Steering latent audio models through interactive machine learning

Gabriel Vigliensoni[1] **and Rebecca Fiebrink**[2]
[1, 2]Creative Computing Institute, University of the Arts London, UK
[1]Centre for Interdisciplinary Research in Music Media and Technology, QC
g.vigliensoni@arts.ac.uk

Abstract

In this paper, we present a proof-of-concept mechanism for steering latent audio models through interactive machine learning. Our approach involves mapping the human-performance space to the high-dimensional, computer-generated latent space of a neural audio model by utilizing a regressive model learned from a set of demonstrative actions. By implementing this method in ideation, exploration, and sound and music performance we have observed its efficiency, flexibility, and immediacy of control over generative audio processes.

Introduction

Recent advances in neural audio synthesis have made it possible to generate audio signals in real time, enabling the use of applications in musical performance. However, exploring and playing with their high-dimensional spaces remains challenging, as the axes do not necessarily correlate to clear musical labels and may vary from model to model. In this paper, we investigate and propose a useful new approach based on interactive machine learning. This approach allows the performer to map the well-known, low-dimensional, human performance space to the high-dimensional generative audio model's latent space by providing training examples that pair the two spaces.

Background

Generative AI audio models

Generative AI audio models provide a data-driven approach to sound generation. These systems are designed to autonomously generate audio signals by learning from existing or custom datasets, capturing the underlying patterns and characteristics of the input data. However, historical systems for generative audio modelling and synthesis, such as WaveNet (Oord et al. 2016) and SampleRNN (Mehri et al. 2017)), have been challenging to integrate into creative environments due to their large computational complexity, poor signal quality, short temporal coherency, and lack of interaction means. Newer neural audio synthesis architectures and systems such as DDSP (Engel et al. 2020) and Jukebox (Dhariwal et al. 2020) have introduced advancements that addressed part of the previously mentioned issues. DDSP can model audio signals using small training datasets and can be steered in real time using pitch and amplitude as generative conditions, but only for monophonic instrument signals. Jukebox can generate a singing voice overlaid on top of complex, polyphonic music signal using text, genre, and artist labels as condition factors, but it requires massive computational power and datasets to be trained and lacks real-time control at generation time. The more recent architecture RAVE (Caillon and Esling 2021) addresses all the aforementioned issues in the context of modelling complex, polyphonic audio signals. However, given the potentially large dimensionality of the learned embedding and also the lack of labels for the latent space axes, there is a need to find a better way for real-time interaction and performing with such models.

Steering Generative AI

Real-time control in neural audio synthesis systems is important as it can enable performers to introduce the long-term temporal coherence often missing in these systems. That is, a generative model producing audio signals with short-term temporal coherence can still be used to generate longer structures if meaningful control is applied during generation. We next describe three main approaches to exerting control on the generative process.

Training data. In creative contexts, the choice of training dataset serves as the primary mechanism through which a human creator specifies what kind of content the machine should generate. This approach is often overlooked due to the extensive data and processing power required by most generative systems. However, working with small-scale datasets has been proposed as a means to allow greater human influence over generative AI systems in creative contexts, better aligning with creators' goals and ways of working (Vigliensoni, Perry, and Fiebrink 2022). In particular, when datasets are small, minor changes, such as the addition or removal of a few training examples, can significantly impact the trained model's behaviour.

Conditioning. In generative tasks, conditioning is a useful approach for controlling the generative process. By passing a certain condition to the network, the system can generate output conditioned on a specific variable. Conditioning can be applied when setting up the generative inference process (e.g., by using the artist or genre labels in Jukebox) or at

inference time (e.g., when conditioning DDSP with pitches and amplitude). In the case of RAVE, the generative process can be indirectly conditioned, such as by using sound content in a timbre transfer task. For instance, a beat track could serve as a MIDI-like clock, and the spectral content of an input signal can condition the output to generate a signal with similar frequency content.

Latent manipulation. This approach involves overriding latent dimension values with user input. For example, the RAVE architecture consists of an encoder that learns to map input audio data to a latent space and a decoder that learns to reconstruct the original data from the latent representation. When performing latent manipulation, one or more latent dimensions' values learned by the RAVE network can be replaced with the output from sliders controlled by a performer. Changes in values can be direct and absolute or relative to those generated by the encoder. In the latter case, arithmetic manipulation can be applied to the encoder output by adding a signal or multiplying it by a variable factor. From a performative perspective, latent manipulation is interesting because the performer can explore how the generative process changes when moving through orthogonal axes in the latent space. This exercise may help identify perceptual labels for specific dimensions. Alternatively, we propose below a novel approach to latent manipulation that uses supervised learning to map the human-performance space to the generative model's latent space.

