# TIMBRE LATENT SPACE: EXPLORATION AND CREATIVE ASPECTS

Antoine CAILLON, Adrien BITTON, Brice GATINET
Institut de Recherche et Coordination Acoustique Musique (IRCAM)
UPMC - CNRS UMR 9912 - 1, Place Igor Stravinsky, F-75004 Paris
{caillon, bitton, gatinet}@ircam.fr

## Introduction

Recent studies show the ability of unsupervised models to learn invertible audio representations using Auto-Encoders **[1]**. They enable high-quality sound synthesis but a limited control since the latent spaces do not disentangle *timbre* properties. The emergence of disentangled representations was studied in Variational Auto-Encoders (VAEs) **[2, 3]**, and has been applied to audio. Using an additional perceptual regularization **[4]** can align such latent representation with the previously established multi-dimensional timbre spaces, while allowing continuous inference and synthesis. Alternatively, some specific sound attributes can be learned as control variables **[5]** while unsupervised dimensions account for the remaining features. New possibilities for timbre manipulations are enabled with generative neural networks, although the exploration and the creative use of their representations remain little. The following experiments are led in cooperation with two composers and propose new creative directions to explore latent sound synthesis of musical timbres, using *specifically designed interfaces* (Max/MSP, Pure Data) or mappings for *descriptor-based synthesis*.

## Method

We introduce two models for learning the generative representations of timbre. One approach leads us to continuous latent sound manipulations, and another to discrete mappings with acoustic descriptors.

*Continuous latent representation of musical timbre.* We use a model based on a regular VAE to learn a latent audio representation that we regularize using both a continuous prior distribution and an adversarial objective. It allows us to explicitly control pre-defined acoustic descriptors such as loudness or spectral centroid during generation. We use an invertible mapping obtained from a Principal Component Analysis (PCA) to project the latent points on a basis that minimizes interdimensional covariance, easing the manipulation of individual dimensions.

*Discrete latent representation of musical timbre.* A variant of the VAE learns a discrete representation by projecting continuous latent codes into a Vector Quantization **[6]** space. We train the VQ-VAE with a disentangled gain in order to decompose the spectral distribution of an instrument into a discrete set of latent codes. As the decoder optimizes reconstruction solely based on these loudness-invariant features, we individually decode each codebook element and compute any corresponding signal descriptors. This enables a direct mapping between acoustic descriptors and the discrete latent representation, performing descriptor-based synthesis and visualizations.

## Results

Unsupervised latent audio representations remain highly-dimensional and only partly disentangled. As a result, sound synthesis and timbre manipulations are tedious and unsuited to creative purposes. One of our contributions is to develop interfaces that facilitate exploring the vast amount of sound variations embedded in a continuous latent space. Another is to learn a discrete timbre representation which eases visualization and control with acoustic descriptors. These tools for interaction and rendering enable a

better understanding of the learned representations, and can be used for *composition* and *generative audio*. Moreover, these new timbre transformations could be useful for perceptual timbre studies.

*Reactive latent audio generation.* This project is based on real-time patches that expose a series of latent features inferred from an input audio signal, allowing the user to perform various mathematical operations before reconstructing an output audio signal. Models were trained using datasets of 1) violin and cello improvisations, 2) classical and popular voices and 3) instruments of the orchestra. Akin to audio effects, they allow modifying the timbre of an individual sound. Moreover, continuous interpolations can be performed across multiple sounds. Some preliminary results and use in composition are shown in the supporting page[1] and repository.

*Analysis and control with audio descriptors.* The VQ-VAE trains on a single instrument dataset (e.g. solo performance recordings) and learns the corresponding latent timbre decomposition. The discrete representation can be mapped to any signal descriptors, such as fundamental frequency or spectral centroid. It allows direct synthesis from an acoustic target, for instance a user-defined pattern or a modulation signal. We can visualize these latent timbre features in a 2-dimensional space and analyze their acoustic distributions. We observe that the unsupervised coordinates only correlate little with local acoustic distances, but they are mapped to consistent spectral distributions and correctly inverted to signal. This analysis is individually performed for several instruments and for the singing voice, with a common set of acoustic descriptors.

## Discussion

This research has studied continuous and discrete latent sound representations as creative tools to explore timbre synthesis. Manipulations are eased by developing specific interfaces and real-time rendering, which greatly enrich composition and sound design possibilities. And in turn, it gives further insights on the generative qualities found in the learned representations, as well as the relevance of their different parameters and controls with respect to the new timbres that are synthesized. When considering a single instrument timbre, the discrete latent space is directly analyzed with any signal descriptors. We can visualize the learned features and their corresponding acoustic distributions. In this setting, timbre synthesis can be performed following various acoustic descriptor targets. This can serve composition, as well as rendering signals with controlled auditory properties which may be useful in order to improve the analysis of timbre perception. Indeed, generative models can morph and create new timbres, by doing so they provide listening stimuli beyond the limits of instrument recorded sound samples. This variety of outcomes will share a path to a better understanding of the efficiency of Variational Auto-Encoders in timbre modelling and perception.

## References

Detailed materials, sounds and complete references on the project page[1] *(until the full-length paper submission)*. **[1]**[Engel et al., 2017], **[2]**[Kingma and Welling, 2014], **[3]**[Higgins et al., 2017], **[4]**[Esling et al., 2018], **[5]**[Bitton et al., 2019], **[6]**[Oord et al., 2018].

[1] please visit https://acids-ircam.github.io/timbre_exploration/