WÆnderer: Projecting and Navigating User-Defined Audio Corpora in Pre-Trained Audio Models

Dominic Thibault
Faculty of Music
Université de Montréal
dominic.thibault@umontreal.ca

David Piazza
Centre for Interdisciplinary Research in Music Media and Technology
david.piazza@umontreal.ca

Abstract

Neural audio synthesis offers powerful ways to create sound, but many systems remain difficult to use as musical instruments. Artist-specific models demand large, curated datasets and training resources, while accessible text-to-audio services give limited control over personal material and often obscure data provenance. We present WÆnderer, a system for real-time sound generation that repurposes a pre-trained variational autoencoder (VAE)—a model that compresses audio into a continuous latent representation and reconstructs it through a decoder—without retraining the audio model. WÆnderer encodes small, artist-defined corpora, organizes their sound segments in an abstract timbre space derived from audio descriptors, and lets performers navigate that space while the VAE reconstructs retrieved latent frames. The system therefore does not generate audio from text or unconstrained latent sampling; it reconfigures a musician's own recorded material within the representational limits of the pre-trained model. A practice-led evaluation identifies three design findings: corpus-aligned retrieval avoids artifacts observed during direct interpolation; decoding-window length trades responsiveness against continuity; and manual, trajectory-based, and morphology-based navigation produce distinct forms of control and sonic organization. This research-creation study contributes an accessible, locally executable approach to neural audio interaction and frames latent-space performance as navigable, corpus-based music rather than autonomous generation.

1 INTRODUCTION

Recent advances in neural audio synthesis (NAS) have significantly expanded the range of tools available for sound generation, enabling the creation of complex audio material through generative audio models. These models are increasingly introduced into musical practice through a variety of interfaces, including web-based platforms, standalone applications, and plugins.

Generative audio models have already been the subject of growing interest within the research community, particularly for the new affordances they introduce in creative and musical contexts. Recent work (Caillon & Esling, 2021) has explored how such models can support real-time synthesis and interaction, as well as how their integration into digital musical instruments reshapes embodied musical practices and performer–instrument relationships. Beyond their technical capabilities, these systems are increasingly considered as instruments in their own right (Zappi & Tatar, 2025), enabling novel forms of timbral exploration, control, and interaction grounded in learned organizations of sound.

Despite these developments, a gap remains between the technical capabilities of generative audio models and their effective integration into musical practice. Widely accessible systems are predominantly based on text-driven interaction with large-scale models (Evans et al., 2024), offering limited control over material specificity and often producing outputs that are detached from the user's own sonic identity[1] (Collins, 2025). At the same time, approaches that afford a higher degree of artistic control typically require training models on curated datasets, involving substantial preparation, technical expertise, and computational resources. This gap is not only identified in the literature but also emerges in artistic and pedagogical contexts. In our teaching, where music students are invited to develop projects that both employ and critically reflect on generative AI systems, we consistently observe difficulties in translating these models into meaningful, controllable, and musically engaging interactions. Rather than functioning as instruments, these systems often remain opaque, difficult to steer in real time, and disconnected from established practices of musical performance.

This paper originates from the observation that many musicians are not primarily seeking tools for generating novel audio content, but rather systems that enable them to interact with, transform, and recontextualize their own sound materials. In this context, the challenge is not only how to generate sound using machine learning, but how to design systems that support meaningful interaction with model-based representations of audio. This shift in perspective suggests reconsidering the role of generative models, not as autonomous producers of content, but as

[1] Though it is important to mention that new forms of creativity are discovered in the process of interacting with text-to-audio models (Coelho, 2025)

structures that can be navigated and appropriated within situated musical practices.

To address this gap, we propose an alternative approach based on the use of pre-trained variational autoencoders (VAEs) as interactive musical instruments. Rather than training models on large, artist-specific datasets, small corpora of user-defined sounds are projected into the latent space of existing models, enabling real-time navigation through perceptually grounded descriptors.

The aim of this paper is to investigate how such an approach can contribute to the development of meaningful and accessible tools for musical creation. In particular, it addresses the following research questions:

- How can recent advances in generative audio models be adapted into accessible and meaningful tools for musical practice?
- To what extent can these models support new forms of interaction and sonic exploration for musicians and composers?
- How can such systems be designed to minimize data, training time, and computational resources while remaining usable and informative for musicians?

