Conv-SINet

Identify Speaker using a convolutional solution

Use case

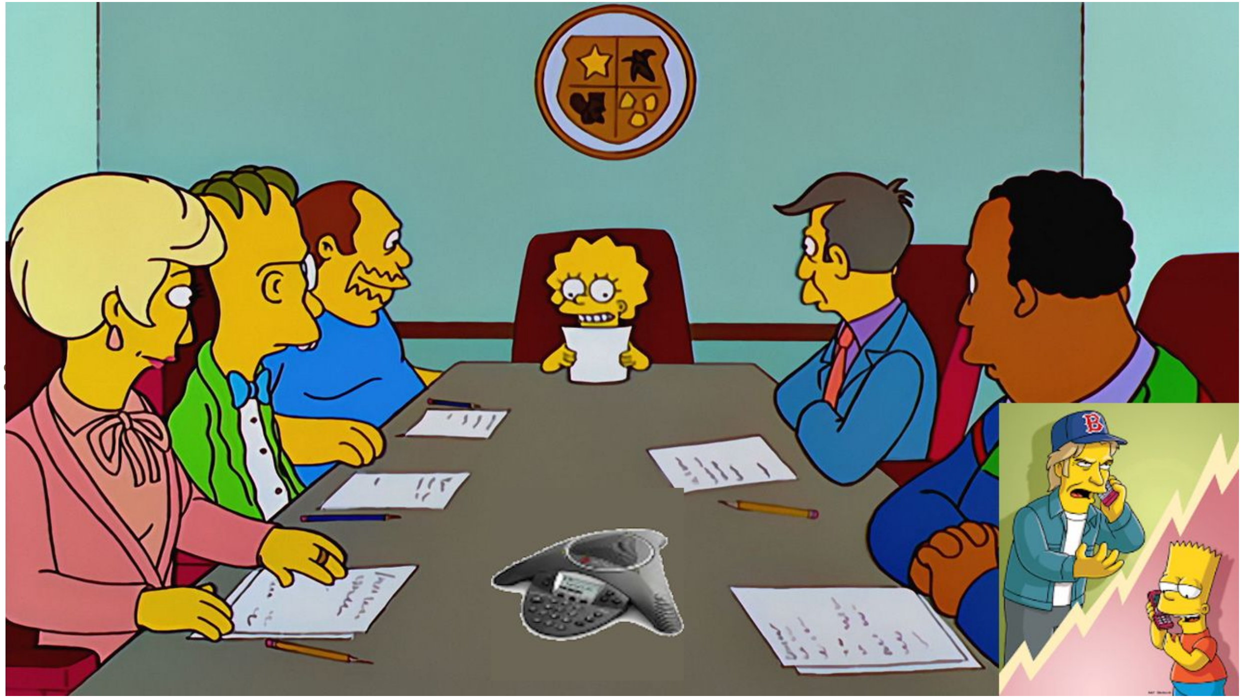
In meeting room:

Théo, Alban, Florian,

Timothé, Michèle, Liza

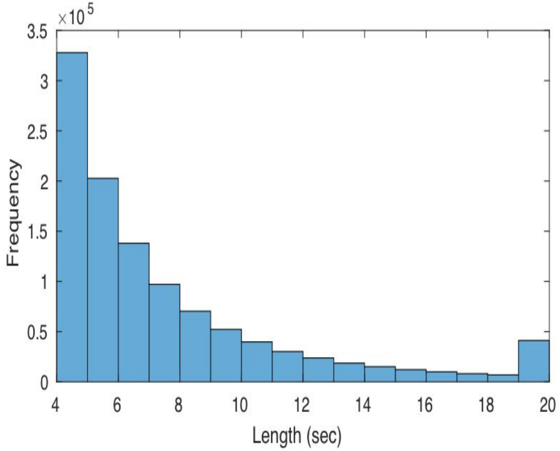
Under personal phone:

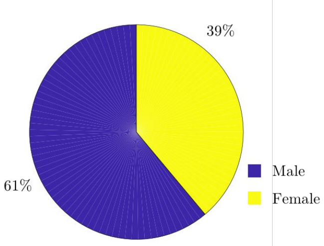
Bart and Simon

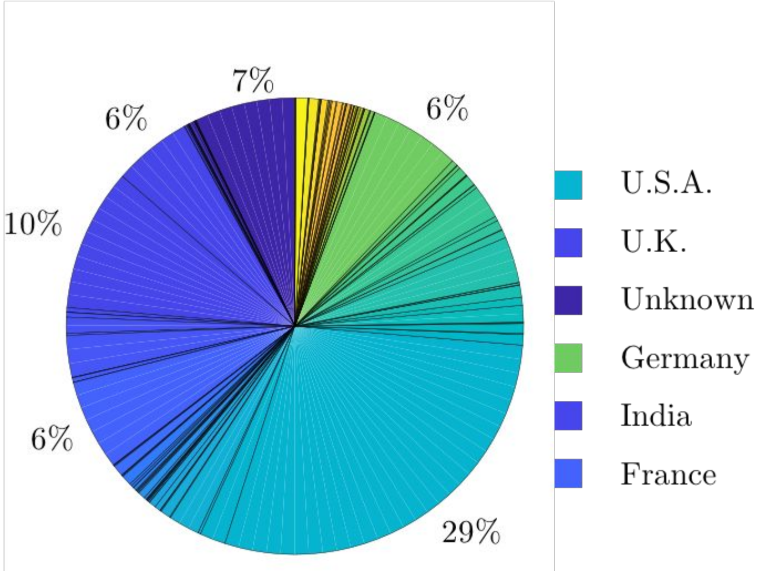


Data set: VoxCeleb1

A wide range of different ethnicities, accents, professions and ages.







