Automatic Speech Recognition for children voices in MathIA solution

Nicolas Tirel

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

22-05-2023

Glossary

ASR Automatic Speech Recognition

CD Context Dependant

DNN Deep Neural Network

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

LM Language Model

RNN Recurrent Neural Network

TTS Text To Speech

VC Voice Conversion

WER Word Error Rate

MathIA

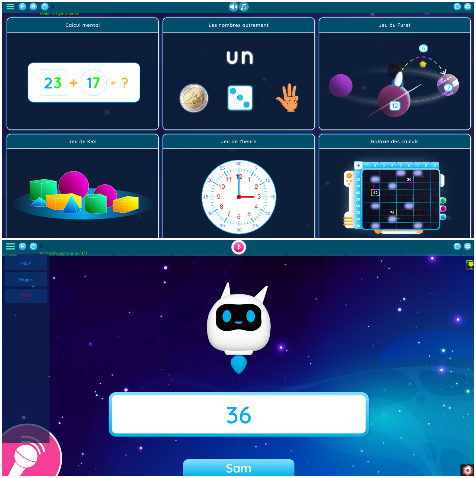


Figure 2: The MathIA interface

Goal and challenges

To develop a solution able to give a result with a lighter open-source solution instead of Microsoft Azure STT service :

- Recognize french and maths vocabulary in a classroom
- We need a corpus with a lot of data closest to the use case
- Requires huge training with big energy consumption

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

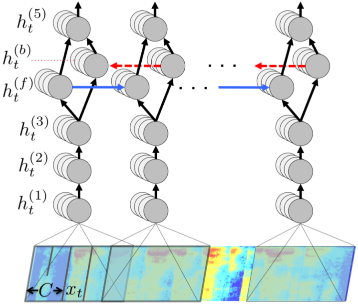


Figure 3: Structure of the RNN model and notation

[Hannun et al., 2014a]

Architecture

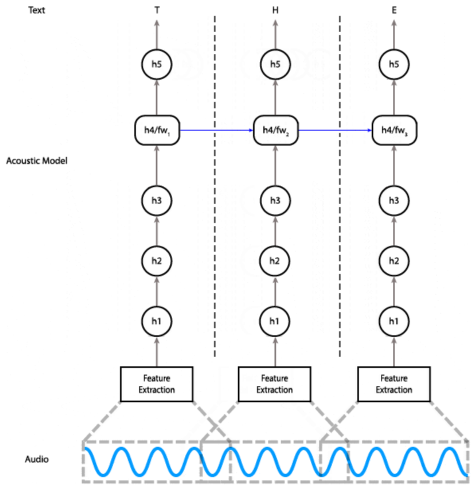


Figure 4: DeepSpeech model by Mozilla's team

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]

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 every two/three months.

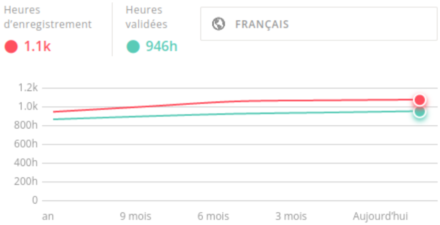


Figure 5: Evolution of the audio recorded and validated in French

Other dataset

- Multilingual LibriSpeech (1100h)
- M-ailabs (315h - 42G)
- Training Speech (180h - 56G)
- Q21_lingua_libre (40h - 6.4G)
- African accented french (15h - 2.2G)
- **mathia (5h - 1.3G)**

We hit around 2.500 hours of audio with CommonVoice included (for +200 GB of data)

Spontaneous dataset

To go further, we need to get a dataset as close as possible to the use case. We decided to validate unlabeled audios from the people using the app with transcription from Microsoft Azure

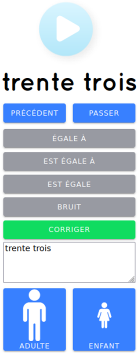


Figure 6: Validate audios and precise if there's noise ("bruit" in french)

Validation and sort

We listen more than 13600 audios :

- 11054 were validated
- 7550 were children voices, 5h18
- 2776 with noise

Previous 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

Best model

Trained with the mix of dataset, then with both **validated audios**, and the **mathia** corpus, using a specific Language Model and the best alpha and beta hyper-parameters