This work is situated within a research-creation framework. WÆnderer was iteratively developed through musical practice, and its architecture and interaction modes were refined in response to observed sonic behaviors. We therefore present both the tool and the affordances encountered so far in our own practice. The study is a preliminary exploration of an audio engine and its modes of interaction, not a finalized system or a general user evaluation.

Through the presentation of WÆnderer and a discussion of its affordances, this paper examines how corpus-based navigation in latent audio spaces can foster alternative modes of interaction with generative audio models. More broadly, it raises the question of how musicians may wish to engage with such systems, and whether the current emphasis on generation aligns with the needs and values of contemporary musical practices.

2 BACKGROUND & CONTEXT

2.1 Corpus-Based and Concatenative Practices

Concatenative synthesis reorganizes short audio fragments drawn from a corpus. Formalized as a musical technique in systems such as CataRT (Schwarz, 2006), it can be understood as timbrally controlled sampling. Its development has progressively shifted emphasis from sound generation toward interaction design, notably by associating audio units with descriptor-based representations (Bell, 2023; Ghisi, 2017; Schwarz, 2012). Sound fragments are analyzed, positioned in multidimensional feature spaces, and selected or recombined in real time according to perceptual similarity (Bell, 2023; Constanzo, 2025; Garber & Ciccola, 2021; Thibault, 2024).

This organization is widely conceptualized as a timbre space (Schwarz et al., 2016), in which distances represent dissimilarities among sounds and support spatial navigation rather than linear sequencing. Dimensionality-reduction techniques such as principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) can project high-dimensional descriptor data into lower-dimensional abstract timbre spaces. In these projections, spatial relations no longer correspond to single descriptors; they reflect statistical structure across the feature set and thereby support exploratory and performative engagement with a corpus. Descriptor spaces are not direct models of auditory perception—timbre is a complex, multidimensional structuring force shaped by perceptual and cognitive processes (McAdams, 2019)—but their approximations can still support intuitive interaction. This tension between representational limitation and musical usability is productive for corpus-based practice.

Recent work further emphasizes the instrumental dimension of these spaces. Systems such as Mosaïque (Thibault et al., 2025) demonstrate how abstract timbre spaces can be operationalized as interactive musical environments, where corpora are explored through continuous navigation rather than discrete triggering. In this context, timbre space becomes the primary interface through which sound is accessed, structured, and performed, enabling forms of musicking grounded in gesture, listening, and spatial exploration.

Taken together, these developments indicate that timbre space is more than a technical construct; it is an emerging musical paradigm in which activity becomes the exploration of a continuous sonic topology. Concatenative synthesis also foregrounds the dataset, or corpus, as a compositional element (Bryan-Kinns et al., 2025), making collection, curation, and organization primary sites of creative agency. Corpus-based approaches are therefore not only methods of sound generation but frameworks for spatial, exploratory, and interaction-driven musical practice.

2.2 Neural Audio Models and Latent Representations

Neural audio synthesis (NAS) has developed rapidly as a field concerned with generating and transforming sound using deep learning. Early approaches relied on autoregressive architectures such as WaveNet (Oord et al., 2016) and SampleRNN (Mehri et al., 2017), followed by variational autoencoders (VAEs), generative adversarial networks (GANs), and, more recently, diffusion models (Božić & Horvat, 2024). Across these architectures, latent representations provide continuous spaces in which audio can be encoded and organized.

2.2.1 What Is a Pre-Trained Variational Autoencoder?

A variational autoencoder is a neural network with two coupled parts. An encoder compresses an audio signal into the parameters of a probability distribution in a lower-dimensional latent space; a sample from that distribution is then passed to a decoder, which reconstructs the signal. During training, the model is optimized both to reproduce its inputs and to keep the latent distribution close to a simple prior. This regularization encourages a continuous space in which nearby latent points can often be decoded into related outputs, although those distances are not guaranteed to correspond to human perceptual similarity (Božić & Horvat, 2024; Caillon & Esling, 2021)

