# Automatic Voice Cloning Across Languages

**Subject description**: We aim to develop a framework that, given an audio segment of speech in a same language from different speakers, a transcript of this segment and a translation of this transcript in a target language, is able to regenerate the audio segment as if each speaker was talking in the target language. This task is to be performed online by the framework but is expected to output an audio of better quality given a longer delay from the source, such that the best performance is obtained when running offline. The quality of an audio sample transferred in a different language relies on the naturalness of the output voice and a meaningful transfer of the speaker features (e.g. someone speaking slowly in French is expected to also speak slowly in English).

**Overview of the implementation**: The core of the project is the voice synthesizer. It will initially be implemented as faithfully as possible to [1]. The framework used for the implementation is not known as of yet. The architecture in [1] has no open-source implementation available yet, but some parts of the architecture do. We will likely adapt some of these implementations to our project. Code to download datasets and to train the models will be made available on the repository.

Only once the synthesizer is operational will we consider improving it so that it can better transfer voice across languages. The model will be integrated along with a speaker diarization model in order to perform automatic voice cloning. Possibly, this diarization model could be derived from the speaker verification model in [1]. This part of the project is considered minor and will only be carried out if the voice synthesizer is properly working.

Eventually, another minor effort of development will be made to create a graphical interface that allows a user to input an audio/video file with subtitles in both languages and that will output the audio segment transfered in the target language. The application would be open-source. This last part of the project is also optional.

# References

[1] Ye Jia, Yu Zhang, Ron J. Weiss et al. **Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis**. June 2018.
`https://arxiv.org/abs/1806.04558`