Our Approach

The primary goal of this project is to devise and implement a real-time solution for steering a generative AI audio model towards a specific creative direction. Since the model has already undergone training, we cannot modify the underlying training data. Therefore, our sole means of interacting with the generative model involve conditioning it with specific features or performing latent manipulation. For example, we can condition the system by exciting the encoder with particular types of sounds, causing them to be projected into specific zones of the embedding and decoding similar sounds. Alternatively, we can perform latent manipulation by overriding the latent dimension values with user input.

The methodology we propose for performing and steering a neural audio model is inspired by research on machine listening systems. In this field, the most promising methods are hybrid systems that combine a data-driven approach informed by models of the perceptual and cognitive processes of the human auditory system (Heller et al. 2023). Similarly, our method to perform with a generative audio system involves utilizing a data-driven autonomous approach to learn the optimal representation for disentangling the audio data (e.g., using RAVE) and, subsequently, we work with the resulting embedding to identify creatively relevant or salient points within that space.

Interactive Machine Learning as a Mapping tool

Art- and music-making are non-teleological and purposeless activities in nature, not problems to be optimized (Audry 2021). As such, our approach to interacting with a neural audio model is centred on the curious and serendipitous exploration of its latent space. However, in order to facilitate more flexible and creative navigation of this space, we have explored the potential of interactive machine learning (IML) to map the human, well-known performance space onto the computer's label-less audio latent space.

IML (Fails and Olsen Jr 2003) is founded on the idea that training can be an incremental process in which the human and the machine collaborate to achieve a specific goal. In contrast to classical machine learning, where interaction with a model begins after it has been trained—usually following an extended offline period during which the algorithm iteratively optimizes to reach a certain model—IML as originally proposed by Fails and Olsen Jr involves a person iteratively experimenting with a machine learning model and tuning or steering its behaviour through changes to its training data. This human-machine interaction can happen over an extended period or in realtime, as the machine learns from human feedback and adjusts its models accordingly.

Such an IML approach has been used to create new gestural musical instruments since the introduction of the Wekinator tool (Fiebrink, Trueman, and Cook 2009), which enables people to iteratively construct and modify mappings from a human control space to sound synthesis parameters, through training examples pairing control-space coordinates and desired synthesis parameter values. To our knowledge, however, this approach has not previously been used to control generative models.

In this paper, we propose the IML paradigm as a tool for steering a generative audio model. The approach involves iteratively supplying training sets consisting of locations in the human-performance space paired with locations in the generative model latent space. We follow these steps:

1. We explore the latent space until identifying a point where an interesting zone emerges in terms of subjective creative possibilities. We describe this point with a descriptive auditory perceptual label. For example, "bright and loud" or "opaque and soft".
2. We select a source point in the performance space that should map to the target point in the latent space. Similar perceptual labels should be clustered together and can follow a meaningful progression in the performance space, such as arranging soft to loud sounds on the horizontal axis and from bright to opaque sounds on the vertical axis of a 2D controller.
3. We repeat the previous two steps as many times as desired, based on our creative aim. Ultimately, we will have a dataset comprising several pairs of source and target points linking the two spaces.
4. We instantiate the learning of a mapping between the performance and the latent space using the built dataset. A regression algorithm learns to map the points between the two spaces. These steps can be repeated to modify the mapping.

The mapping between the two spaces is shown in Fig. 1. The figure shows how a vector describing a point in "performance space" on the left is mapped onto a higher dimensional space given a series of models learned via regression.