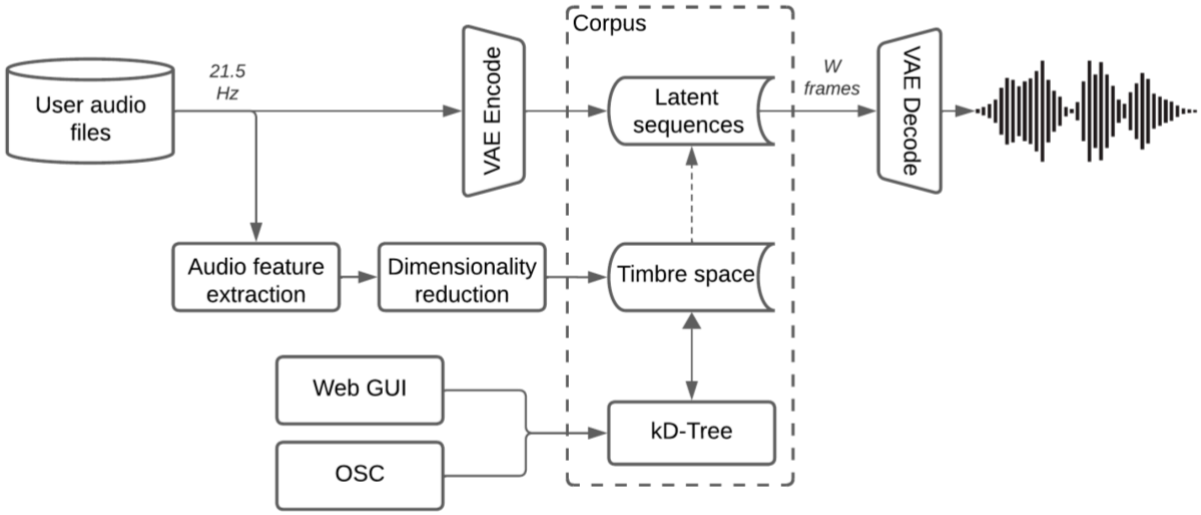

Figure 1 Audio is processed along two paths: VAE encoding for synthesis (top) and descriptor extraction for navigation (bottom). The reduced timbre space (kD-tree) is accessed via WebUI/OSC to retrieve latent frames, which are projected onto the corpus manifold, assembled into sequences, and decoded to audio.

Here, pre-trained means that the encoder-decoder pair has already been trained on a large external audio collection and its weights are reused without further adaptation. WÆnderer passes an artist's own corpus through the fixed encoder, stores the resulting latent frames, and later sends selected frames to the fixed decoder in real time. The model therefore acts less as an autonomous composer than as a learned representational medium: it reconstructs personal material, while its architecture and original training data still shape the sound.

Beyond their computational role, latent spaces have increasingly been approached as interfaces for musical interaction (Tahiroglu & Wyse, 2024; Yee-King, 2022), enabling navigation, manipulation, and performance within learned representations of sound. This paradigm has been particularly supported by systems such as RAVE (Caillon & Esling, 2021), which have made latent space exploration accessible to musicians and fostered its adoption in artistic contexts. Concepts such as "soundwalking" in latent spaces (Scurto & Postel, 2023) further emphasize exploratory and embodied interaction, framing these spaces as environments in which gestural and perceptual relationships to sound can be developed (S. J. Zheng et al., 2025).

Despite these advances, latent spaces remain difficult to navigate: they are typically high-dimensional, entangled, and poorly aligned with perceptual attributes, leading musicians to rely on interpolation, random sampling, or trial and error. Dimensionality reduction, PCA-based interfaces (Horta Valenzuela & Tomás, 2025), nn_terrain (S. Zheng et al., 2024), and interactive machine-learning mappings (Vigliensoni & Fiebrink, 2024) can improve access, but they do not remove two constraints.

However, a fundamental limitation persists: most approaches require training or adapting models for specific datasets. Training models such as RAVE demands substantial computational resources and curated corpora, which can restrict accessibility. Moreover, latent spaces are not inherently aligned with perceptual organization; studies have shown that they are often dominated by factors such as pitch, while timbral characteristics remain entangled and difficult to control (Cameron & Blackwell, 2026).

In parallel, recent neural text-to-audio models rely on pretrained generalist audio VAEs trained on large and diverse datasets (Božić & Horvat, 2024; Evans et al., 2024; Liu et al., 2023), enabling them to reproduce a wide range of sounds. While primarily designed for offline generation, these models embed rich latent representations that may be repurposed for alternative forms of interaction.

2.3 Hypothesis

We hypothesize that generalist pre-trained audio VAEs, such as those used in text-to-audio systems, can be repurposed as real-time navigable environments for musical interaction. Rather than retraining the audio model on an artist-specific dataset, this approach operates within its existing representation to engage a user-defined corpus. The practical challenge is to adapt a model designed for offline generation to responsive performance and rapid musical iteration.

