Children Speech Recognition system in a classroom context with energy consumption consideration

Nicolas Tirel

GreenAI U.P.P.A. x Prof en Poche

03-28-2022

Glossary

ASR Automatic Speech Recognition

STT Speech To Text

DNN Deep Neural Network

RNN Recurrent Neural Network

CD Context Dependant

GMM Gaussian Mixture Model

HMM Hidden Markov Model

WER Word Error Rate

LVCSR Large Vocabulary Continuous Speech Recognition

LM Language Model

1. Motivations
   - Goal
   - Challenges
2. Litterature
3. Data & Tools
4. Results
5. Improvements & next steps
6. Energy and emission

Goal

Replace Microsoft Azure STT service by an open-source solution that can run offline

- Be able to recognize and understand children speech with science vocabulary in a classroom

Goal

Replace Microsoft Azure STT service by an open-source solution that can run offline

- Be able to recognize and understand children speech with science vocabulary in a classroom
- Make sure to keep a low energy consumption for training and inference

Goal

Replace Microsoft Azure STT service by an open-source solution that can run offline

- Be able to recognize and understand children speech with science vocabulary in a classroom
- Make sure to keep a low energy consumption for training and inference
- Provide an embedded solution for smartphone

1. Motivations
   - Goal
   - Challenges
2. Litterature
3. Data & Tools
4. Results
5. Improvements & next steps
6. Energy and emission

Challenges

- We need a lot of data

Challenges

- We need a lot of data
- Corpus must be the closest to the use case

Challenges

- We need a lot of data
- Corpus must be the closest to the use case
- Requires a lot of training, and therefore more energy consumption

Challenges

- We need a lot of data
- Corpus must be the closest to the use case
- Requires a lot of training, and therefore more energy consumption
- The model and Language Model are oversized for smartphone

GMM-HMM approach

Hidden Markov Models (HMMs) provide a simple and effective framework for modelling time-varying spectral vector sequences. As a consequence, almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs. [Gales and Young, 2007]

DNN and End-to-end innovation

Experiments on a challenging business search dataset demonstrate that CD-DNN-HMMs can significantly outperform the conventional context-dependent Gaussian mixture model (GMM)-HMMs, with an absolute sentence accuracy improvement of 5.8% and 9.2% (or relative error reduction of 16.0% and 23.2%) over the CD-GMM-HMMs [Dahl et al., 2014]

This paper presents a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. [Graves and Jaitly, 2014]

DeepSpeech

Baidu Research Silicon Valley AI Lab

DeepSpeech: Scaling up end-to-end speech recognition

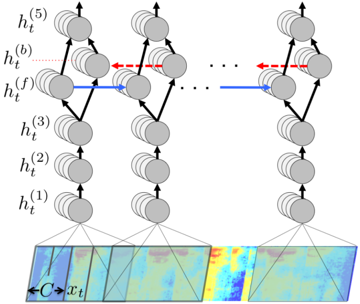

Figure 2: Structure of the RNN model and notation

[Hannun et al., 2014]

Children SR in particular

Children speech recognition is challenging mainly due to the inherent high variability in childrens physical and articulatory characteristics and expressions. [Shivakumar and Georgiou, 2020]

End-to-end architectures trained on large amounts of adult speech data can help performance on children speech. Addition of large amounts of adult speech is found to benefit more when the acoustic mismatch is large between children and adults. Although, adaptation of acoustic model on children speech helps, the recognition performance remains more than 6 times worse compared to adult ASR. [Shivakumar and Narayanan, 2021]

Energy and carbon footprint E2E ASR

This work investigates for the first time the carbon cost of end-to-end automatic speech recognition (ASR). [...] With this study, we hope to raise awareness on this crucial topic and we provide guidelines, insights, and estimates enabling researchers to better assess the environmental impact of training speech technologies [Parcollet and Ravanelli, 2021]

1. Motivations
2. Litterature
3. Data & Tools
   - Dataset
   - DeepSpeech by Mozilla
   - AIPowerMeter and Wattmeter
4. Results
5. Improvements & next steps
6. Energy and emission