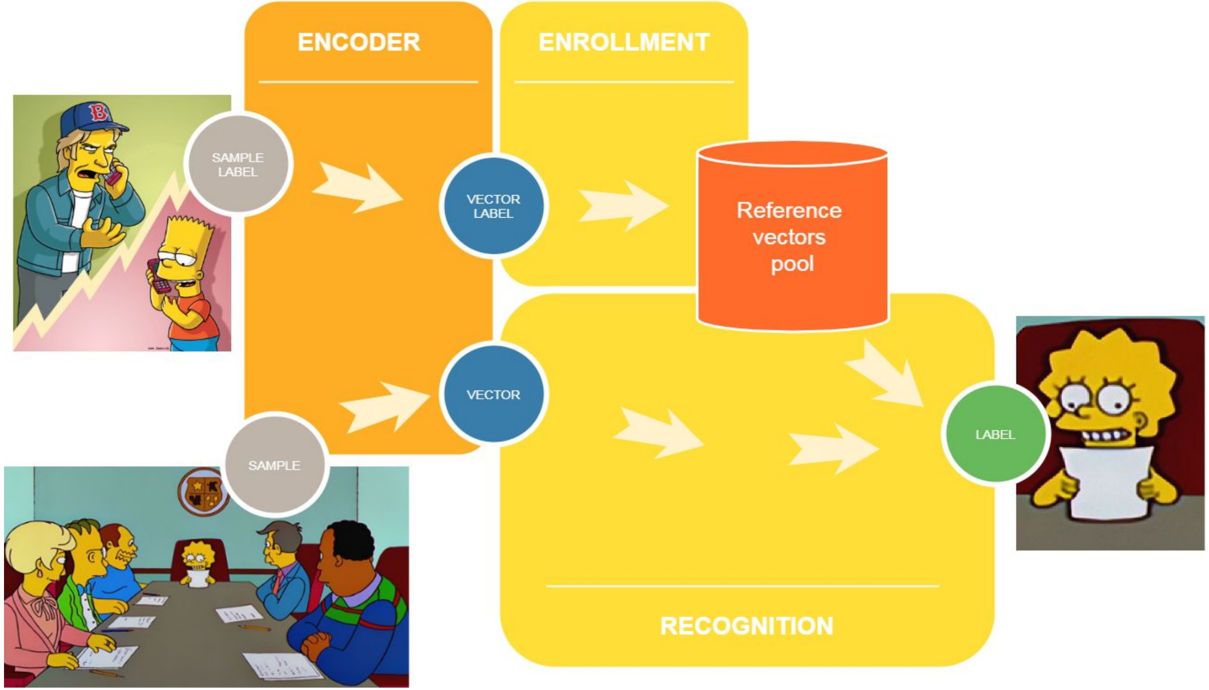
Cross validation and data set

VS

The David data-set split.

Uniform background sound
for some speakers.