3 SYSTEM OVERVIEW

WÆnderer, the proposed system, reuses a pre-trained audio VAE as a structure for real-time interaction with user-defined collections of sounds, or corpora. This is achieved through a pipeline (see Figure 1) that links descriptor-based navigation to latent space decoding.

To complement the written description of the system, readers are invited to visit the companion website https://lfo-lab.github.io/WAEnderer/ , where audiovisual demonstrations make it possible to see and hear the system in action. These materials provide concrete examples of the interactions between the interface, the audio corpus, the descriptor space, and the resulting sound transformations discussed throughout this section.

3.1 System Pipeline

The system operates by linking a user-defined audio corpus to real-time decoding through a series of analysis, indexing, and retrieval stages. During preprocessing, audio files are segmented and encoded into sequences of latent vectors using a pre-trained VAE. For the Stable Audio Open model (Evans et al., 2024) used in this work, latent representations are produced at a fixed rate of approximately 21.5 Hz, yielding a temporally structured latent sequence for each input signal.

In parallel, each audio segment is described through computed features (Peeters et al., 2011), including spectral centroid, spread, flatness, and rolloff; mel-frequency cepstral coefficients (MFCCs); chroma; pitch estimates; and loudness. These features form a descriptor space that captures salient characteristics of the corpus. PCA or UMAP then projects this high-dimensional space into a low-dimensional abstract timbre space, which serves as the main navigation interface.

Each point in this abstract timbre space is associated with corresponding latent frames derived from the corpus encoded through the VAE, establishing a mapping between descriptor-based queries and latent representations. At runtime, interaction consists of traversing this control space according to the navigation strategies described in 3.2 Modes of Navigation. Rather than generating unconstrained latent trajectories, the system performs nearest-neighbor retrieval in the reduced feature space, selecting latent frames whose descriptors are proximal to the current control position.

The retrieved latent frames are then assembled into variable-length temporal windows and decoded as fast as possible, allowing the system to balance responsiveness and continuity under real-time constraints. Continuous audio output is produced through adaptive windowed decoding and crossfading between successive segments, ensuring perceptual continuity while preserving low-latency interaction.

3.2 Modes of Navigation

The system supports multiple interaction modes that define how trajectories are generated within the abstract timbre space and how corresponding latent frames are selected. These modes provide complementary approaches to navigating the corpus, ranging from direct user control to partially or fully automated behaviors.

3.2.1 Manual Mode

In manual mode, navigation is driven directly by the performer through a four-dimensional control interface (X, Y, Z, and a color gradient dimension). The position in the abstract timbre space determines the retrieval of latent frames via nearest-neighbor search. Two mechanisms are used to introduce variation: alternating between k-nearest neighbors to avoid repeated selection of identical frames, and applying Gaussian perturbation to retrieved latent vectors prior to decoding. These processes operate at the level of successive decoding windows and affect the selection and variation of latent frames over time.

3.2.2 Exploration Mode

In contrast, exploration mode introduces algorithmic navigation based on learned or heuristic latent trajectories. When a trained model is available, a gated recurrent unit (GRU) predicts successive latent frames from the current state. These predictions are not decoded directly; each is projected back onto the corpus by retrieving a nearby encoded frame, so the output remains grounded in the dataset. In the absence of a trained model, heuristic latent motion provides a similar prediction-retrieval trajectory.

3.2.3 Reorganized Morphologies Mode

Finally, reorganized mode operates at a higher structural level by traversing a graph of variable-length sonic morphology units derived from the corpus. These units are segmented according to morphological criteria inspired by Denis Smalley's concept of spectromorphology (Smalley, 1997) , reflecting patterns of energy, gesture, and temporal evolution in sound. Units are connected through a transition model that can be either heuristic or learned, forming a navigable structure over the corpus. Navigation in this mode consists of selecting successive units based on candidate scoring, optionally incorporating learned transition probabilities. The resulting trajectories emphasize larger-scale continuity and transformation, enabling the reorganization of sonic material according to morphological relationships rather than solely timbral proximity.

3.3 Control Interfaces

Interaction with the system is implemented through a set of real-time control interfaces designed for performance and experimentation. The system is deployed as a Python application, where a main script initializes the audio engine and concurrently launches a Web-based interface (see Figure 2) and an Open Sound Control (OSC) server.