Main corpus

CommonVoice : a crowdsourcing project from Mozilla with the motivation to build a high quality, publicly open dataset. It has been started in early 2019, and get updated half a year

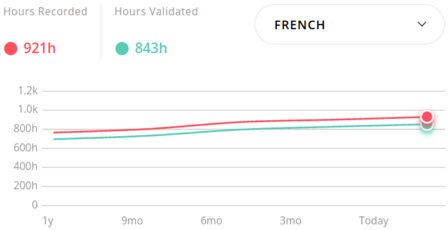

Figure 3: Evolution of the audio recorded and validated in French

Other dataset

- TranscriptionsXML MEFR (300h - 87G)
- M-ailabs (190h - 21G)
- Training Speech (180h - 56G)
- Q21_lingua_libre (40h - 6.4G)
- African accented french (22h - 2.2G)
- **mathia (5h - 1.3G)**

We hit around 1.000 hours of audio with CommonVoice included (for 200 GB of data)

Research of new data

- Multilingual LibriSpeech (MLS) (1076h - 63G)
- TED-lium3 (452h - 59G)
- TCOF (146h - 50G)
- Att-hack (28h - 11G)
- SIWIS (10h - 3.4G)

Now with the most updated CommonVoice version, all included reach 3000 hours of audio (for 500-600 GB)

Architecture

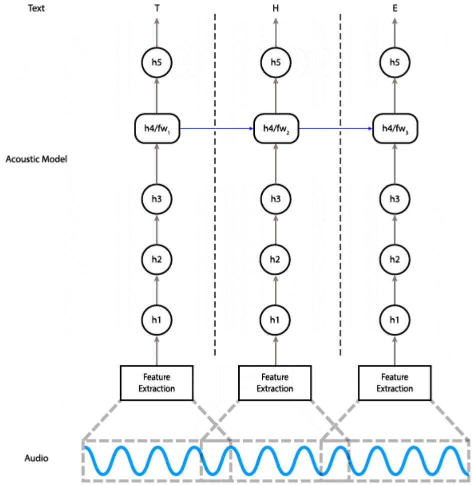

Figure 4: DeepSpeech model by Mozilla's team

Parameters

- Alphabet

Parameters

- Alphabet
- Language Model (N-grams)

Parameters

- Alphabet
- Language Model (N-grams)
- Data separated in three parts (train, dev and test separated 80-10-10 most of the time)

Parameters

- Alphabet
- Language Model (N-grams)
- Data separated in three parts (train, dev and test separated 80-10-10 most of the time)
- Hyper-parameters (epochs, learning rate, batch size...)

Mathia demonstration

Let's see how it looks like in the Mathia project !

Figure 5: Mathia : the clever assistant for mathematics

1. Motivations
2. Litterature
3. Data & Tools
   - Dataset
   - DeepSpeech by Mozilla
   - AIPowerMeter and Wattmeter
4. Results
5. Improvements & next steps
6. Energy and emission

AIPowerMeter

AIPowerMeter is a solution internally developed to track the power of the CPU and GPU. It uses the informations provided by Intel through RAPL, and nvidia-smi for the GPU, a linux command that shows a lot of information about running processes that are using the GPU.

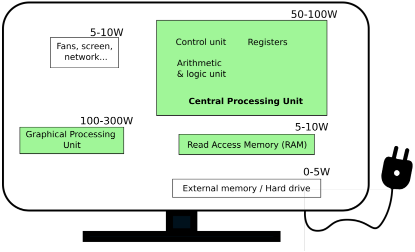

Figure 6: Sources of energy consumption in a computer

Wattmeter

In addition, the machine used for all my work at Prof en Poche is pluged to a wattmeter which measures the power used by the whole machine instead of only the CPU/GPU. We just have to integrate over time to get the energy consumption in Joules or Watt-hours.

