Transformer-Based Audio Generation Conditioned by 2D Latent Maps: A Demonstration

Christian Limberg, Zhe Zhang, Marc A. Kastner
National Institute of Informatics (NII), 101-8430 Chiyoda-ku, Tokyo, Japan
Technische Universität Berlin, Straße des 17. Juni, 10623 Berlin, Germany

Web-App

Abstract

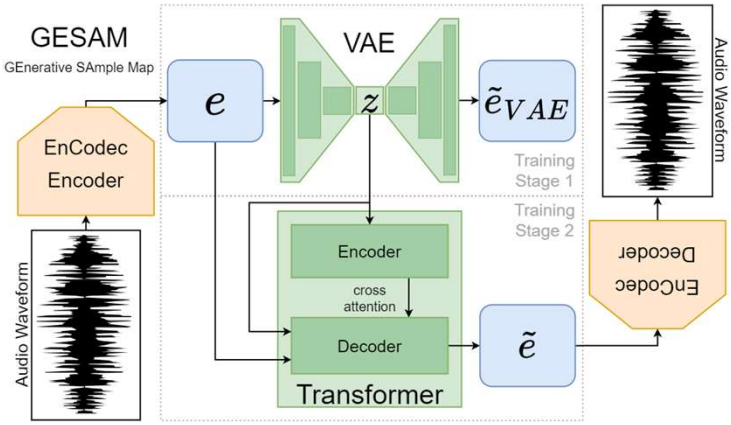
This paper presents a demonstration of an improved framework for audio sample generation using interactive 2D latent maps. Building upon the foundational work *Mapping the Audio Landscape for Innovative Music Sample Generation*, we enhance the framework by introducing visualization techniques for exploring the 2D audio landscape through different audio features such as energy and bandwidth. Additionally, we train a t-SNE embedding over these features to create a more abstract visualization of the audio samples on the map. This demo also significantly improves usability and user interactivity, allowing for a more intuitive and efficient exploration of the generated audio samples. The demo showcases these improvements in real-time, providing users with an enhanced novel interface for generating high-quality audio samples.
Demo URL: https://limchr.github.io/gesam_demo/

Objective

- Introduced a framework to generate high-quality audio samples for music production by leveraging a 2D latent map.
- Aimed to simplify the process of exploring audio samples through intuitive visual interfaces, catering to musicians and sound designers.

Methodology

- Utilized Variational Autoencoders (VAEs) to encode audio samples into a 2D latent space. (Stage 1)
- Trained a Transformer model to generate audio conditioned by the 2D latent map, enabling seamless transitions across the audio landscape. (Stage 2)



Innovations

- Proposed the use of a visually navigable 2D latent space to represent audio features, allowing users to explore audio intuitively.
- Combined unsupervised representation learning (via VAEs) with powerful generative capabilities (via Transformers).

Enhanced Visualizations

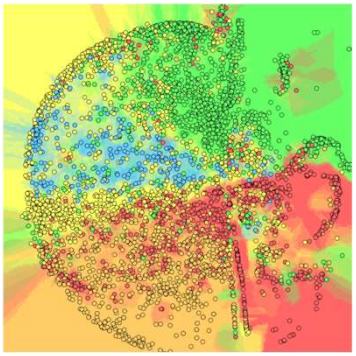
For our interactive web-based application showcasing our approach, we generated a grid of 200x200 audio samples using our trained model. To create a user-friendly interface, we visualize the rendered audio samples using three distinct approaches.

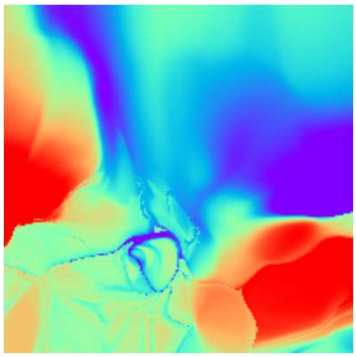
Classifier Map (left image)

We trained a classifier that can distinguishes the classes we trained the transformer model with. We color the embedding space with the corresponding class color by using a 5 nearest neighbor classifier, where we mix the colors of the 5 nearest neighbors for creating a visual transition between classes.

Audio Feature Maps (right image)

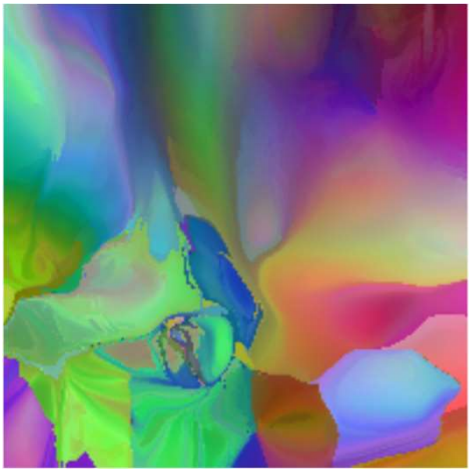
We computed 5 different audio features (like energy, bandwidth, ...) and use a color map for visualizing each one as a background image for orientation.





T-SNE Map

Based on the 5 audio features we trained a 3D T-SNE embedding which we map to RGB space for visualizing a richer, but more abstract map of the sound scape.



Acknowledgements

This work was supported by a fellowship within the IFI program of the German Academic Exchange Service (DAAD).

TECHNISCHE UNIVERSITÄT BERLIN

Inter-University Research Institute Corporation
Research Organization of Information and Systems
National Institute of Informatics

DAAD Deutscher Akademischer Austauschdienst
German Academic Exchange Service

Christian Limberg
cnlimberg@gmail.com