Robust Generative AI pipeline for voice generation.

D1.3 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.

Current status.

The original RVC model has been effectively deployed and examined at the local level. To facilitate smoother operations with the model, a custom wrapper has been developed specifically for the launch and inference of the RVC. Moreover, 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, our objective is to optimize the data preprocessing stage by improving the slicing algorithm employed for audio signal segmentation during the model's training phase. The aim is to achieve a higher degree of precision and reliability in extracting speech fragments, thereby enhancing the overall performance during training. This analytical approach will allow for the retrieval and combination of information from random slices to assemble a new and improved audio output.

  Although various permutations of split lengths and shuffling algorithms have been explored, thus far no substantial improvements have been observed. Consequently, a hypothesis has emerged suggesting that overlaying heterogeneous speech segments prior to training may yield more favorable outcomes. This approach intends to leverage the complementary characteristics of different speech fragments, potentially leading to enhanced results.

- Secondly, we are engaged in a comprehensive exploration of potential enhancements to the existing 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. In addition to this, we propose the integration of an intermediate feature extractor, such as an Embedding or LSTM layer, which would intake initial samples as input and supplement the VAE with additional features. At present, we are actively assessing the effectiveness of this approach, albeit grappling with the challenge of selecting a suitable comparison metric for our proposed solution.

- This novel model is designed to receive input directly from the RVC, with the primary objective of enhancing the overall quality of the generated output. Through the implementation of sophisticated filtering techniques, we anticipate refining the audio output to achieve heightened clarity and improved coherence. This postprocessing filtering model represents a crucial step in our ongoing pursuit of optimizing the performance and fidelity of our system.

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.

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