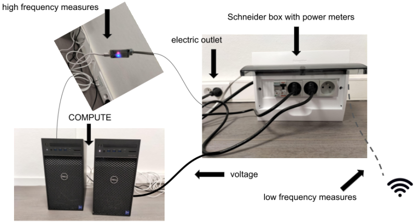

Figure 7: Wattmeter installation with low and high frequency measures

1. Motivations
2. Litterature
3. Data & Tools
4. Results
   - Best models
   - Dashboard
5. Improvements & next steps
6. Energy and emission

Original best model

Trained with Lingua Libre, African Accented, CCPMF, training speech, M-AILABS, **mathia** and CommonVoice v5, therefore fine-tuned with **mathia** corpus

Score

WER: 0.186813, CER: 0.127046, loss: 14.883443

Current best model

Trained in three steps decreasing learning rate each time and for 40 epochs :

- CommonVoice 8 only with a learning rate of 0.001
- CommonVoice and **mathia** with a learning rate of 0.0001
- **mathia** only with a learning rate of 0.00005

Score (for a total of 28.25 kWh consumed)

WER: 0.187479, CER: 0.123425, loss: 12.353087

Dashboard

For any past or current training, we want to know :

- Parameters used (epochs, learning rate, dropout rate etc..)

Dashboard

For any past or current training, we want to know :

- Parameters used (epochs, learning rate, dropout rate etc..)
- Data for test, validation and training

Dashboard

For any past or current training, we want to know :

- Parameters used (epochs, learning rate, dropout rate etc..)
- Data for test, validation and training
- Total consumption of the GPU and/or the whole machine

Dashboard

For any past or current training, we want to know :

- Parameters used (epochs, learning rate, dropout rate etc..)
- Data for test, validation and training
- Total consumption of the GPU and/or the whole machine
- Results of the model in WER and CER (Word/Character Error Rate

Dashboard

For any past or current training, we want to know :

- Parameters used (epochs, learning rate, dropout rate etc..)
- Data for test, validation and training
- Total consumption of the GPU and/or the whole machine
- Results of the model in WER and CER (Word/Character Error Rate
- Upload and see the result of an audio

Dashboard demo

All of that information are grouped in a dashboard. We can compare any model, but as well follow the power consumption of the current training in real time !

Again, let's see what it looks like !

Ideas to improve our results

- Incorporate new dataset in the training

Ideas to improve our results

- Incorporate new dataset in the training
- Train with other hyperparameters

Ideas to improve our results

- Incorporate new dataset in the training
- Train with other hyperparameters
- Update use case with more recent utterances

Ideas to improve our results

- Incorporate new dataset in the training
- Train with other hyperparameters
- Update use case with more recent utterances
- Implement coqui.ai

Ideas to improve our results

- Incorporate new dataset in the training
- Train with other hyperparameters
- Update use case with more recent utterances
- Implement coqui.ai
- Look closer to the poor transcripts

Embedded solution

Give a try with pruning and sparsity solutions to reduce space and time computation.

The goal as well is to make the solution embedded, we need therefore to reduce the size of the model. Thanks to recent work published on coqui blog; we can reduce the size from 188 to 47 MB, but the main problem remaining is the Language Model with **685MB** !!

1. Motivations
2. Litterature
3. Data & Tools
4. Results
5. Improvements & next steps
6. Energy and emission

Visualize energy consumption

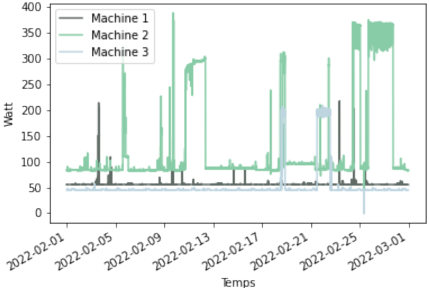

Figure 8: Active power of February

Energy consumption (in kWh/GJ)

Machine 1 : 37.55/1.35 - **2** : 85.86/3.09 - **3** : 35.06/1.26

Some orders of magnitude in energy

The three machines consumed in total 158.47 kWh or 5.7 GJ for the period. To visualize it, that represents :

- 2.88 times the annual consumption of numeric services per capita in the EU-28 [Bordage et al., 2021]

Some orders of magnitude in energy

The three machines consumed in total 158.47 kWh or 5.7 GJ for the period. To visualize it, that represents :