The Web interface provides both control and visualization of the system. It supports three primary operational modes: corpus preprocessing, model training (GRU-based navigation and MLP-based morphological transitions), and sound generation. In generation mode, the interface enables navigation across the three modes described in 3.2 Modes of Navigation. A central component of the interface is the interactive 3D visualization of the corpus embedding, which allows users to explore and manipulate the abstract timbre space by controlling a cursor within it. Control parameters are specific to each navigation mode, allowing tailored interaction with manual, exploration, and reorganized behaviors.

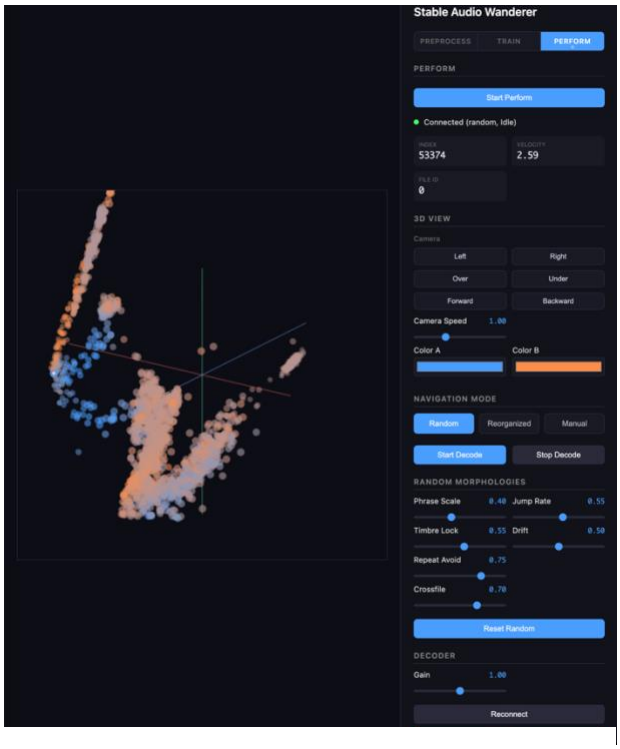

Figure 2 Web interface of WÆnderer, showing the 3D visualization of the abstract timbre space and controls for corpus processing, model training, and real-time navigation.

In parallel, the OSC interface exposes a network-based control layer through a dedicated input port, receiving structured messages that map to navigation and decoding parameters. These messages enable external systems to control cursor position, modify navigation behavior, and adjust synthesis in real time. Widely used in computer-based musical interfacing and digital musical instruments, OSC provides a practical bridge between standalone tools and creative coding environments such as Max, Pd, and SuperCollider. By exposing the main navigation and synthesis parameters of WÆnderer as OSC endpoints, the system can be integrated as a modular component within an artist's setup, allowing latent-space navigation to interact with controllers, live-coding systems, spatialization patches, and other real-time processes.

4 NAVIGATING THE CORPUS

WÆnderer combines two representations of sound—a descriptor-based organization of a corpus and its reconstruction through a pre-trained VAE—to support musical interaction. In our preliminary practice, its instrumentality depended on two design objectives: musicians should be able to prepare and revise a corpus quickly, and they should be able to control the resulting audio in real time.

4.1 Rapid Iteration

First, the system was conceived to support rapid and efficient interaction, enabling iterative exploration of its sonic behavior. Our aim was to provide a tool in which musicians could promptly provide a dataset and start playing with the resulting sound engine.

In practice, the pipeline of WÆnderer was designed to preprocess audio corpora and train the associated navigation models within a short time frame. For a dataset of 60 minutes of audio, preprocessing requires just over five minutes (5:12), while the training of the navigation models—including the UMAP embedding for manual mode, the GRU model for exploration mode, and the MLP model for reorganized morphologies—takes approximately 2 minutes and 20 seconds on an M1 Max MacBook Pro. We consider this duration sufficiently short to support iterative workflows, allowing musicians to construct a dataset and begin navigating the resulting abstract timbre space with minimal delay.

This objective was informed by prior work with Mosaïque (Thibault et al., 2025), where immediate control and feedback were essential for understanding and appropriating corpus-based concatenative synthesis. It also aligns with observations by Bryan-Kinns et al. (2025) and Tremblay et al. (2019) on accessible, iterative engagement with sound materials, while contrasting with approaches such as RAVE (Caillon & Esling, 2021) that depend on lengthy model training before a timbre space can be explored.