General architecture



deep learning part

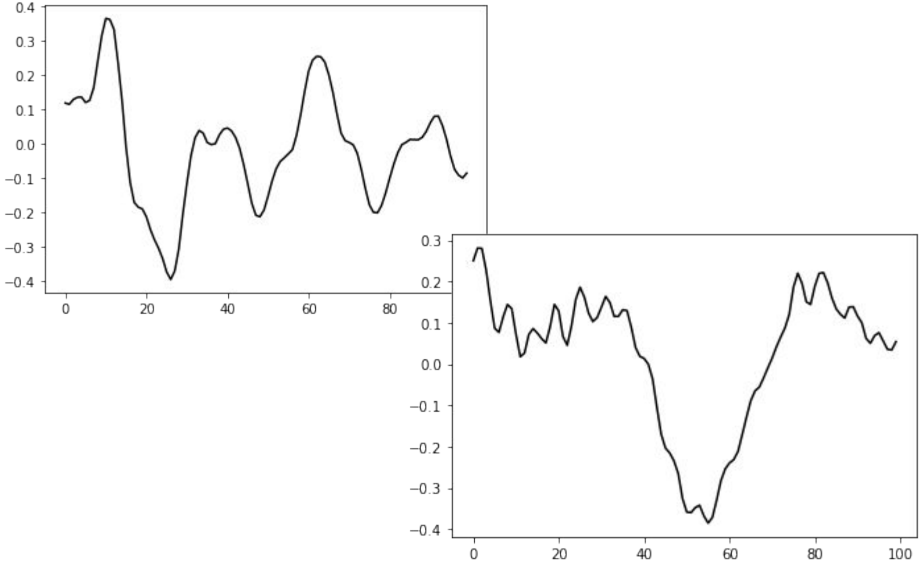
Trained one time for all

machine learning part

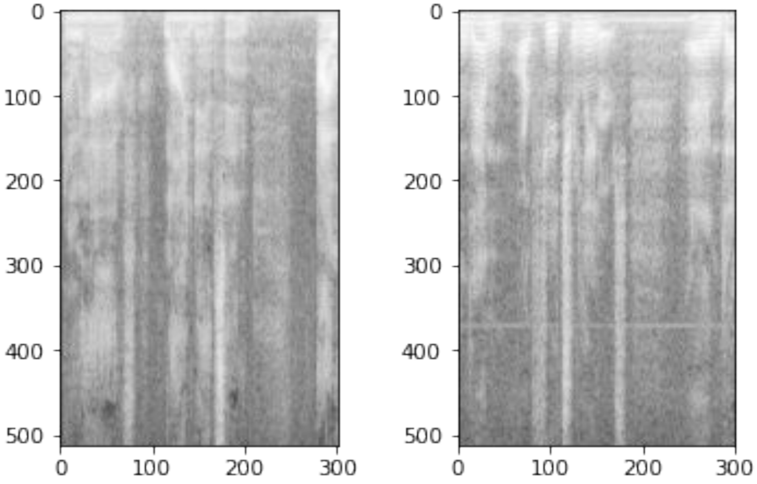
Trained by company

Data set: time vs STFT

16 kHz raw sample



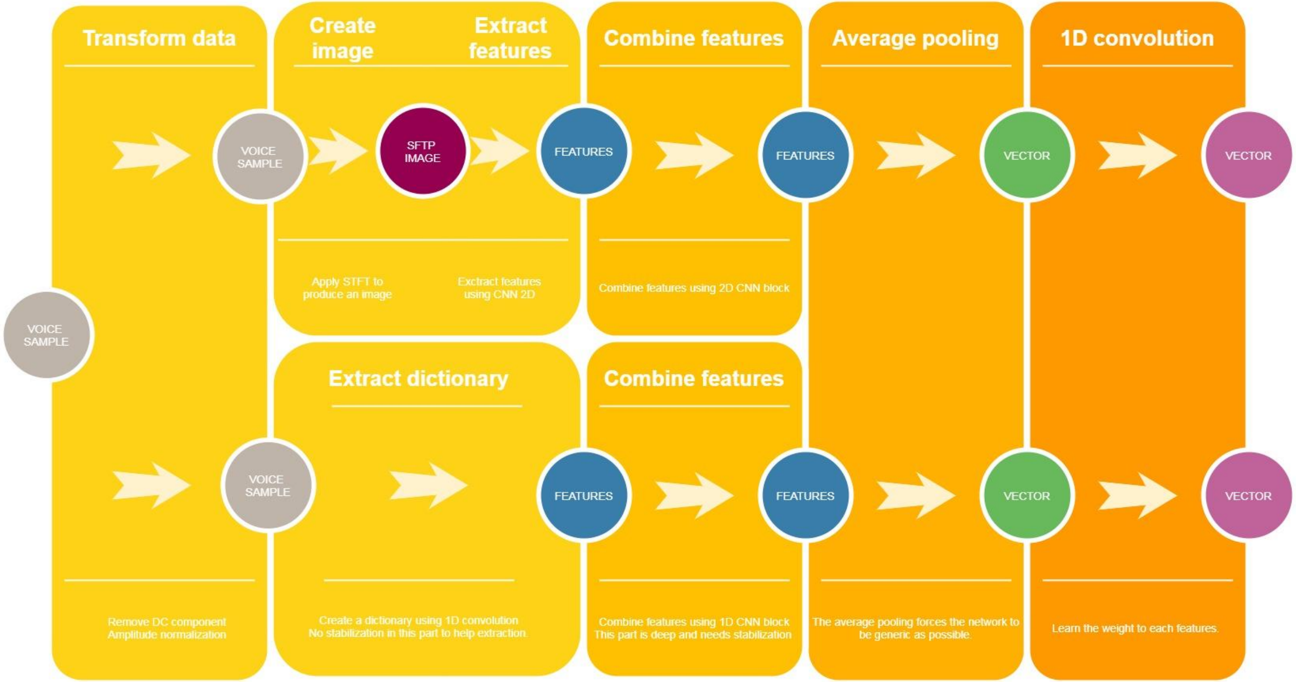
STFT



The encoding

Frequency encoding

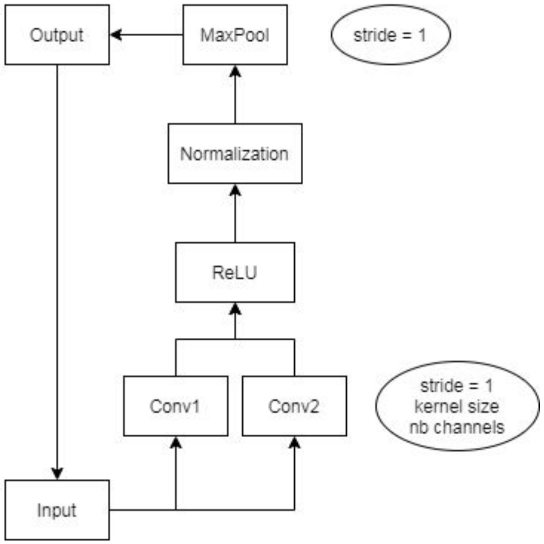
Time encoding



Convolution block

These blocks could stacked to adjust complexity of neural network.
If NN is deep we should activate the inception mechanism.

