Timbre Latent Space: Exploration and Creative aspects

Antoine CAILLON, Adrien BITTON, Brice GATINET, Philippe ESLING

Institut de Recherche et Coordination Acoustique Musique (IRCAM) Sorbonne Université - CNRS UMR 9912 - 1, Place Igor Stravinsky, F-75004 Paris

Models

We propose two models based on the Variational Auto-Encoder (VAE) [1], which learn a latent representation of an audio dataset by jointly optimizing two functions used to **analyse** (encoding) and **synthesize** (decoding) audio.

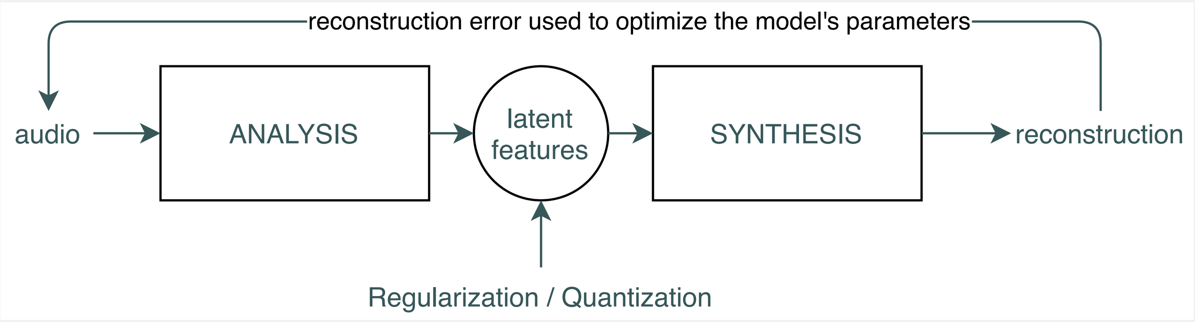

Figure: Overall architecture shared by both models

This **invertible** representation is generally of **lower dimensionality** than audio, but its use as a synthesis tool in a creative process remains complicated. In this work we explore interactions either based on a **continuous** latent representation or a **discrete** set of latent features.

Continuous latent space

The continuous model is composed of two main blocks: a mel-spectrogram VAE and a mel-spectrogram to waveform model [2]. We train the VAE using an objective composed of a **reconstruction loss** and a **regularization loss**, itself being the addition of a prior regularization and a domain adaptation loss [3]. The obtained regularization loss ensures that the latent space is smooth and loudness invariant.