- 2.88 times the annual consumption of numeric services per capita in the EU-28 [Bordage et al., 2021]
- 1.56 times the consumption of my apartment in the same period

Some orders of magnitude in energy

The three machines consumed in total 158.47 kWh or 5.7 GJ for the period. To visualize it, that represents :

- 2.88 times the annual consumption of numeric services per capita in the EU-28 [Bordage et al., 2021]
- 1.56 times the consumption of my apartment in the same period
- 1042.57 hours (or 43+ days non-stop) of streaming video with a 50" TV, Wifi, 4K [(IEA), 2020]

Some orders of magnitude in energy

The three machines consumed in total 158.47 kWh or 5.7 GJ for the period. To visualize it, that represents :

- 2.88 times the annual consumption of numeric services per capita in the EU-28 [Bordage et al., 2021]
- 1.56 times the consumption of my apartment in the same period
- 1042.57 hours (or 43+ days non-stop) of streaming video with a 50" TV, Wifi, 4K [(IEA), 2020]
- 1800 kettle uses (3 people can drink 21 teas every day) [Murray et al., 2016]

And in CO2 equivalent

According to the ADEME, it represents an emission of 9.5 kgCO2e [ADEME, 2020b]. In order to visualize, we release the same amount of CO2e with :

- Between 1 and 18 meals (1.3 with animal dominant, and 18.6 with vegetarian diet) [ADEME, 2017]
- 98 km with a new car in average [ADEME, 2020a]
- Buying a new polo [ADEME, 2018]

To conclude

If you want to go further and take concrete actions :

- Measure your carbon footprint
- Become a player of the change : participate in The Climate Fresk, change your diet to have an impact 10 times more important than shutting down the 3 machines [Dugast and Soyeux, 2019], Spread the Word
- Read the GIEC/IPCC reports (and bonpote, Le réveilleur, Pour un réveil écologique)

All models pollute [Parcollet and Ravanelli, 2021]

Thanks!

References I

[ADEME, 2017] ADEME (2017).
Approche repas moyen français.
Source here.

[ADEME, 2018] ADEME (2018).
Modélisation et évaluation du poids carbone de produits de consommation et biens déquipements.
Source here.

[ADEME, 2020a] ADEME (2020a).
Evolution du taux moyen d'émissions de co2 en france - véhicules particuliers neufs vendus en france.
Source here.

References II

[ADEME, 2020b] ADEME (2020b).
Mix réseau électrique - france continentale - moyen.
Source here.

[Bordage et al., 2021] Bordage, F., de Montenay, L., et al. (2021).
Le numérique en europe : une approche des impacts environnementaux par l'analyse du cycle de vie.
Source here.

[Dahl et al., 2014] Dahl, G. E. et al. (2014).
Context-dependent pre-trained deep neural networks for large vocabulary speech recognition.

[Dugast and Soyeux, 2019] Dugast, C. and Soyeux, A. (2019).
Faire sa part ?
Source here.

References III

[Gales and Young, 2007] Gales, M. and Young, S. (2007).
The application of hidden markov models in speech recognition.

[Graves and Jaitly, 2014] Graves, A. and Jaitly, N. (2014).
Towards end-to-end speech recognition with recurrent neural networks.

[Hannun et al., 2014] Hannun, A. Y. et al. (2014).
Deepspeech: Scaling up end-to-end speech recognition.

[(IEA), 2020] (IEA), G. K. (2020).
The carbon footprint of streaming video: fact-checking the headlines.
Source here.

References IV

[Murray et al., 2016] Murray, D. et al. (2016).
Understanding usage patterns of electric kettle and energy saving potential.

[Parcollet and Ravanelli, 2021] Parcollet, T. and Ravanelli, M. (2021).
The energy and carbon footprint of training end-to-end speech recognizers.

[Shivakumar and Georgiou, 2020] Shivakumar, P. G. and Georgiou, P. (2020).
Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations.

References V

[Shivakumar and Narayanan, 2021] Shivakumar, P. G. and Narayanan, S. (2021).
End-to-end neural systems for automatic children speech recognition: An empirical study.