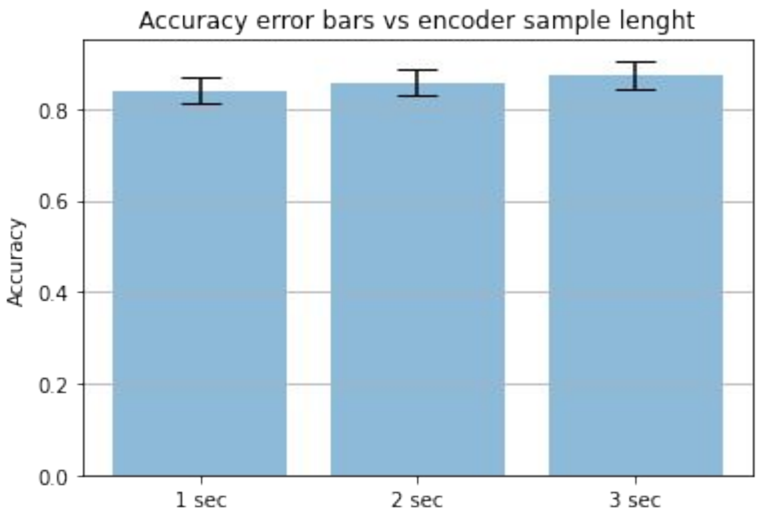
Conv-D block comes from VGG-net architcture.



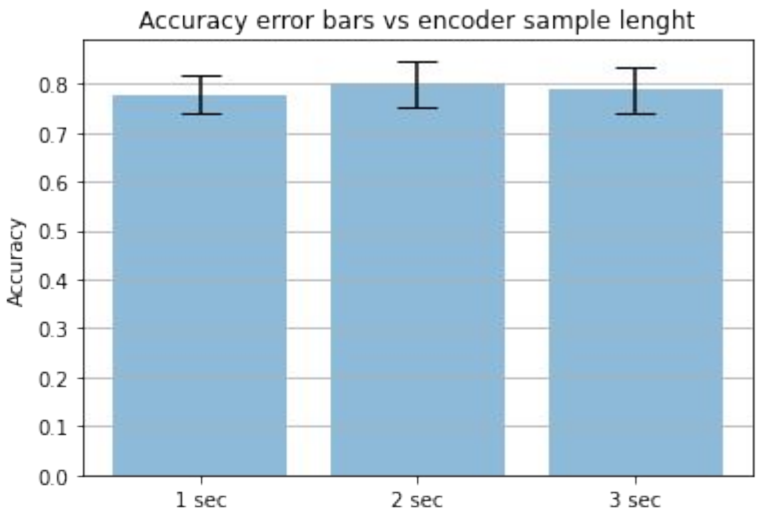
Inception use:
The first convolution have a kernel size equal to 1.
These features are simpler than features generated by a larger kernel size.
This inception stabilize the block.

