and all files from MLS, VoxForge and APASCI, and contains 186,778 utterances/speech files totalling 372.62 hours of transcribed speech. 19,199 utterance/files were filtered out from the training set totalling 97.77 hours of removed speech. This is due to the fact that in some dataset, mainly in MLS and VoxForge, there were some utterances longer than 16.7 seconds, a time limit hard coded into NeMo in order to keep the model computationally tractable. We checked also that transcriptions contain only the 34 standard characters from the Italian alphabet (26 lowercase letters plus six accented characters, the apostrophe and the space) as it is a standard practice in ASR to lowercase transcriptions and to remove any punctuation mark not strictly useful or relevant to help the recognition.

With regard to decoding and rescoring, NeMo offers various possibilities:

- **Greedy Decoding**. This method simply computes the most likely sequence of characters, also called as the "best-path decoder", given the audio input.

- **Beam Search Decoding**. Beam Search Decoding (BSD) is another way of decoding model prediction that leads to better results than the greedy search. BSD, instead of choosing al-

[10] https://ngc.nvidia.com/catalog/collections/nvidia:nemo_asr

ways the best prediction at each step, considers the top-K hypothesis having the highest probabilities, where K is the so called *beam size*. For all the subsequent experiments we used beam_size=1024, beam_alpha=1.0 and beam_beta=0.5 (see NeMo documentation).

Language Models (LM) have shown to help the accuracy of ASR models when combined to BSD. NeMo currently supports the following two approaches to incorporate language models into the ASR models through BSD:

- **N-gram Rescoring**. In this approach, an N-gram Language Model is trained on text data, then it is used in fusion with beam search decoding to find the best candidates. The beam search decoders in NeMo support language models trained with the KenLM library (Heafield et al., 2013). We used this library code for building a 3-gram and a 6-gram LM using the 165-million-token-version of the CORIS corpus[11] (Rossini Favretti et al., 2002) specially cleaned and prepared for this task.
- **Neural Rescoring**. In the neural rescoring approach a neural network is used to give scores to a candidate text transcript predicted by the decoder of the ASR model. The top K candidates produced by the beam search decoding are given to a neural language model to rank them. This score is usually combined with the scores from the beam search decoding to produce the final scores and rankings. NeMo neural LMs are based on the Transformer sequence-to-sequence architecture like those described in (Vaswani et al., 2017). Again, we used the CORIS corpus described above to train an Italian neural LM from scratch and, after a month of training, we reached a perplexity of 29.30.

Given such possibilities, we fine tuned the STT_en model on a single V100 GPU using our joined dataset described above and the MCV7 validation and test set respectively for early stopping the training process and to evaluate all models. The hyperparameters we modified w.r.t. the original English model, and contained in the model itself, are listed in Table 1.

As notable exception to the NVIDIA suggested procedure for fine tuning a model, we have to report that we obtained the best results by unfreezing the encoder and letting it to slightly adapt the extracted speech features to the new language, namely Italian, that certainly share most of the sounds with the starting English model STT_en, but contains also specific sounds (e.g. [ɲ] and [ʎ]) that may require small adaptations.

[11] Corresponding to the 2021 brand new update.

Par.	**Value**
train_ds.batch_size	96
validation_ds.batch_size	4
optim.lr	0.0012
optim.betas	[0.8,0.5]
optim.weight_decay	0.001
optim.warmup_steps	500
optim.sched.min_lr	1e-10
trainer.precision	16
trainer.amp_level	O1

Table 1: Hyperparametrs modified during the fine-tuning process w.r.t. the STT_en_Quartznet15x5 model.

Table 2 outlines our results after a complete fine tuning of the end-to-end ASR model using the Italian dataset described before and applying different decoding and rescoring schemas. The improvement obtained with the fine-tuning process, when compared to the original model delivered by NVIDIA is relevant, but not so big, while when applying the BSD with the two rescoring algorithms the WER metric improve of 40% w.r.t. the greedy decoding schema.