The system treats the dataset as a central component of the instrument, positioning the corpus as a material to be constructed, curated, and explored. In this context, preparing a dataset is not only a technical prerequisite but a compositional act, shaping the sonic possibilities of the system. This perspective extends observations from corpus-based approaches, where the selection and organization of sound materials directly influence the resulting interaction. By enabling rapid iteration on datasets without requiring extensive model training, WÆnderer reinforces this relationship and supports a workflow in which corpus design and musical exploration are closely intertwined.

4.2 Navigation as a Musical Paradigm

WÆnderer can be understood as an instrument defined by relations among corpus, control, and response: the corpus supplies the material, the abstract timbre space supplies a navigable control structure, and the VAE reconstructs selected latent frames within the constraints of its learned representation. Instrumentality arises from this coupling rather than from a fixed synthesis process.

The system suggests a alternative to how neural audio models can be engaged in musical contexts. Rather than approaching these models as generators of new audio material, WÆnderer frames interaction as a process of navigation within a structured space of existing sounds. This approach allows navigation in an abstract timbre space, based on perceptual organization, while benefitting from model-based reconstruction.

In this context, the system does not aim to maximize generative capacity. Instead, it operates through the retrieval and reorganization of corpus-derived latent material, constraining the output to regions of the latent space associated with encoded audio. This approach departs from prompt-based generation and unconstrained latent synthesis, and instead defines a mode of interaction grounded in reconstructive navigation, where variation

emerges from the exploration of structured relationships within the corpus.

All three modes of navigation presented in 3.2 Modes of Navigation operate within this shared framework of abstract timbre space navigation and latent frame retrieval. Together, they define a spectrum of interaction strategies combining direct control, stochastic variation, and structure-aware sound organization.

Taken together, these elements suggest that navigation itself can function as a primary musical gesture. Musical form emerges from movement within a space defined by both the dataset and the model, positioning latent space not as a site of generation, but as an environment for structured exploration and interaction.

5 DISCUSSION

5.1 Practice-Led Evaluation and Design Findings

The development of WÆnderer was closely informed by iterative listening and evaluation of its sonic behavior. These observations played a central role in shaping both the system architecture and the design of its navigation modes. In particular, the use of the Stable Audio Open VAE revealed characteristic artifacts when navigating the latent space directly. Interpolation between latent vectors often produces a perceptible hum, which is outside of the corpus and overwhelmingly distracting. This behavior was a determining factor in the decision to constrain playback to corpus-aligned latent frames, ensuring that decoded audio remains anchored in stable regions of the latent space.

The three navigation modes exhibit distinct sonic characteristics. In manual mode, the system enables controlled navigation of timbre through the abstract timbre space in four dimensions. While this allows for exploration of perceptual similarity, the resulting sound often remains relatively static, and the source corpus is not always clearly recognizable. This behavior is comparable to latent navigation in systems such as RAVE, where interpolation produces hybrid timbres that do not necessarily preserve the identity of the original material. Movement across the four-dimensional control space can nevertheless reveal morphologies that resemble those of the corpus. The size of the decoding window plays a critical role: shorter windows introduce perceptible repetition in the sound produced, while longer windows produce smoother audio at the expense of responsiveness.

The exploration mode was developed in response to these observations. By introducing trajectory-based navigation through a GRU trained on latent sequences, the system generates continuous motion within the corpus. Although the process can be jittery and lead to rapid transitions, it results in a stronger perceptual connection to the source material, with recognizable fragments emerging more frequently. User controls allow partial steering of the model's behavior, modulating its tendency toward stability or activity. This mode offers a dynamic form of exploration, while also revealing limitations in terms of control and consistency.

The reorganized morphologies mode was developed as a further response to the behavior of the exploration mode. The observed instability of trajectory-based navigation motivated an approach based on the segmentation of the corpus into sonic morphology units, inspired by spectromorphological theory (Smalley, 1997). This mode aims to organize sound according to perceptual structures rather than local similarity alone, enabling the retrieval of characteristic temporal patterns present in the corpus. Initial results suggest that this approach can produce more coherent large-scale organization of sound material. However, further work is required to develop robust methods for segmentation, representation, and control of these structures.

Finally, the evaluation of the system also highlights the influence of the underlying VAE model. The variability between pretrained VAEs suggests that such models may themselves constitute distinct sonic entities, shaped by their architecture and training data. From this perspective, choosing a VAE may become analogous to selecting an instrument, where different models afford different forms of interaction and sonic behavior. Preliminary experiments conducted with alternative models, such as ɛar-VAE (Wang et al., 2025), indicate perceptible differences in the resulting audio, supporting this observation. Future work could explore the development of community-trained VAE models, allowing practitioners to navigate more ethical latent spaces.