Score (for a total of 25,93 kWh consumed)

WER: 9.03%, CER: 5.73%, loss: 08.99 - **Old dataset**
WER: 11.22%, CER: 9.02%, loss: 07.31 - **New dataset w/o noise**
WER: 23.30%, CER: 19.99%, loss: 16.20 - **New dataset w/ noise**

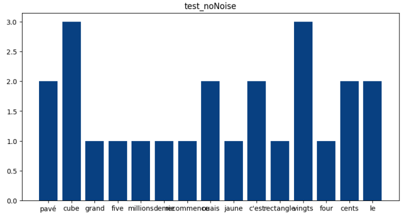
Azure model

We can compare using the result from Azure with the new audios

Score

WER: 06.19%, CER: 04.56% - **New dataset w/o noise**
WER: 20.94%, CER: 17.43% - **New dataset w/ noise**

Pitfalls



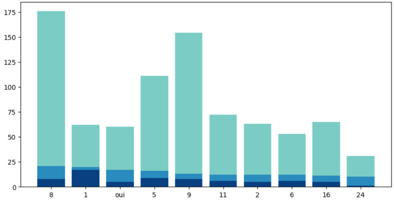


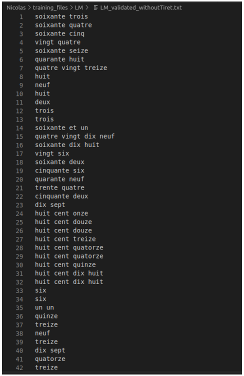
Figure 7: Wrong words and answers

What is a LM ?

A Language Model is created using a corpus of text, gets a sentence as input and returns the probability of the last word given all the previous words. It was used in 2014 for decoding CTC output with an important improve : an acoustic model could go from a WER of 35.8% to 14.1% [Hannun et al., 2014b] Really good explanation can be found here

Number LM

Once we know the specific vocabulary, i.e. be able to recognize numbers, yes, no, and some geometric shapes, we can write all of them in a file, and convert them using KenLM toolkit.


Nicolas > training_files > LM > LM_validated_withoutTiret.txt

```
soixante trois
soixante quatre
soixante cinq
vingt quatre
soixante seize
quarante huit
quatre vingt treize
huit
neuf
huit
deux
trois
trois
soixante et un
quatre vingt dix neuf
soixante dix huit
vingt six
soixante deux
cinquante six
quarante neuf
trente quatre
cinquante deux
dix sept
huit cent onze
huit cent douze
huit cent douze
huit cent treize
huit cent quatorze
huit cent quatorze
huit cent quinze
huit cent dix huit
huit cent dix huit
six
six
un un
quinze
treize
neuf
treize
dix sept
quatorze
treize
```

Figure 8: LM from all the validated transcription

Voice Conversion

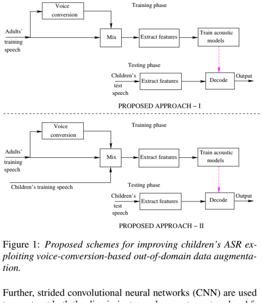


Figure 9: Voice conversion (VC) is a technique for transforming the non/para-linguistic information of given speech while preserving the linguistic information[Shahnawazuddin et al., 2020]

CycleGan-VC2

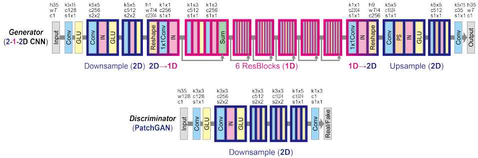


Figure 10: A CycleGAN learns forward and inverse mappings simultaneously using adversarial and cycle-consistency losses.[Kaneko and Kameoka, 2017]

