Explainability Paths for Sustained Artistic Practice with AI

AUSTIN TECKS, THOMAS PESCHLOW, and GABRIEL VIGLIENSONI, Concordia University, Canada

The development of AI-driven generative audio mirrors broader AI trends, often prioritizing immediate accessibility at the expense of explainability. Consequently, integrating such tools into sustained artistic practice remains a significant challenge. In this paper, we explore several paths to improve explainability, drawing primarily from our research-creation practice in training and implementing generative audio models. As practical provisions for improved explainability, we highlight human agency over training materials, the viability of small-scale datasets, the facilitation of the iterative creative process, and the integration of interactive machine learning as a mapping tool. Importantly, these steps aim to enhance human agency over generative AI systems not only during model inference, but also when curating and preprocessing training data as well as during the training phase of models.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Interaction design**; **Visualization**; • **Applied computing → Arts and humanities**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: explainable AI (XAI), artificial intelligence (AI), sound art, generative arts, human-computer interaction (HCI), neural audio synthesis, interaction design

ACM Reference Format:
Austin Tecks, Thomas Peschlow, and Gabriel Vigliensoni. 2024. Explainability Paths for Sustained Artistic Practice with AI. In *2nd International Workshop on Explainable AI for the Arts (XAIxArts), ACM Creativity and Cognition (C&C '24), June 23, 2024, Chicago, USA.* , 4 pages.

1 INTRODUCTION

The rapid proliferation of pre-trained generative audio models belies the minimal adoption these tools have seen as part of sustained sound and music practices. In step with prevailing AI trends, popular tools like Suno[1] and Stable Audio[2] have adopted text-based user interfaces favouring broad appeal and immediate access at the expense of human agency and interpretability [12]. This trend is highlighted by the widespread use of natural language conditioning, which, despite broadening user access, presents limitations for sound artists and musicians who require precise control to maintain long-term temporal coherence in the sonic output of generative systems [14]. Moreover, in a bid to offer wide-ranging user experiences, generative audio models such as Jukebox and Suno have been trained on unfathomably large datasets, the former boasting a dataset of more than 1.2 million songs [6]. Although this may facilitate the generation of novel and unexpected hybrids across music genres and provide satisfying responses to common prompts, the practical value of these innovations for artists seeking a sustained practice remains questionable.

While the conditions presented above have successfully piqued the public's interest in generative AI art, our ongoing research-creation practice with neural audio synthesis illuminates alternative paths, prioritizing explainability for sustained artistic practice. This paper highlights our work preparing intimately curated, small-scale datasets for

[1] https://suno.com/
[2] https://stableaudio.com/

Manuscript submitted to ACM

training and implementing neural audio synthesis models with RAVE [2]. Throughout this process of data curation and preprocessing, training, and model inference, we have identified the following provisions for improved explainability:

- Improving agency through human-scale models and artist-curated datasets
- Extending the iterative process beyond inference to curation and training
- Defining the performance space through interactive machine learning [7]

In Sections 2, 3, and 4, we will define the above-mentioned provisions, and in Section 5, we will draw from our research-creation practice, illustrating the potential of XAI to reconcile the opaqueness of generative AI with the demands of sustained artistic practice.

2 HUMAN-SCALE MODELS, ARTIST-CURATED DATASETS

The suitability of AI-driven generative audio for sustained creative work depends on artist control over training material, which in turn often relies on the viability of smaller-scale datasets. While mainstream machine learning models, pre-trained for public use, thrive on large amounts of data, artists usually find greater utility in models with a more narrow scope, facilitating a deeper connection to the training material as well as the ability to more effectively steer the model towards a creative objective [15]. Notable examples of extremely narrow and focused datasets are Holly+, a sonic digital likeness of Holly Herndon capable of reconstructing her voice [8], and the early work by Dadabots, where single albums are used as training data to generate music "within the limited aesthetic space of the album's sound" [3]. Similarly, using smaller datasets can enhance transparency, agency, and bolster an artist's confidence and trust in a model, thereby addressing a major challenge artists and the public face in adopting AI-driven technologies [5].

3 EXTENDING THE ITERATIVE PROCESS BEYOND INFERENCE

Iteration is an essential process for artistic development. It facilitates the emergence of novel and meaningful insights [4] and is a powerful force for creativity [13]. Perhaps the most appealing affordance provided by pre-trained models is the ability to jump straight into an iterative, inferential process within a satisfactory time frame, a process enabled by externalized computational power. This experience, however, usually confines the creative process to iterative conditioning on text prompts. Furthermore, artists may find such a model inadequate in providing the conditions required for fruitful creative iteration due to a lack of fine control of the generative process, its intrinsic non-causality, and the lack of long-term temporal coherence. We propose that artists would benefit most from machine learning models that support an iterative process both during the inference and training phases. This provision is interdependently related to the scale of the model, as smaller-scale datasets potentially yield shortened training times.

4 DEFINING THE PERFORMANCE SPACE THROUGH INTERACTIVE MACHINE LEARNING

We also link explainability in creative AI practices to a practitioner's ability to steer the models during performance. At training time, artists guide the learning process by curating the data for training the system and by continuously observing and adjusting the process. However, at inference time, due to the stochastic nature of training, artists may have to perform with models whose axes and parameters are unknown. To address this issue, we propose a regressive approach where we map the familiar human space to the computer's latent space, using interactive machine learning as a mapping tool [14]. This method involves exploring the latent space, identifying points of interest, and mapping these to specific points in the performance space. By doing so, we define the space rather than merely explaining it. This strategy has proven effective for real-time interaction with generative models, even those that are highly dimensional.