Encoding results

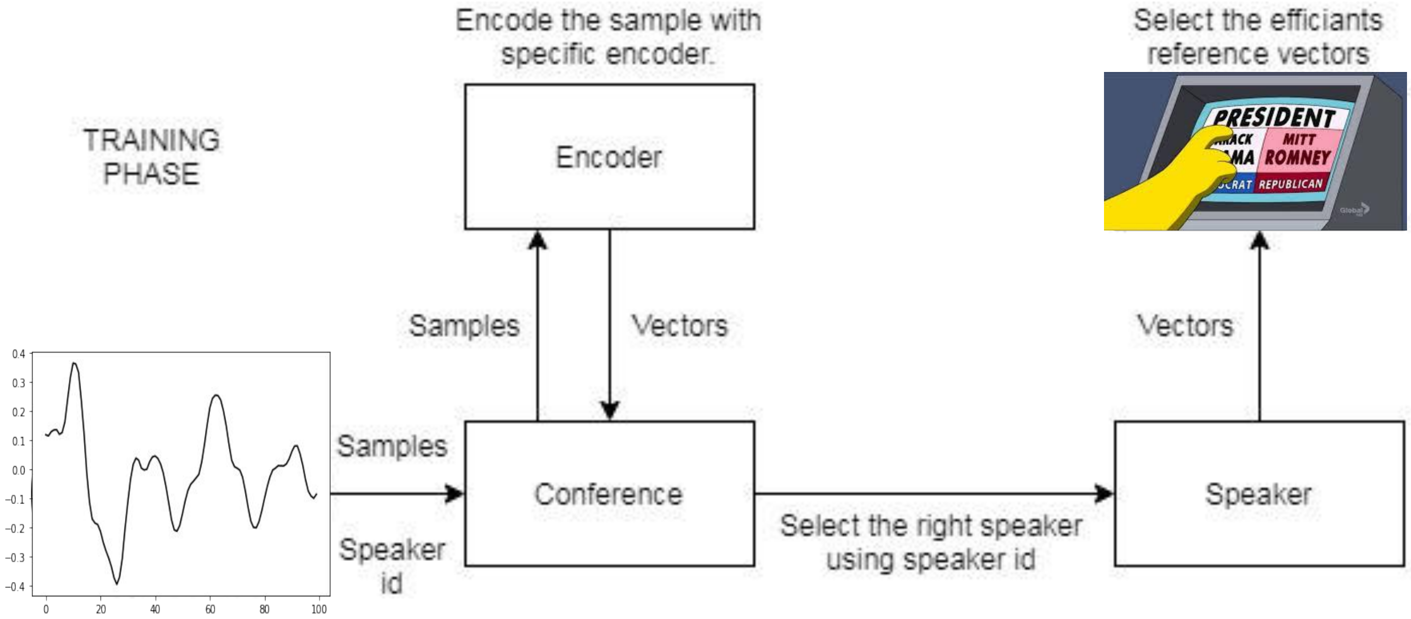
Frequency encoder



Full time encoder

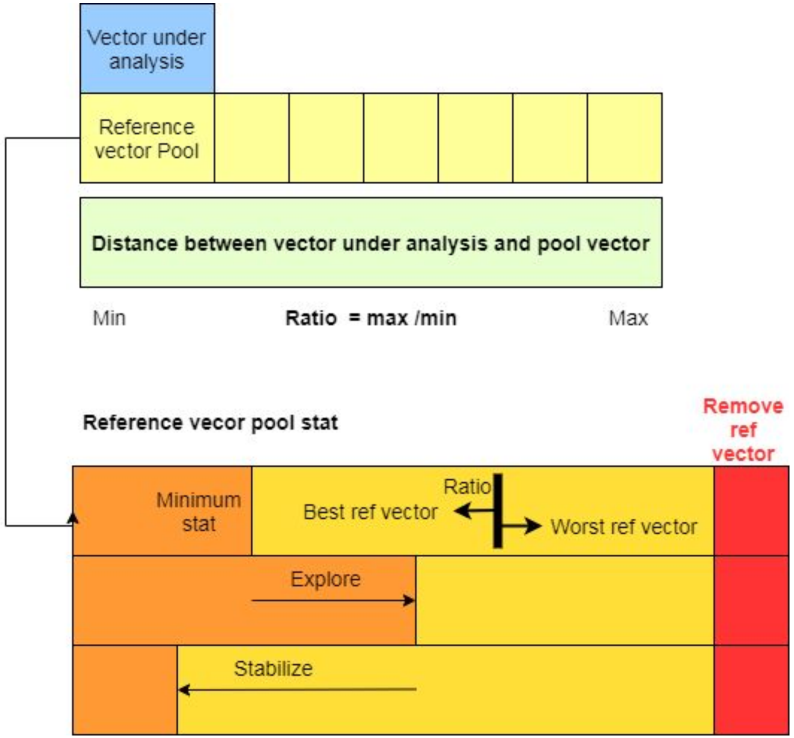


The enrollment

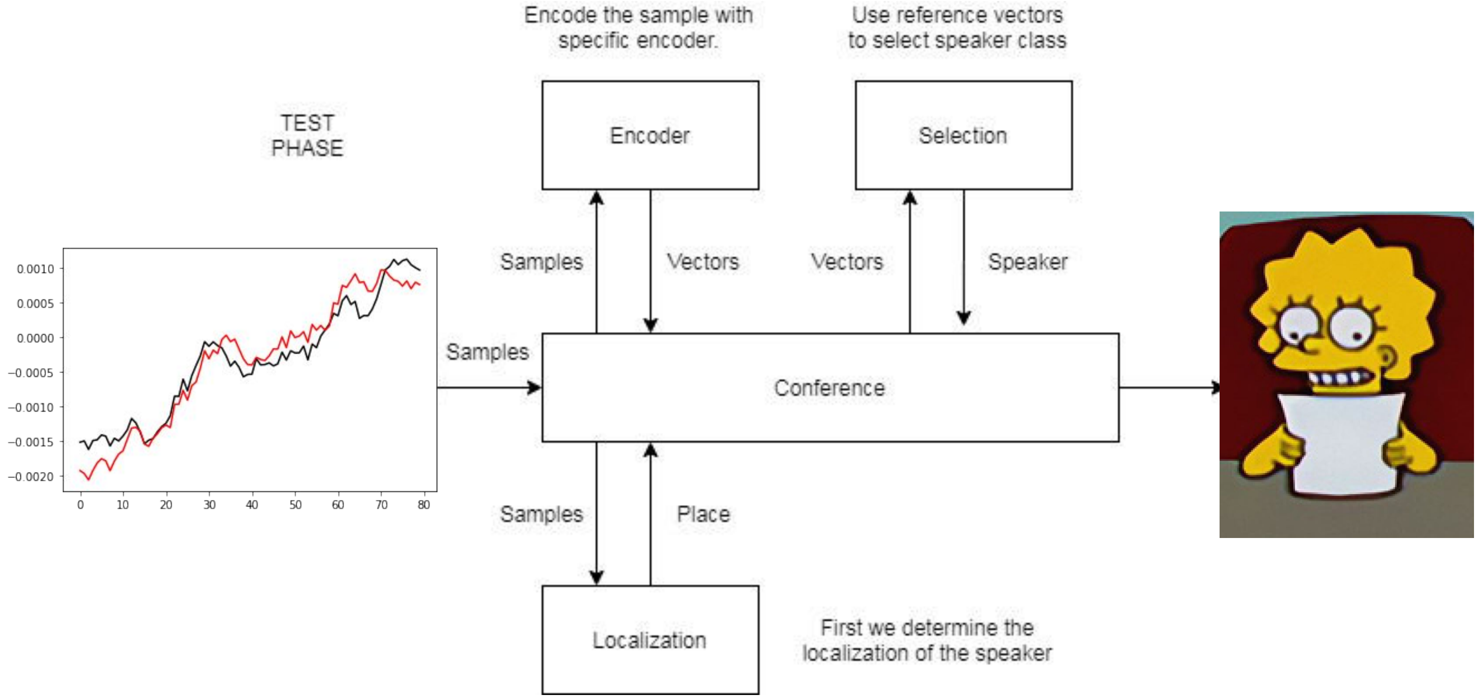


Deeper in election

When stat is close to minimum, the ref vector is in a safe place.
It is not possible to exceed the minimum.
If the ratio is big, the best vector is really good, he moved further away from the red zone.
In red zone the ref vector is removed and replaced by the vector under analysis.
We only remove ref vector if the ratio is more than 2. Don't remove vector if ratio is too small.
One more thing, don't add vector under analysis is they are too close.
When mean label is wrong, we need to more explore.
When mean label is right, we need to stabilize.



