This is a draft for both the title of the thesis and its formal subject description.

Automatic voice cloning across languages

Short 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 were 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 with 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).

Formal description: Let an audio segment $a_s \in A$, a set of speakers from this segment S_a, a source language l_s and a target language l_t. Let $t(a_s)$ be the transcript of a_s, i.e.

$$t : A \to (S_a \times P_s)^n$$

Where P_s is the set of all sentences in the source language and n is the total number of lines said in a_s. We define a synthesis function $f : (S_a \times P_s)^n \to A$ such that $f(T(a_s)) \approx a_s$.