System	**Valid.**	**Test**
Baseline (NVIDIA STT_it)		
Greedy Decoding	15.64/4.00	16.90/4.46
BSD & 3-gram Resc.	10.79/3.18	11.59/3.54
BSD & 6-gram Resc.	10.77/3.17	11.57/3.53
BSD & Neural Resc.	9.54/ -	10.51/ -
NVIDIA STT_en + Our Retraining		
Greedy Decoding	14.86/3.78	15.82/4.14
BSD & 3-gram Resc.	10.41/2.97	10.96/3.27
BSD & 6-gram Resc.	10.36/2.95	10.94/3.26
BSD & Neural Resc.	**9.04/ -**	**9.67/ -**

Table 2: WER/CER results (in percentage) on Mozilla Common Voice v7.0 (MCV7) validation and test sets.

6 Conclusions

This paper presented work in progress for the construction of a reliable and performing ASR system for Italian adults' speech. Thanks to the NVIDIA NeMo package, we were able to produce a very strong baseline reaching a WER = 9.67% over the MCV7 test set.

This is only the beginning of our work, as any change in the kind of speech used to train the system could degrade the whole performance, but, having used a collection of four different datasets containing thousands of different speakers and speech utterances for setting up such ASR system, we believe that the result should be robust enough. Unfortunately, the lack of a standardised benchmark for Italian does not allow for a quantitative and objective evaluation of this statement.

End-to-end character ASR model, and its improvement on WER, is only part of the game: the work on decoding and rescoring procedures produced much more improvements. Thus, the most important "take home lesson" is certainly to focus on the development of high performance LM specifically tuned for ASR.

All the models presented in this paper as well as the scripts and additional codes for using NeMo and generating the results will be made available[12].

Acknowledgements

We acknowledge the CINECA[13] award no. HP10C7XVUO (project QT4CLML) under the ISCRA initiative, for the availability of HPC resources and support.

References

B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. 1994. Speaker Independent Continuous Speech Recognition Using An Acoustic-Phonetic Italian Corpus. In *Proc. of the 3rd International Conference on Spoken Language Processing - ICSLP '94*, pages 1391–1394, Yokohama, Japan.

R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

P. Cosi. 2008. Recent Advances in Sonic Italian Children's Speech Recognition for Interactive Literacy Tutors. In *Proc. 1st Workshop on Child, Computer and Interaction (WOCCI '08)*, Chania, Crete, Greece.

P. Cosi. 2015. A kaldi-dnn-based asr system for italian. In *Proc. 2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5.

P. Cosi and J.P. Hosom. 2000. High performance "general purpose" phonetic recognition for italian. In *Sixth International Conference on Spoken Language Processing, ICSLP 2000/Interspeech 2000, Beijing, China, October 16-20, 2000*, pages 527–530. ISCA.

P. Cosi, M. Nicolao, G. Paci, G. Sommavilla, and T. Tesser. 2014. Comparing open source ASR toolkits on Italian children speech. In *Proc. 4th Workshop on Child Computer Interaction (WOCCI 2014)*.

P. Cosi and B.L. Pellom. 2005. Italian children's speech recognition for advanced interactive literacy tutors. In *Proc. Interspeech 2005*, pages 2201–2204.

R. Gretter. 2014. Euronews: a multilingual benchmark for ASR and LID. In *Proc. of the 15th Annual Conference of the International Speech Communication Association - INTERSPEECH 2014*, pages 1603–1607, Singapore. ISCA.

A.Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A.Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

K. Heafield, I. Pouzyrevsky, J.H. Clark, and P. Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg.

[12] https://github.com/ftamburin/ItaNeMoASR

[13] https://www.cineca.it/en

2020. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *CoRR*, abs/2005.04290.

S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 6124–6128.

J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J.M. M. Cohen, H. Nguyen, and R.T. Gadde. 2019. Jasper: An End-to-End Convolutional Neural Acoustic Model. In *Proc. Interspeech 2019*, pages 71–75.

V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proc. of , 21st Annual Conference of the International Speech Communication Association (Interspeech 2020)*, pages 2757–2761, Shanghai, China.

R. Rossini Favretti, F. Tamburini, and C. De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, and T. McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 27–38. Lincom-Europa, Munich.

R. Tedesco, S. Cenceschi, and L. Sbattella. 2018. Verso il riconoscimento automatico della prosodia. In *Proc. AISV 2018*, pages 433–439.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems - NIPS'17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

C. Weibin. 2019. *Phoenix: Deep Speech Based Automatic Speech Recognition System for Italian Language*. Master Thesis, Politecnico di Milano.