TF-Encodec: High Fidelity Neural Audio Compression using Time-Frequency Representation

Despite impressive recent advances with neural networks, generating the waveform in audio synthesis remains a challenge in machine learning. As a result, recent works typically use specific encoders such as EnCodec to synthesize sound and avoid generating the waveform. Training audio encoders, however, is challenging due to the oscillatory and multi-scale nature of sound, making it data and computation-heavy. In this project, we investigate an alternative technique to build an encoder that requires less memory and computation.

Instead of generating the waveform directly, we propose generating a Short Time Fourier Transform (STFT) representation of the waveform. The STFT magnitude is a much more stable sound representation that does not oscillate in time. However, previous attempts to produce high-quality audio using the STFT often failed because of the difficulty to recover the phase from the STFT magnitude. Therefore, additionally to the magnitude, we suggest producing the phase gradients with the decoder. Then, the waveform can be recovered using phase integration and STFT inversion.

Since EnCodec uses a discriminator based on the STFT, we believe the computation will be significantly reduced compared to the original EnCodec, making it possible to train with a smaller dataset and less compute while still achieving similar quality.

**Requirements:** Ability to work independently and determined to obtain results, creative thinking, knowledge with Machine Learning and Python.

We will have weekly meetings to address questions, discuss progress and think about future ideas.

**Contact**
- Luca Lanzendörfer : lanzendoerfer@ethz.ch, ETZ G93
- Nathanaël Perraudin : nathanael.perraudin@sdsc.ethz.ch