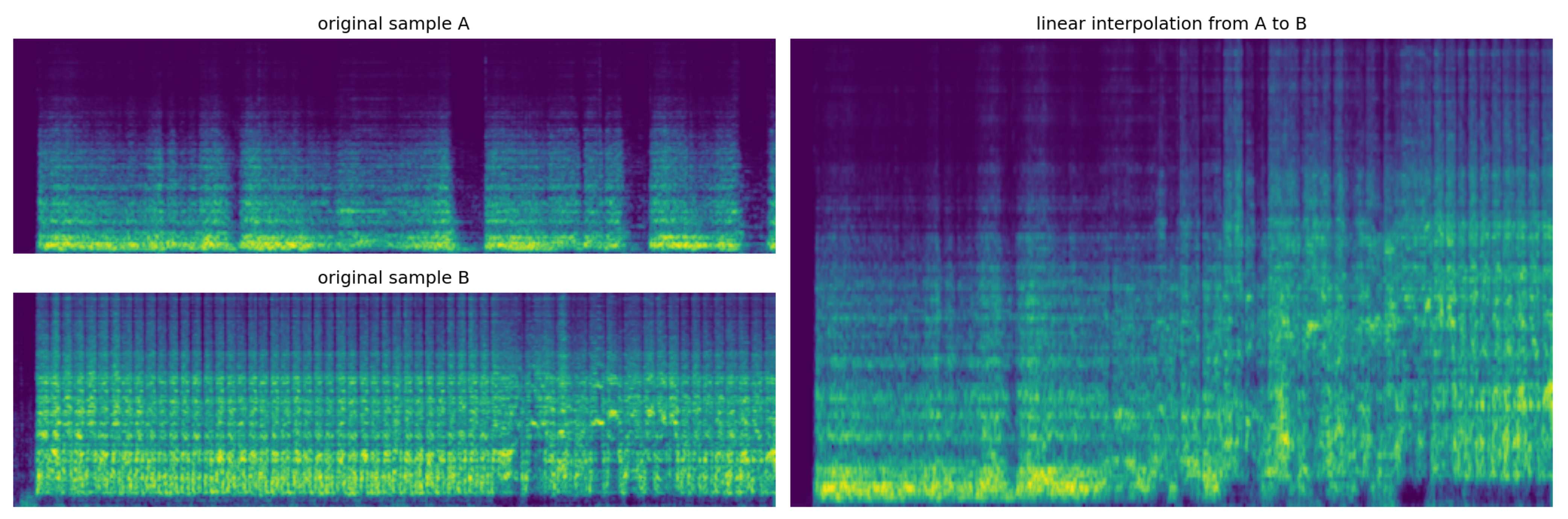

Figure: Time linear interpolation between two audio samples

Offline generation

Max/MSP interface designed to help the process of encoding and decoding audio. We added several tools like **manual deformation of latent series** and an **interpolation plane**.

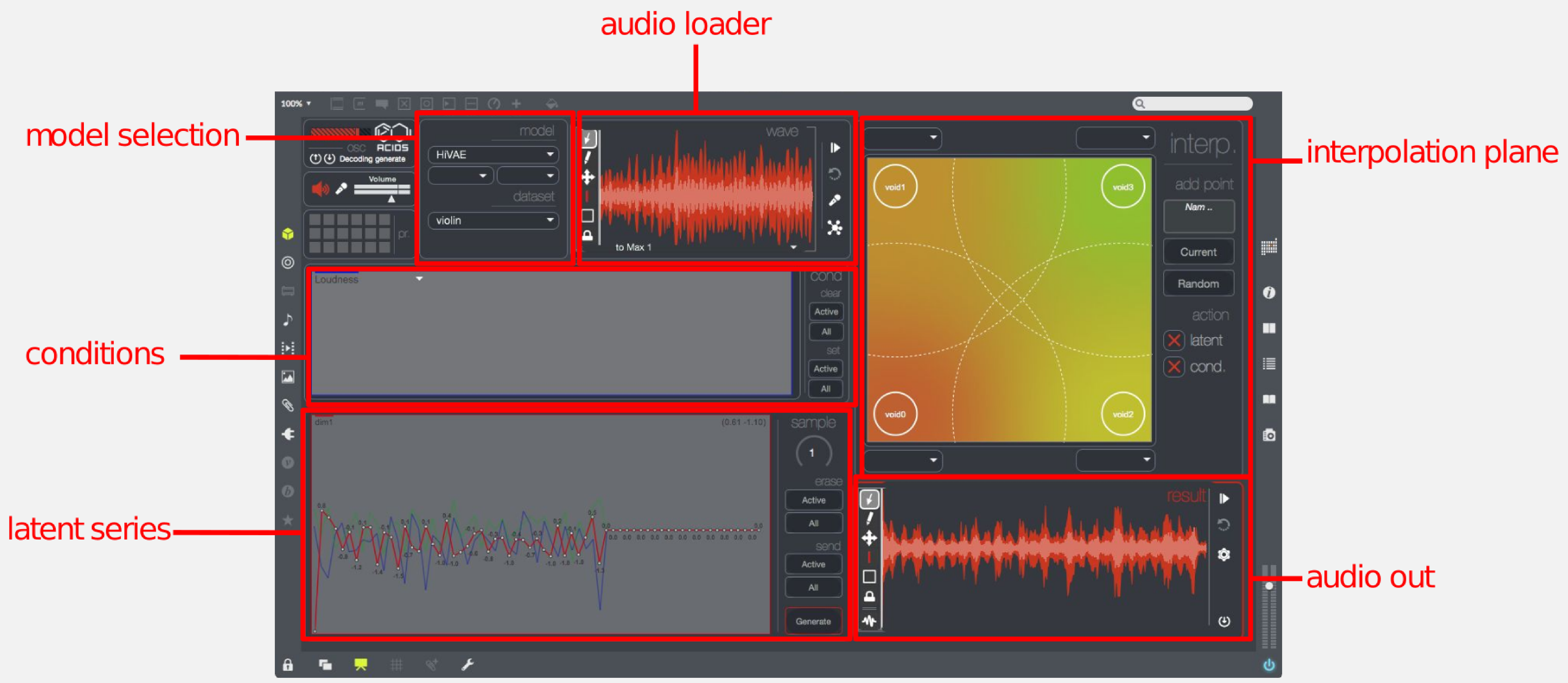

Figure: Max / MSP interface for offline generation

Discrete latent space

The discrete model is based on a Vector-Quantized VAE [4] for frame-wise processing of raw waveform. Each signal window is **analysed** and **quantized** with the nearest latent vector, also invariant to audio levels. Once trained, we can analyse each individual latent feature and compute some corresponding acoustic descriptor values. This provides a **mapping** that allows direct **descriptor-based** synthesis, by matching a given descriptor target with the series of nearest latent features.