The recognition

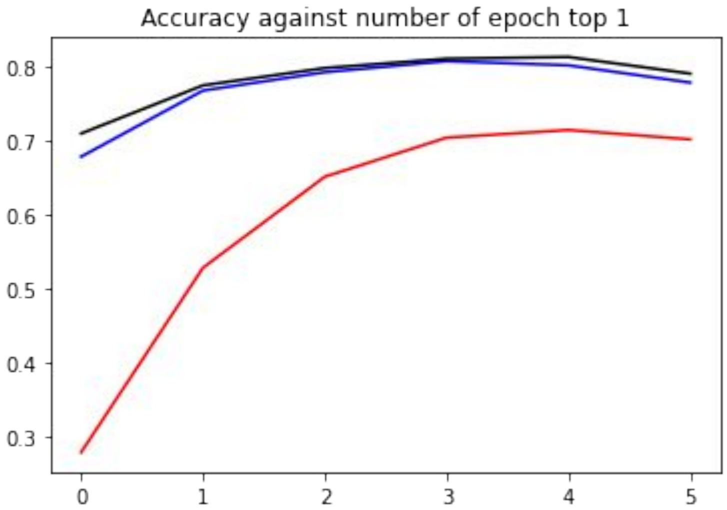


Reference pool: impact of election

Pool size vs accuracy

Pool size	start acc	max acc	
1	0.5037	0.5037	Base line == No election
2	0.5454	0.5543	+0.0089
5	0.6785	0.6968	+0.0183
10	0.7179	0.7896	+0.0717
20	0.7596	0.8603	Max impact of election + 0.1007

The strategy: mean vs top 4 vs best



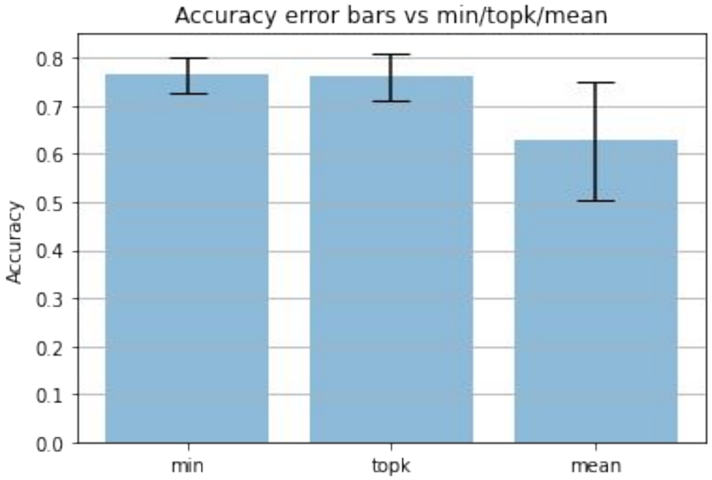
best: OK only best matching is used.

top 4: OK we have enough matching vector.

mean: vectors without matching pattern weighed down the results.