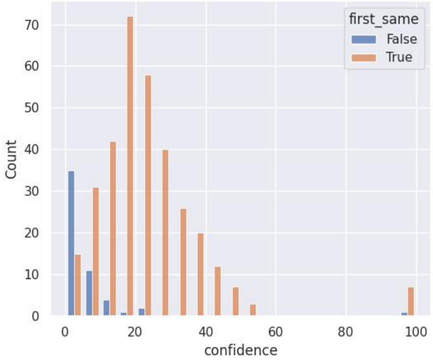
Confidence from the Coqui model

The confidence evaluation of every candidates for this transcription is roughly the sum of the acoustic model logit values for each timestep/token that contributed to the creation of this transcription.

```
[start]---------===============-----------[end]
Result 0 confidence =-3.7841262817382812 text : sept in digit : 7 text original : sept in digit : 7
Result 1 confidence =-20.998090744018555 text : sept  in digit : 7  text original : sept in digit : 7
Result 2 confidence =-25.09174346923828 text : et in digit : et text original : sept in digit : 7
Result 3 confidence =-27.75187110900879 text : cent in digit : 100 text original : sept in digit : 7
Result 4 confidence =-32.596405029296875 text : est in digit : est text original : sept in digit : 7

Recognized Text: sept
Confidence :  21.3076171875
```

Figure 11: For 5 candidates, we measure the distance between first candidate and the first **different** answer in the 4 candidates left




```
Only numbers : Counting only the first answer, coqui has a precision of 86.05 % for numbers, counting all candidates : 94.57 %
```

Figure 12: Confidence evaluation with the test no noise dataset

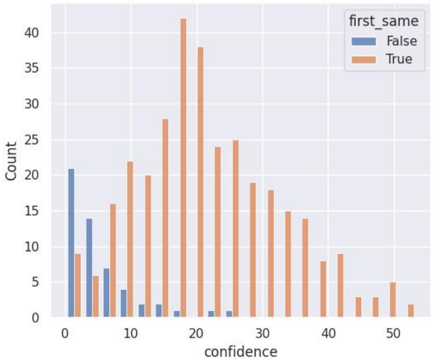


Figure 13: Remove noise

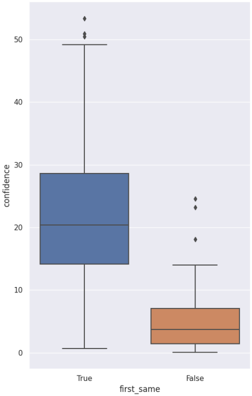


Figure 14: Confidence for good or bad first candidate

```
With an elimination of 95.0% of bad answers, we keep 70.968% of good answers
```

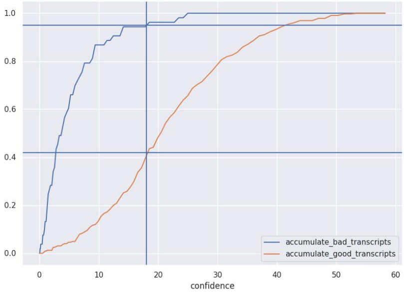


Figure 15: With a confidence of 18, we keep a very good score

Everything together !

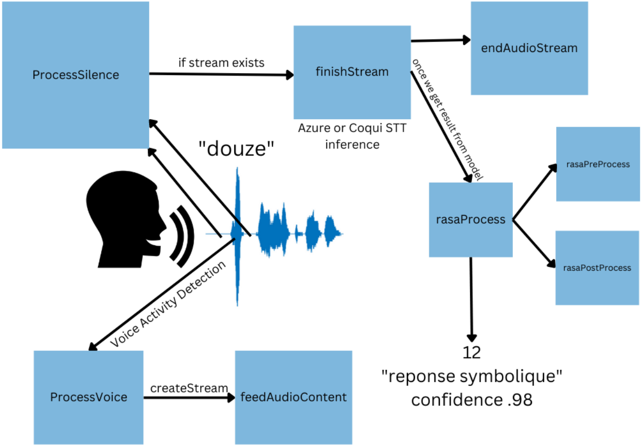


Figure 16: Production schema

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]

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.

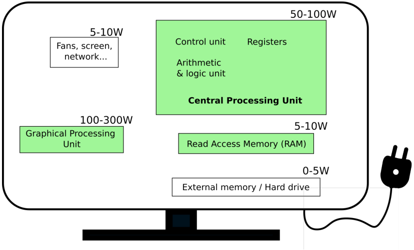


Figure 17: Sources of energy consumption in a computer

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.