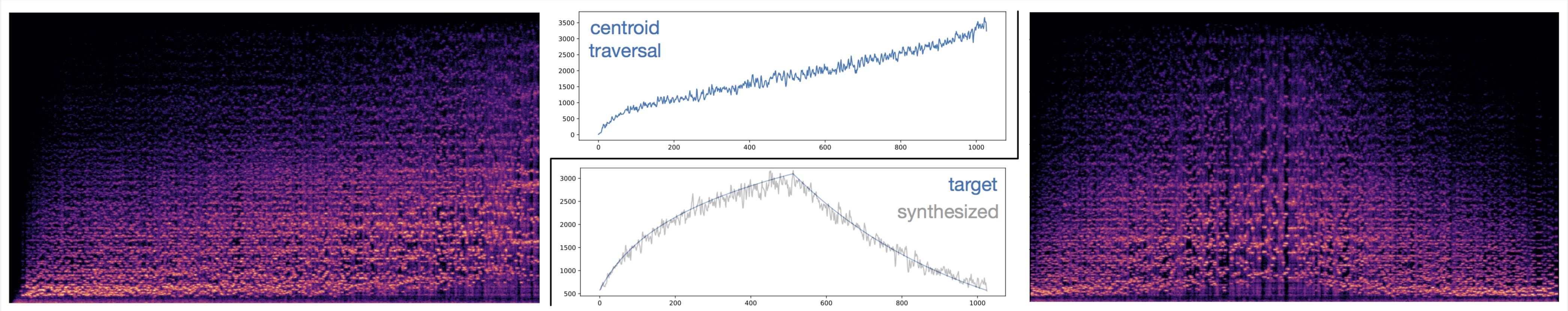

Figure: Left: traversal of the discrete representation in the increasing order of the spectral centroid. Right: Example of descriptor-based synthesis.

Online generation

In order to allow a **realtime interaction** with the model, we abstracted the encoder-decoder pair as PureData signal objects, allowing their use inside a complex composition workflow.

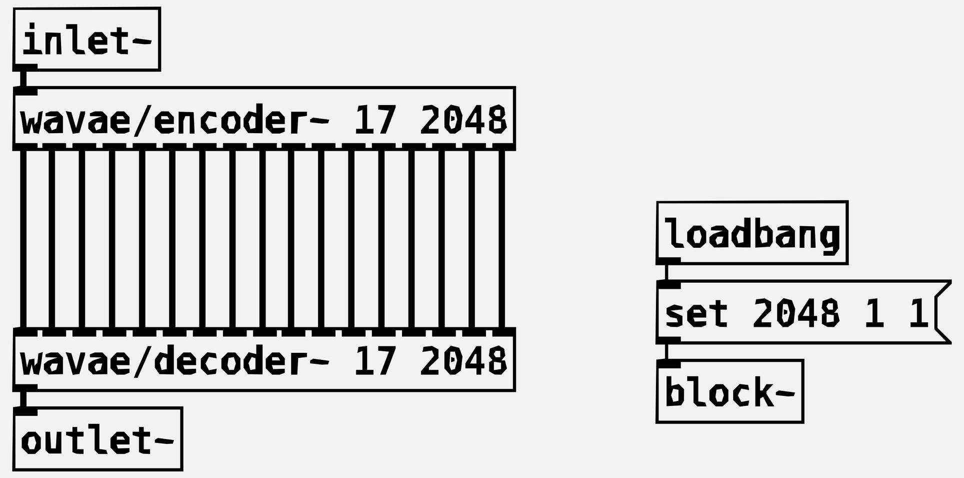

Figure: PureData encoder / decoder objects

Visit our website !

References

[1] Diederik P. Kingma et al. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations*. 2014.

[2] Kundan Kumar et al. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis". In: *Advances in Neural Information Processing Systems 32*. 2019.

[3] Yaroslav Ganin et al. "Unsupervised Domain Adaptation by Backpropagation". In: vol. 37. Proceedings of Machine Learning Research. 2015.

[4] Aaron van den Oord et al. "Neural Discrete Representation Learning". In: *Advances in Neural Information Processing Systems 30*. 2017.

Contact Information

- Web: https://acids-ircam.github.io/timbre_exploration/
- Email: {caillon, bitton, gatinet, esling}@ircam.fr