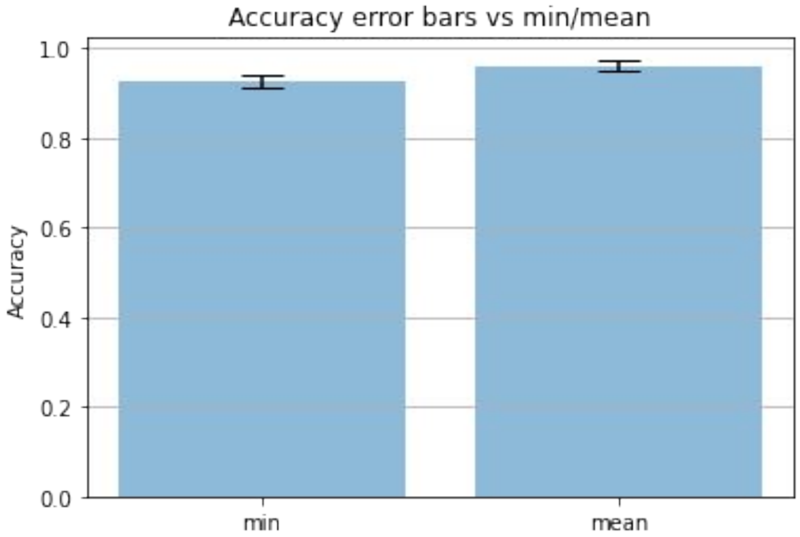
Identification results

Error bar for one sample under analysis



mean acc	topk acc	min acc
0.784	0.853	0.871

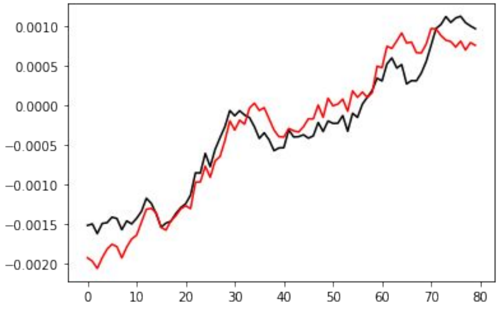
min	mean
0.973	0.988



Error bar for 3 samples under analysis

The localization

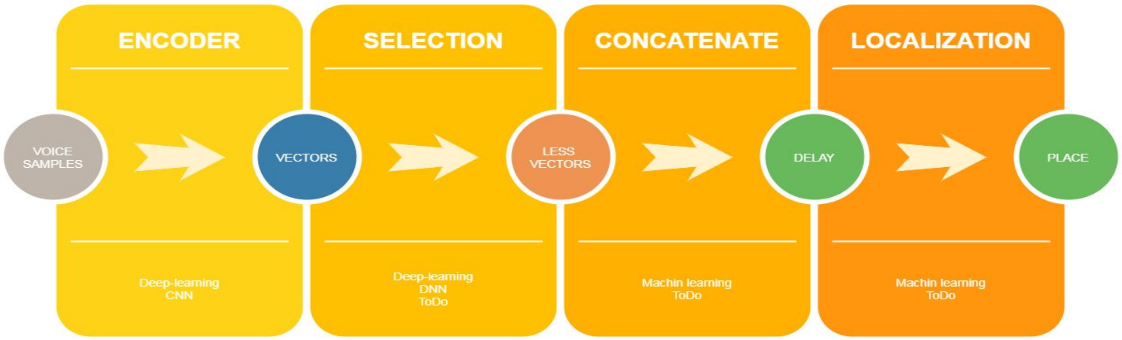
2 mics signal



Encoder accuracy

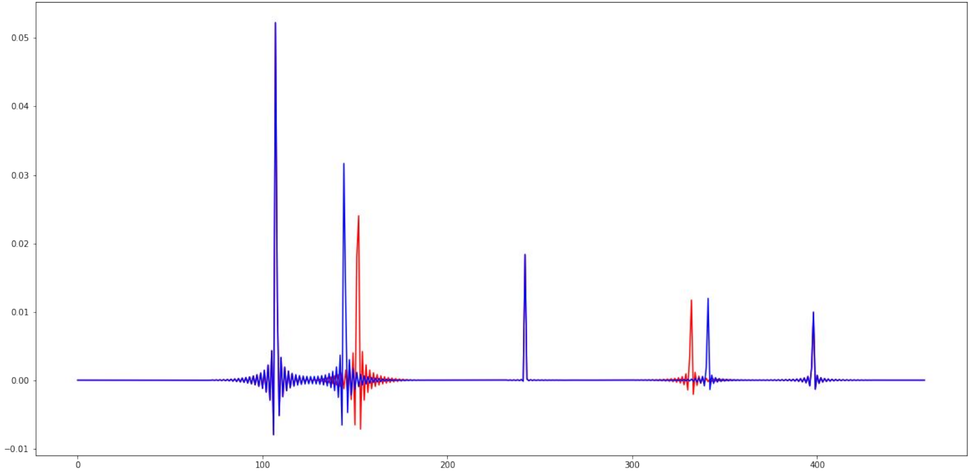
vector length	accuracy
16	96.80%
32	97.75%
64	97.74%
128	98.27%

Global architecture for localization

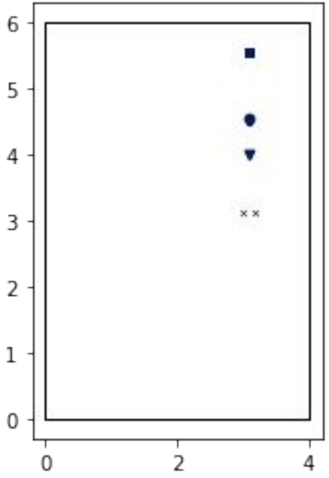


Localization II

Room Imputional Response



Octopus with 3 mics



Room simulation

Result : encoder genericity

Sample size	train acc	test acc
3	0.958	0.886
2	0.941	0.874
1	0.922	0.866

Encoding accuracy

short sample => more genericity

Accuracy for a speaker

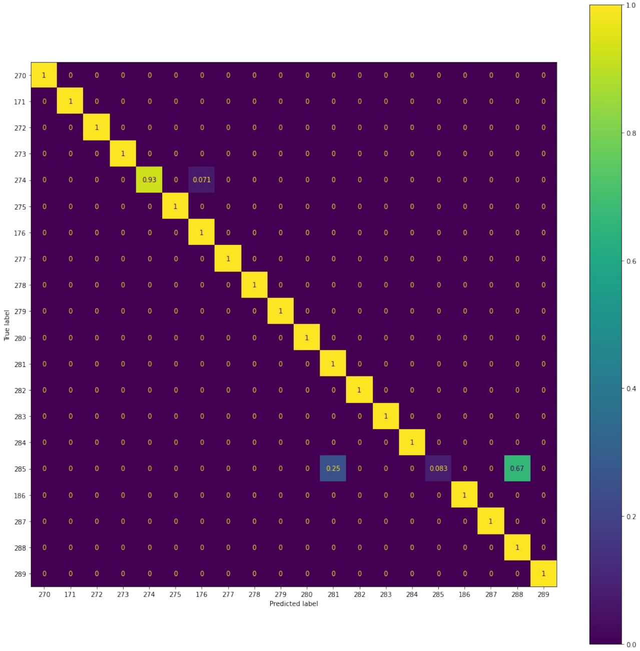
sample sz	3	2	1
topk acc	0.825	0.813	0.853

Learning speed of encoding

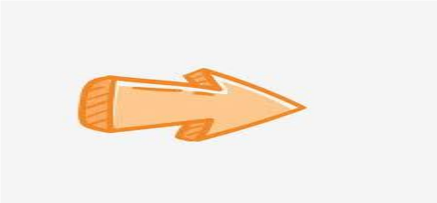
Slow learning => Deep learning

Sample size	best test acc	best epoch	0.86 epoch
3	0.886	25	5
2	0.877	21	8
1	0.863	32	32

