Real-Time Voice Cloning

We developed a three-stage deep learning framework that performs voice cloning in real-time. This framework is the result of a 2018 paper from Google for which there existed no public implementation before ours. From an utterance of speech of only 5 seconds, the framework is able to capture in a digital format a meaningful representation of the voice spoken. Given a text prompt, it is able to perform text-to-speech using any voice extracted by this process. We reproduced each of the three stages of the framework with our own implementations or open-source ones. We implemented efficient deep learning models and adequate data preprocessing pipelines. We trained these models for weeks or months on large datasets of tens of thousands of hours of speech from several thousands of speakers. We analyzed their capabilities and their drawbacks. We focused on making this framework operate in real-time, that is, to make it possible to capture a voice and generate speech in less time than the duration of the generated speech. The framework is able to clone voices it has never heard during training, and to generate speech from text it has never seen. We made our code and pretrained models public, in addition to developing a graphical interface to the framework, so that it is accessible even for users unfamiliar with deep learning.

Author: Corentin Jemine *(C.S. bachelor 2014-2017, master in Data Science 2017-2019)*)

Supervisor: Prof. Gilles Louppe

Academic year: 2018-2019