Robust Generative AI pipeline for voice generation.

D1.1 progress submission

Introduction.

The generation of voice is a highly complex problem that is deeply intertwined with the sequential processing of time series data. Voice generation, also known as speech synthesis, involves the creation of human-like speech from textual input. This field encompasses various subfields such as text-to-speech (TTS) synthesis and voice cloning, which have wide-ranging applications in industries like entertainment, telecommunications, and assistive technology.

Currently, the existing models for voice generation are plagued by their lack of modularity and excessive complexity, making them difficult to understand and extend. As a result, researchers and practitioners in the field are faced with the challenging task of investigating and exploring alternative approaches and solutions. In this research project, we aim to address this problem by thoroughly examining existing methodologies and developing a modular pipeline.

Technologies.

In order to tackle this endeavor, we have chosen Python as our primary programming language. Additionally, we will be utilizing the PyTorch framework for machine learning, which has gained significant popularity in recent years due to its flexibility, ease of use, and excellent support for building and training neural networks.

While the specific open source models to be employed in this project are yet to be determined, we have identified several promising candidates for further research and evaluation. These models include TorToeSe, a state-of-the-art text-to-speech (TTS) system that utilizes duration modeling and spectral features for more natural-sounding speech synthesis. Another model worth investigating is so-vits-svc, which combines TTS and speech vector concatenation techniques to generate high-quality voice samples. Additionally, we will explore the possibilities offered by the HUBERT model, a self-supervised representation learning framework that can be adapted for voice generation. Finally, the RVC model has already been launched and initial results have been obtained, prompting further investigation and refinement.

Data.

One advantage of generative models for voice generation is that they do not necessarily require vast amounts of data for training purposes. In fact, existing research suggests that training a voice generation model can be accomplished with as little as three minutes of speech data, while more extensive training sets of up to 10 hours can yield even better results. To ensure the availability of suitable and diverse training data, we will curate and collect our own dataset, thereby enabling us to have full control over the quality and characteristics of the information utilized for training our models.

Current status.

At this stage, our research efforts are concentrated on carefully studying the architectures and methodologies of existing voice generation models. This preliminary analysis will serve as the basis for estimating the scope of work required to develop our modular pipeline. Additionally, we have already initiated the implementation of the RVC model and have made significant progress in assessing its performance and identifying potential areas for improvement.

Moreover, we are actively engaged in the task of integrating various architectural components from different models to achieve enhanced results. This entails investigating the compatibility of different modules, identifying potential synergies, and fine-tuning the interplay between these building blocks to optimize the overall performance of our voice generation system.

Acknowledgements and Team.

Our research team comprises three members:

- Leon Parepko (l.parepko@innopolis.university) - Advanced Machine Learning, scientific research.
- Polina Lesak (p.lesak@innopolis.university) - Machine Learning engineer
- Darina Merzakreeva (d.merzakreeva@innopolis.university) - Machine Learning engineer

For additional information about our work, please visit: https://github.com/Leon-Parepko/Audio-Generation