5.2 Data, Models and Reuse

The approach proposed with WÆnderer places particular emphasis on the relationship between datasets, models, and musical practice. By enabling musicians to work with artist-defined corpora, the system foregrounds the role of the dataset as a primary creative material. This stands in contrast to many contemporary generative systems, where audio is produced from large-scale datasets that remain opaque to the user and outside their control.

At the same time, the system relies on a pre-trained VAE, itself trained on extensive audio data. In this sense, the approach does not fully escape the limitations associated with large-scale model training, and can be understood as operating within the representational constraints of such datasets. The latent space being navigated reflects statistical structures learned from data that the user does not necessarily own or control. This introduces an inherent tension between the use of personal corpora at the interaction level and the dependence on external data at the model level.

However, by reusing existing models rather than training new ones, the system aligns with perspectives associated with resource-aware or "Small AI" approaches, which emphasize the reuse of pre-trained models, local inference, and reduced computational cost. This strategy allows musicians to engage with complex models without requiring extensive training resources, while also mitigating the environmental impact associated with repeated model training.

6 CONCLUSION

This article has presented WÆnderer, a system that turns personal sound corpora and a pre-trained audio VAE into a real-time navigable musical environment. Audio descriptors organize the corpus into an abstract timbre space; performers move through that space; and the

decoder reconstructs retrieved, corpus-aligned latent frames. Manual, trajectory-based, and morphology-based modes offer different balances of direct control, variation, and temporal organization within the same reconstructive framework.

Our practice-led evaluation establishes a preliminary set of design findings rather than a general user study. Corpus-aligned retrieval avoided distracting artifacts encountered during direct interpolation; decoding-window length exposed a responsiveness-continuity tradeoff; and the three modes produced distinct relationships to source recognizability and formal motion. Structured studies with sound artists are needed to assess usability, appropriation, and creative value beyond the authors' practice.

The project originates from our own artistic practice, within a research-creation framework in which technical development and musical experimentation are closely intertwined. The system was developed not only as a proof of concept, but as a tool intended to be appropriated by other creators. As such, future developments will also consider community-oriented outcomes, including the sharing of corpora, interaction strategies, and potentially model configurations.

From a technical perspective, several extensions are currently under consideration. We aim to explore the integration of timbre transfer functionalities within the proposed framework, leveraging the latent representation to enable transformations between different sound sources while maintaining real-time interaction.

More broadly, this work contributes to ongoing efforts to rethink the role of artificial intelligence in musical practice. By reusing pre-trained models and focusing on interaction rather than generation, the system allows musicians to engage directly with their own sonic materials while benefiting from the representational capacities of large-scale models, without requiring additional training. In this sense, WÆnderer proposes a shift from generative paradigms toward interactive, corpus-based, and reconstructive approaches to sound creation.

ACKNOWLEDGMENTS

The research-creation that produced the WÆnderer system and this article was supported by a grant from the Fonds de recherche du Québec (https://doi.org/10.69777/368430)

REFERENCES

Bell, J. (2023). Maps as Scores: "Timbre Space" Representations in Corpus-Based Concatenative Synthesis. *International Conference on Technologies for Music Notation and Representation (TENOR)*. https://hal.science/hal-04182706

Božić, M., & Horvat, M. (2024). *A Survey of Deep Learning Audio Generation Methods* (arXiv:2406.00146). arXiv. https://doi.org/10.48550/arXiv.2406.00146

Bryan-Kinns, N., Wszeborowska, A., Sutskova, O., Wilson, E., Perry, P., Fiebrink, R., Viglіensoni, G., Lindell, R., Coronel, A., & Correia, N. N. (2025). Leveraging small datasets for ethical and responsible AI music making. *Proceedings of the 20th International Audio Mostly Conference*, 70–81. https://doi.org/10.1145/3771594.3771601

Caillon, A., & Esling, P. (2021). *RAVE: A variational autoencoder for fast and high-quality neural audio synthesis* (arXiv:2111.05011). arXiv. https://doi.org/10.48550/arXiv.2111.05011

Cameron, J., & Blackwell, A. (2026). *Evaluating Latent Space Structure in Timbre VAEs: A Comparative Study of Unsupervised, Descriptor-Conditioned, and Perceptual Feature-Conditioned Models* (arXiv:2603.16713; Version 1). arXiv. https://doi.org/10.48550/arXiv.2603.16713

Coelho, G. (2025). *AI in Music and Sound: Pedagogical Reflections, Post-Structuralist Approaches, and Creative Outcomes in Seminar Practice*. Zenodo. The AI Music Creativity Conference (AIMC). https://doi.org/10.5281/zenodo.16946639

Collins, N. (2025). *Unstable Audio: Code Bending Text-to-Music Generation*. AES International Conference on Machine Learning and Artificial Intelligence for Audio.

