SPEECH RECOGNITION UNDER CONSTRAINT

MATZ, O.; EL OUAZZANI, M.; GANTET, J.; DIARRA, N.; DEGUILHEM, B.
R&I Department @ Capgemini Engineering – DEMS

CONTEXT & OVERVIEW

Currently, Speech Recognition solutions that can be used in production are owned by large groups (IBM, AWS, Microsoft, Google, etc.) and are available as cloud services. In most cases, these offers do not meet the needs of manufacturers because (i) they do not guarantee data confidentiality (manufacturers refuse to send their private and sensitive data to the cloud), (ii) the cost of use is important and (iii) they are not adapted to the specific business vocabulary of the manufacturer.
In this context, the development of a voice recognition solution deployed locally, requiring little data labeled for training and robust to noisy environments represents a real challenge in the field of industry.

OBJECTIVES

The main objectives of this work are to develop an end-to-end voice recognition solution:

1. From a small amount of labelled data (<100h)
2. Robust to noisy environments
3. Capable of running on local equipment in real time (onboard hardware, PC fixed or portable,…
4. And with an acceptable precision (i.e., WER <10% on real conditions data).

BIBLIOGRAPHY

Q. Xu et al., *'Self-Training and Pre-Training Are Complementary for Speech Recognition'*, 2020.

D. Amodei et al., *'Deep Speech 2: End-to-End Speech Recognition in English and Mandarin'*, 2015.

A. Baevski et al. *'wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations'*, 2020.

A. Défossez et al. *'Demucs: Deep Extractor for Music Sources with Extra Unlabeled Data Remixed'*, 2019.

A. Défossez, et al. *'Real Time Speech Enhancement in the Waveform Domain'*, 2020.

Contact: olivier.matz@capgemini.com

NOISY ENVIRONMENT

Speech Recognition algorithms are benchmarked on "laboratory conditions" datasets which explained the excellent accuracy reported.
⇒ The environmental noise can raise the WER up to 80%

Our work on the denoising task is inspired by audio source separation methods initially developed in the musical field

LOW RESOURCE DATA

Marketed voice recognition solutions and most of the state-of-the-art works are based on supervised approaches which require many thousands of hours of labelled audio.
The need for such databases represents a major obstacle to the development of voice recognition solutions, especially in an industrial context (the cost of acquiring such databases, the lack of data, the resources required for storage and training, ...)

CHALLENGE TO SOLVE

DATA CENTRIC

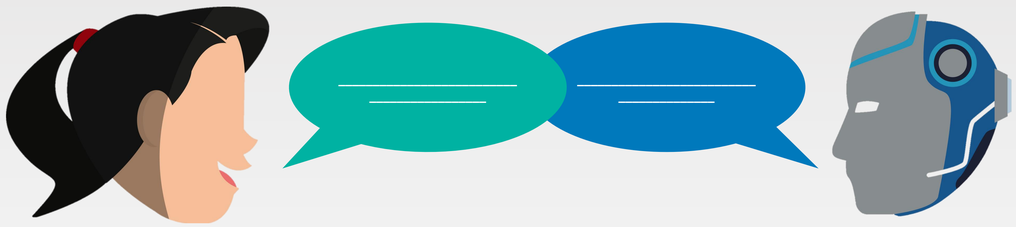
Our works highlight the gap between benchmark and real-life data resulting in a nonsignificant decrease of accuracy
⇒ non transferability to the state-of-the-art model in the real world

Our methods are based on a data centric approach where we build realistic dataset for both denoising and speech to text aera.

BRINGING AI ON EDGE

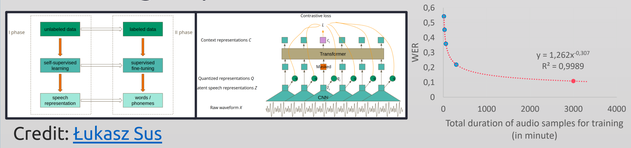
To respond to this problem, we are working on two main axes:

1. Development of a light model by reducing the model complexity and its number of parameters.
2. By reducing the size of the model via model reduction methods such as pruning and quantization.



FRAMEWORK, ARCHITECTURE & PRELIMINARY RESULTS

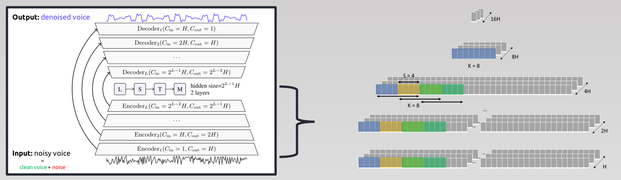
Self-supervised learning:
Our approach is based on the recent work of Baevski et al. using a pre-training step via a self-supervised approach to learn a contextual latent representation of an audio sample which will then be fine-tuned on a Speech to Text task. This work is done within FAIRSEQ, and training are performed on AWS and GENCI.



Credit: Łukasz Sus

Our first results highlight that only 50 hours of labelled speech allow to reach a WER of 10%.

Denoising:
Our work consist to adapt approaches based on the separation of audio sources that have proven their worth in the musical field with an architecture CNN+LSTM compatible with audio real time analysis.



Our first results demonstrate the lack of representativeness of reference dataset with a decrease of efficiency up to 50%.

	Valentini*		S-Noise200		Noise400	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
noisy	1,97	0,92	1,09	0,74	1,49	0,85
dns48	2,93	0,95	1,45	0,81	1,84	0,88
dns64	2,91	0,95	1,45	0,81	1,91	0,89
master64	3,07	0,95	1,47	0,81	2,07	0,89

*Valentini-Botinhao, *'Noisy Speech Database for Training Speech Enhancement Algorithms and TTS Models'*, 2017

As a part of our data centric approach, we have built 2 in house datasets of noisy speech:
– 200h of white noise
– 400h of more than 100 types of noise

Capgemini engineering