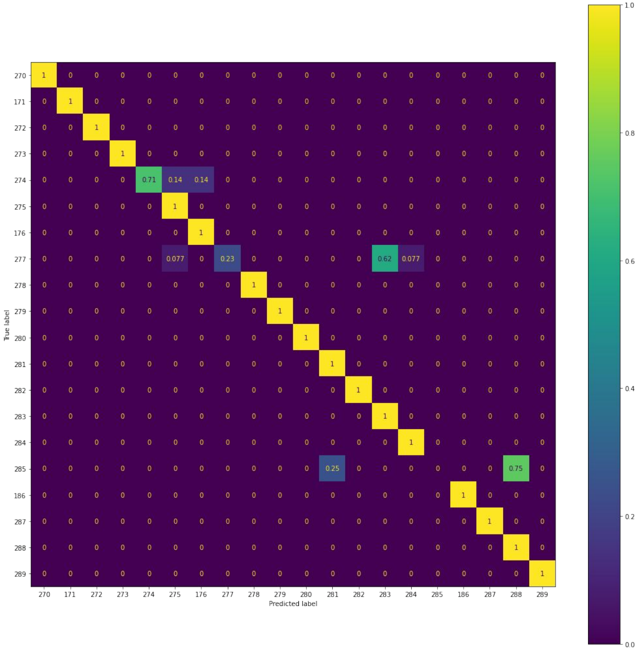
Results : confusion matrix



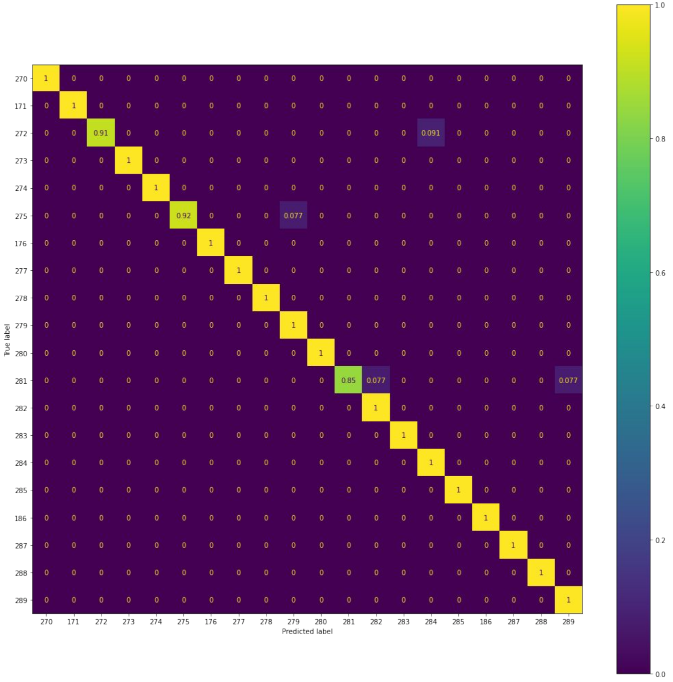
Confusion matrix:
20 speakers
1 sample of 3 seconds
See speakers 285



Election

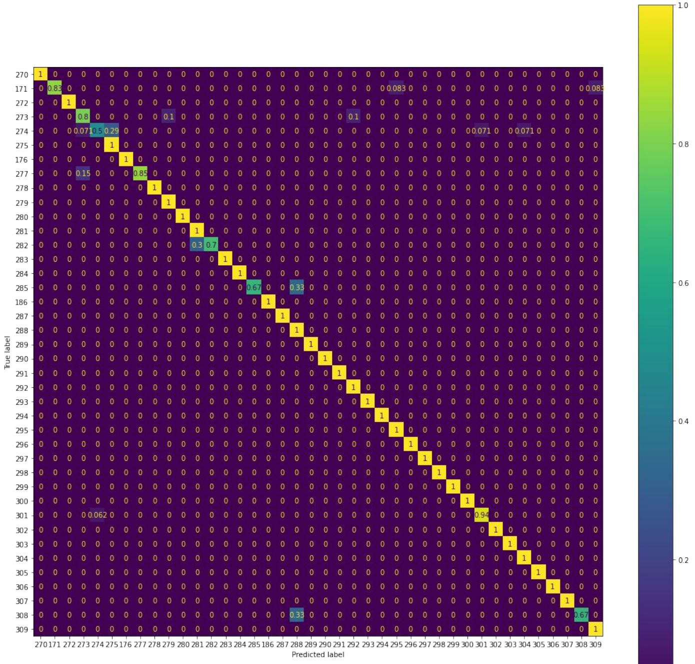


20 speakers vs 40 speakers

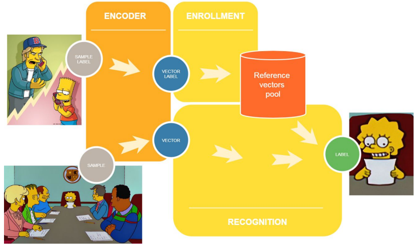


Confusion matrix:
After 3 samples of 3 seconds

For 20 speakers:
best accuracy = 0.988
For 40 speakers:
best accuracy = 0.951



Conclusion



specific encoder
point of interest selection
concatenation of results

Deep network
are lazy

Mixing deep learning
and ML works

N'oubliez pas de vous abonner a ma chaine.

OR

OR