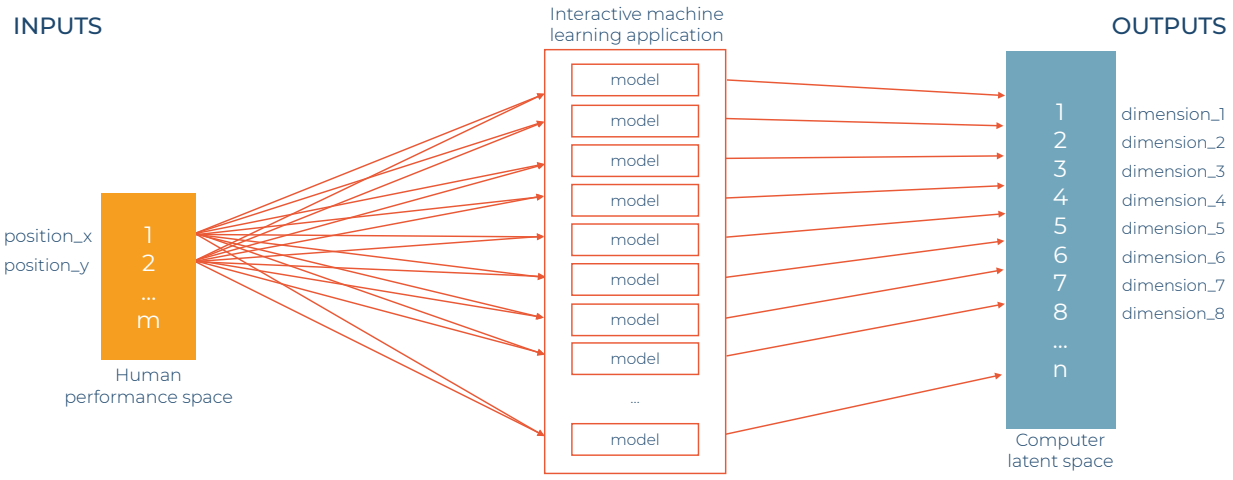

Figure 1: Interactive machine learning as a mapping tool. The low-dimension human-performance space is mapped to the high-dimensional computer-latent space. The mapping is done through a regressive task using a supervised learning approach.

For example, the input values could be the (x, y) coordinates when using a mouse on a canvas or an XY grid controller, or six values $(x_1, y_1, z_1, x_2, y_2, z_2)$ if using a six degrees of freedom controller such as a Gametrak.[1] We create one regression model per dimension of latent space, rather than one multi-dimensional model outputting a full latent space vector, to keep each modeling task simpler and thus trainable with fewer examples.

Experiments, Use, and Reflection

In our experiments, we have used the Wekinator and the FluCoMa (Tremblay, Roma, and Green 2021) toolboxes as frameworks for learning regression models using a supervised approach. Given that the number of training examples we have used is typically small (in the order of a few dozen), only a shallow (1 or 2 hidden layers) multilayer perceptron neural network is needed, facilitating very fast training and retraining. We have applied this method to map performance spaces where gestures are captured from on-screen and physical/gestural controllers using an arbitrary number of degrees of freedom (in our experiments, 2, 3, 6, and 15). These gestures have benn then mapped to steer RAVE latent audio models, encompassing a range from 4 to 64 latent dimensions.

In Figure 2, we present a graphical user interface of an instance of our approach using RAVE inside MaxMSP, and the FluCoMa package to learn a mapping between a human-performance space (a 2D mouse canvas in this case) to the computer-latent space (8D in this case). Once a mapping is learned, the selected zones of the latent space are mapped onto the performance space, and the performer plays the performance space.

Some of this experimentation has taken place in ideation and live performance contexts as part of the first author's preparation for Visiones Sonoras 18,[2] beat-based electronic music performances, both solo and in a duo with sound artist dedosmuertos, in which IML-generated models were employed for real-time gestural control of RAVE. We have also tried this setup in DJ sessions where the digital turntable's output signal has been timbre-transferred using audio models and our IML-enabled manipulation of the latent space.

Our approach has allowed us to interact and play with latent audio models in a straightforward and flexible way. In particular, it has enabled us to move between distant points in the latent space efficiently and relatably in the human performance space. Given the small amount of training data needed to learn the mappings, we have even retrained the system during the performance. The mappings between the spaces are not discrete but continuous, resulting in additional control as we can engage in constant subtle modulation of the latent space decoding, leading to continuously changing audio signals. In our experiments, we have experienced the immediacy of our approach to control over the generative audio process.

The most significant drawbacks we have experienced in performance are the latency of the generative system, which introduces a delay between the human gesture and the resulting action, and the potential existence of problematic zones in the latent space that can lead to unexpected loud sounds. While the former issue is inherent to digital audio buffering, we have addressed the latter by employing heavy limiting, rehearsing, and familiarizing ourselves with the spaces.

A video demonstrating the training of a mapping model and its use in performing with a high-dimensional audio latent space using a mouse and a Gametrak controller can be accessed at `https://bit.ly/iccc2023`.

