Robust Generative AI pipeline for voice generation.

Leon Parepko (l.parepko@innopolis.university), Ayaz Sunagatulin (a.sunagatullin@innopolis.university)

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.

Current status.

The current stage of our work involves the comprehensive analysis of existing methodologies, wherein we have successfully examined and assessed various approaches that could potentially enhance the performance of our model. In order to expand our understanding and ensure a well-rounded perspective, we have specifically chosen to base our research on three meticulously selected scholarly articles. These articles, relevant to our research objectives, have been cataloged and made readily accessible via our dedicated repository on GitHub under the "Articles" folder.

The original RVC model has been successfully launched and tested locally. In order to enhance the performance of the model, we are currently focusing on three key aspects: improving data preprocessing, enhancing the model itself, and developing a postprocessing filtering model.

- Firstly, we aim to enhance the data preprocessing stage by refining the slicing algorithm used to segment the audio signal during training. The goal is to extract speech fragments more accurately and reliably, which would result in better performance during inference mode. This analytical approach will allow for the retrieval and combination of information from random slices to assemble a new and improved audio output.
- Secondly, we are exploring ways to improve the model architecture. Through an initial analysis of the RVC architecture, we have identified that increase of dimensionality of the conditional VAE component of the VITS model, which serves as the basis for RVC,

may yield positive results. However, due to the complexity of the reverse engineering process, we have not yet achieved significant advancements in this area.

- Lastly, we are in the process of developing a postprocessing filtering model. This model will take input from the RVC and focus on increasing the overall quality of the output. By applying advanced filtering techniques, we aim to refine the audio output and improve its clarity and coherence.

  Moving forward, our plan is to finalize the aforementioned aspects and integrate them into a unified and comprehensive model. By combining these improvements, we anticipate achieving enhanced performance and overall effectiveness in the RVC model.

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