Automatic speech recognition (ASR)

Duration: 3/4 months

Where: Italian Institute of Technology (IIT), Center of Translational neurophysiology for Speech and Communication (CTNSC@UniFe) and University of Ferrara, Section of Human Physiology.

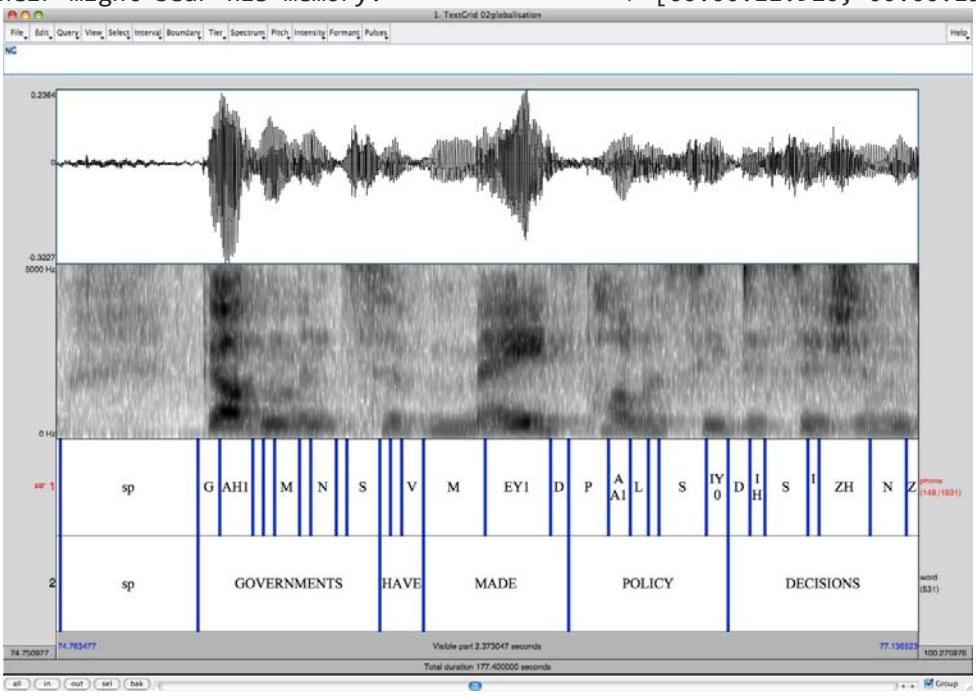
The student will be involved in a larger project related to the development of an innovative ASR system, named neuro-driven ASR. The student will be trained and will work under the direct supervision of expert researcher in the ASR field.

The problem: To build an ASR system the first thing we need is a structured database. This database is built with audio and text files via **forced alignment**. Given an audio file containing speech, and the corresponding transcript, computing a **forced alignment** is the process of determining, for each fragment of the transcript, the **time interval** (in the audio file) containing the spoken text of the fragment. A text fragment can have arbitrary granularity:

- a paragraph,
- a sentence,
- a portion of a sentence (i.e., a group of words),
- a word, or
- a phoneme (i.e., a single sound).

For example, given a text file and an audio file, a force alignment at verse-level can be the following:

```
1                                                   => [00:00:00.000, 00:00:02.640]
From fairest creatures we desire increase,          => [00:00:02.640, 00:00:05.880]
That thereby beauty's rose might never die,         => [00:00:05.880, 00:00:09.240]
But as the riper should by time decease,            => [00:00:09.240, 00:00:11.920]
His tender heir might bear his memory:              => [00:00:11.920, 00:00:15.280]
```



Typical applications of forced alignment include Audio-eBooks, closed captioning, and automating the creation of training data for automated speech recognition systems. The student will be trained to use the Kaldi ASR toolkit.
Knowledge in: basic Linux scripting (bash), python