Some key insights from this experimentation include: (i) Using shallow neural networks such as those in Wekinator and FluCoMa was adequate for building useful mapping functions that usually matched our intention. (ii) Even minimal training data (e.g. a few dozen examples) usually suffices to create a useful and playable mapping between the two spaces. The small size enables the training and retrain-

[1] `https://en.wikipedia.org/wiki/Gametrak`

[2] `https://en.cmmas.com/vs18`

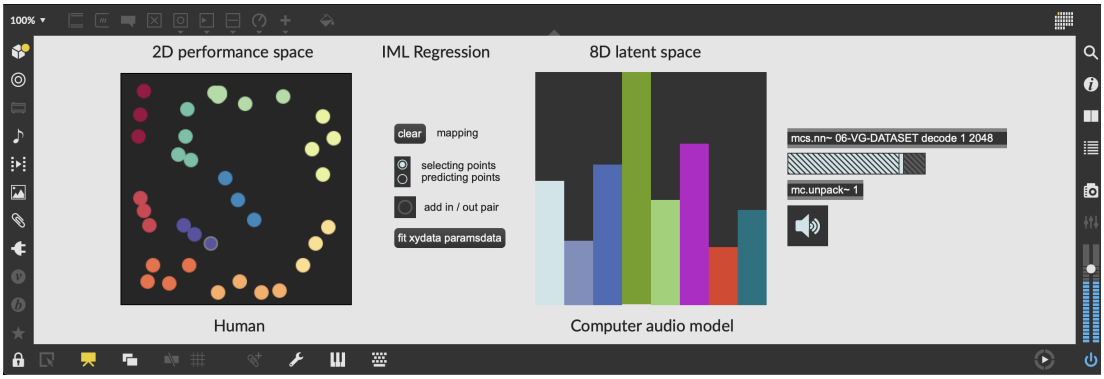

Figure 2: Graphical user interface showing the mapping between the human-performance space to a higher-dimensional latent audio model. In this example, an 8-dimensional space is controlled by means of a 2D space.

ing of models even at performance time. (iii) Sometimes, due to the small amount of data, our method yields models that do not perfectly match intentions. However, these individually crafted models of interaction can still prove to be useful and inspiring in a creative context. (iv) This approach facilitates creation of control trajectories that allow for drastic or smooth transitions between points in the latent space. (v) The IML approach to mappings promotes fast prototyping, flexibility in mappings creation, and immediacy of control. (vi) In performance, this approach allows us to overcome the issue of short temporal coherence often found in generative neural audio systems. Because a performer has control over the generative process, they can maintain a longer window of coherence and manipulate sound and music motives and tension more effectively. This can be achieved, for instance, by revisiting or introducing new timbres or motifs.

Conclusion

We have described how IML can enable performers to map from real-time control vectors—from on-screen or physical controls—to creatively relevant or salient points within a latent space. We have found that IML can be an effective tool for enabling real-time, performative interactions with generative models, even when the latent space of a model is high-dimensional and its dimensions do not neatly correspond to perceptual attributes.

Acknowledgments

This research has been supported by Fonds de recherche du Québec – Société et culture (FRQSC) through a research creation grant (Ref. 2022-B5R-310633). Important parts of this work used Digital Research Alliance of Canada's High Performance Computing resources.

References

Audry, S. 2021. *Art in the Age of Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Caillon, A., and Esling, P. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. arXiv:2005.00341.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations (ICLR 2020)*.

Fails, J. A., and Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 39–45.

Fiebrink, R.; Trueman, D.; and Cook, P. R. 2009. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of the 9th International Conference on New Interfaces for Musical Expression*, 280–285.

Heller, L. M.; Elizalde, B.; Raj, B.; and Deshmukh, S. 2023. Synergy between human and machine approaches to sound/scene recognition and processing: An overview of ICASSP special session. In *Special session on "Synergy between human and machine approaches to sound/scene recognition and processing" at the 2023 ICASSP*. arXiv:2302.09719.

Mehri, S.; Kumar, K.; Gulrajani, I.; Kumar, R.; Jain, S.; Sotelo, J.; Courville, A.; and Bengio, Y. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. In *International Conference on Learning Representations (ICLR 2017)*.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A generative model for raw audio. arXiv:1609.03499.

Tremblay, P. A.; Roma, G.; and Green, O. 2021. Enabling programmatic data mining as musicking: The Fluid Corpus Manipulation toolkit. *Computer Music Journal* 45(2):9–23.

Vigliensoni, G.; Perry, P.; and Fiebrink, R. 2022. A small-data mindset for generative AI creative work. In *Proceedings of the Generative AI and Computer Human Interaction Workshop (GenAICHI, CHI 2022 Workshop)*.