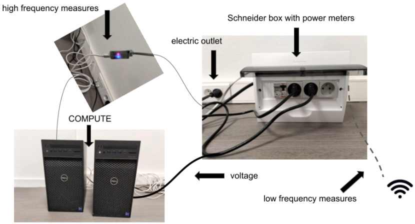


Figure 18: Wattmeter installation with low and high frequency measures

Training has a huge impact

When using our three machines, we can see a huge increase during training, and one of them consumes around 100 kWh for a single training. The emission related is highly dependant of the country of production, in France with 60 grams per kWh we get 6 kg of CO2e emissions, but if we did this in Poland it rises to 73 kg ! [Ritchie et al., 2020]

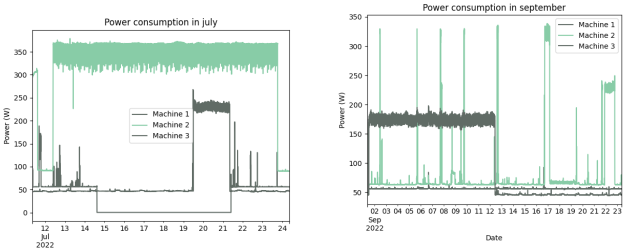


Figure 19: Power consumption of three machines in July and September

Some orders of magnitude in energy

During one month, 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]

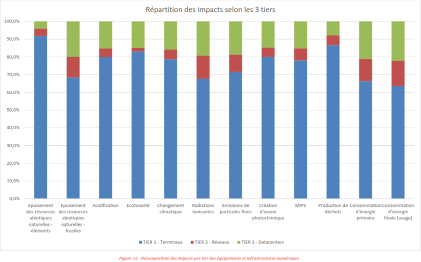
And in CO2 equivalent

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]

Electricity is not the only impact

When talking about numeric emission, we always think about the electricity or the data centers, but we need to think also about the fabrication process, which is responsible of 80% of the footprint in the life cycle assessment [Déragne and Mouneu, 2020]



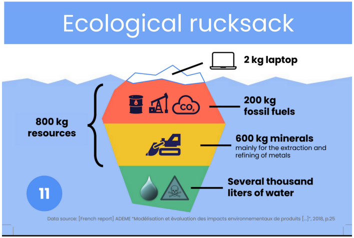


Figure 20: Planet boundaries are not only about CO2e [et ARCEP, 2022]

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, or The Digital Collage, keep your numeric equipment 10 years at least, change your diet to have an impact 10 times more important than shutting down the 3 machines [Dugast and Soyeux, 2019], think systemic !
- Read the IPCC reports, "L'âge des low tech" Philippe Bihouix, watch "Ruée minière au XXIè siècle : jusqu'où les limites seront-elles repoussées ?" - Aurore Stephant at USI...

"Sans un plan de sobriété, les impacts environnementaux du numérique tripleront dici 2050" (Ademe et Arcep) [Breteau, 2023]

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.

[Breteau, 2023] Breteau, L. (2023).
Etude greenit.fr.
Source here.

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

References III

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

[Déragne and Mouneu, 2020] Déragne, A. and Mouneu, Y. (2020).
The digital collage.
Source here.

[et ARCEP, 2022] et ARCEP, A. (2022).
Evaluation de l'impact environnemental du numérique en france et analyse prospective.
Source here.

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

References IV

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

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

[Hannun et al., 2014b] Hannun, A. Y., Maas, A. L., Jurafsky, D., and Ng, A. Y. (2014b).
First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns.

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

References V

[Kaneko and Kameoka, 2017] Kaneko, T. and Kameoka, H. (2017).
Parallel-data-free voice conversion using cycle-consistent adversarial networks.

[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.

References VI

[Ritchie et al., 2020] Ritchie, H., Roser, M., and Rosado, P. (2020).
Energy.
Our World in Data.
https://ourworldindata.org/energy.

[Shahnawazuddin et al., 2020] Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., and Ahmad, W. (2020).
Voice conversion based data augmentation to improve childrens speech recognition in limited data scenario.

[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 VII

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