5 CASE STUDY

As a case study, we provide insights derived from our experience training generative audio models on a dataset composed of archival recordings from the Museo de la Memoria y los Derechos Humanos in Santiago, Chile.[3] In our overview, we examine the crucial creative steps in the development of this project: data-preparation, training, and performance.

5.1 Data Preparation

Data preparation entails curating, classifying, and normalizing a given dataset. This process initially precedes training but exists within an iterative data preparation/training/implementation cycle, thus providing ample opportunity for applied human agency. In preparing our data, we curated three separate datasets from about 35 hours of viable audio recordings. These datasets were organized based on their historical and semantic content as well as their sonic coherence, resulting in categories of public recordings, music, and home recordings. From here, the recordings were processed and normalized to enhance fidelity and achieve a baseline of intra-dataset spectral congruency.

Normalizing the audio data and sequestering the recordings into separate datasets according to their sonic characteristics facilitate an optimal training process. If done effectively, this process can improve the reliability of a generative model, allowing us to better anticipate its behavior, an essential component for the foundation of trust and transparency [10]. Relatively short training cycles enable us to alter the decisions made at this stage with each training iteration, effectively compounding human agency within the data preparation process.

5.2 Training

RAVE's training process, composed of clearly segmented phases with distinct task orientations, allows for meaningful alterations to the model over the iterative process, further enabling human agency. The first training phase is based on a variational autoencoder (VAE) [11]. It focuses on the mathematical optimization of a compression and decompression process in which the size of the bottleneck, or the dimensionality of the latent space, is an important hyperparameter that can both be set by the user, or automatically derived by the model by virtue of the model being what is called a disentangled variational autoencoder [1]. The second training phase is based on a generative adversarial network (GAN) [9]. This phase focuses on the perceptual optimization of the output by fine-tuning the model through a process consisting of translating the encoding of noisy data into sound pertaining to the original dataset.

For the VAE phase, a 5 million-step training phase was deemed optimal for all three datasets despite their sonically distinct nature. This may be due to the shared prominence of the human voice and the relative absence of discontinuous transient-rich material. In the GAN phase, we maintain consistency between datasets, ranging from one to two million steps. This phase often presents the most significant fidelity improvements. However, prolonged training in this phase can produce a smoothing effect on output, potentially diminishing certain sound characteristics and introducing out of domain sonic artifacts to generation output.

5.3 Performance

One feature that sets RAVE apart from other generative audio models is its ability to perform model inference in real time. We chose a latent space size of eight dimensions to provide a nuanced understanding of each axes' impact on output while allowing diverse sound reconstruction. We have experimented with various gestural interfaces to control RAVE models. For this project, we chose to use our face as the performance space, utilizing Google's MediaPipe

[3] https://archivoradial.museodelamemoria.cl/

Face Mesh model[4] to embody this interaction. We use the interactive machine learning approach to map our facial landmarks and movements to the model's latent space. A video demonstrating the degree of steerability achieved with the provisions stated in this paper can be watched at `https://media.vigliensoni.com/video/xaixarts2024`.

6 CONCLUSION

Through our research-creation practice, we have identified reliable methods to enhance explainability and steerability throughout all stages of interaction with a generative audio model. Smaller datasets, supported by models like RAVE, allow artists to work with datasets with which they are intimately familiar. Extending the iterative process to training compounds artist agency and improves a model's explainability. Additionally, the inherent iterative nature of interactive machine learning and its mapping capabilities enable artists to define axes and zones for exploration in generative models. This allows them to create causal gestures and produce sound and music with long-term temporal coherence. We believe that sustained artistic practice benefits significantly from the explainable pathways these provisions supply.

ACKNOWLEDGMENTS

This project has been funded by Concordia University's FRDP grant. Important parts of this work used Digital Research Alliance of Canada's High Performance Computing resources.

REFERENCES

[1] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-VAE. arXiv:1804.03599

[2] Antoine Caillon and Philippe Esling. 2021. RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis. https://doi.org/10.48550/arXiv.2111.05011 arXiv:2111.05011.

[3] C. J. Carr and Zack Zukowski. 2018. Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands. In *Proceedings of the 6th International Workshop on Musical Metacreation (MUME 2018).*

[4] Joel Chan and Christian D. Schunn. 2015. The Importance of Iteration in Creative Conceptual Combination. *Cognition* 145 (Dec. 2015), 104–115. https://doi.org/10.1016/j.cognition.2015.08.008

[5] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction* 39, 9 (2023), 1727–1739.

[6] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. https://doi.org/10.48550/arXiv.2005.00341 arXiv:2005.00341.

[7] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03).* Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

[8] Freethink. 2023. AI Is Changing Music Forever | Holly Herndon and Mat Dryhurst. https://www.youtube.com/watch?v=qPW_rdUgV_8

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661

[10] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Online) *(FAccT '21).* ACM, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[11] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114

[12] Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli. 2022. An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing* 470 (Jan. 2022), 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

[13] R. Keith Sawyer. 2021. The Iterative and Improvisational Nature of the Creative Process. *Journal of Creativity* 31 (Dec. 2021), 100002. https://doi.org/10.1016/j.yjoc.2021.100002

[14] Gabriel Vigliensoni and Rebecca Fiebrink. 2023. Steering Latent Audio Models through Interactive Machine Learning. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23).* Waterloo, ON, 19–23.

[15] Gabriel Vigliensoni, Phoenix Perry, and Rebecca Fiebrink. 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).* Online.

[4] `https://developers.google.com/mediapipe/solutions/vision/face_landmarker/index`