Constanzo, R. (2025). *Data Knot* (Version v1.0) [Max]. https://github.com/rconstanzo/data-knot/

Evans, Z., Parker, J. D., Carr, C. J., Zukowski, Z., Taylor, J., & Pons, J. (2024). *Stable audio open*. https://doi.org/10.48550/arXiv.2407.14358

Garber, L., & Ciccola, T. (2021). *AudioStellar* [Computer software]. MUNTREF Centro de Arte y Ciencia.

Ghisi, D. (2017). *Music across music: Towards a corpus-based, interactive computer-aided composition* [Phdthesis, Université Pierre et Marie Curie - Paris VI]. https://theses.hal.science/tel-01880061

Horta Valenzuela, M., & Tomás, E. (2025). Nebula: A PCA-Based Method to Explore RAVE-Encoded Audio Representations. In *Proceedings of the 22nd Sound and Music Computing Conference (SMC2025), Graz, July 2025*. Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz. Sound and Music Computing 2025 (SMC 2025), Graz, Austria. https://doi.org/10.5281/zenodo.15839719

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., & Plumbley, M. D. (2023). AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *Proceedings of the 40th International Conference on Machine Learning*, 21450–21474. https://proceedings.mlr.press/v202/liu23f.html

McAdams, S. (2019). Timbre as a Structuring Force in Music. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 211–

243). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_8

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., & Bengio, Y. (2017). *SampleRNN: An Unconditional End-to-End Neural Audio Generation Model* (arXiv:1612.07837). arXiv. https://doi.org/10.48550/arXiv.1612.07837

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio* (arXiv:1609.03499). arXiv. https://doi.org/10.48550/arXiv.1609.03499

Schwarz, D. (2006). *Real-time Corpus-based Concatenative Synthesis with CataRT*. 9th Int. Conference on Digital Audio Effects (DAFx-06).

Schwarz, D. (2012). *The Sound Space as Musical Instrument: Playing Corpus-Based Concatenative Synthesis*. 250—253.

Schwarz, D., Roebel, A., Yeh, C., & Laburthe, A. (2016). *Concatenative Sound Texture Synthesis Methods and Evaluation*. 217–224.

Scurto, H., & Postel, L. (2023). Soundwalking Deep Latent Spaces. *Proceedings of the 23rd International Conference on New Interfaces for Musical Expression (NIME'23)*. https://hal.science/hal-04108997

Smalley, D. (1997). Spectromorphology: Explaining sound-shapes. *Organised Sound*, *2*(2), 107–126. https://doi.org/10.1017/S1355771897009059

Tahiroglu, K., & Wyse, L. (2024). *Latent Spaces as Platforms for Sonic Creativity*. International Conference on Computational Creativity 2024.

Thibault, D. (2024). Mosaïque—Concatenative Synthesis Instrument for the Practicing Musicians. *AIMC 2024 (09/09 - 11/09)*. https://aimc2024.pubpub.org/pub/buh7kcah/release/1

Thibault, D., Piazza, D., Allard, M., & Arseneault, M. (2025, September 10). *Navigating Abstract Timbre Spaces: Instrumental Affordances of Concatenative Synthesis in Mosaïque*. Zenodo. The AI Music Creativity Conference (AIMC). https://doi.org/10.5281/zenodo.16946851

Viglensoni, G., & Fiebrink, R. (2024). Data- and interaction-driven approaches for sustained musical practices with machine learning. *Journal of New Music Research*, *53*(1–2), 19–32. https://doi.org/10.1080/09298215.2024.2442361

Wang, K., Wu, Z., Zhou, D., Lin, R., Dai, J., & Jiang, T. (2025). *Back to Ear: Perceptually Driven High Fidelity Music Reconstruction* (arXiv:2509.14912). arXiv. https://doi.org/10.48550/arXiv.2509.14912

Yee-King, M. (2022). Latent Spaces: A Creative Approach. In C. Vear & F. Poltronieri (Eds.), *The Language of Creative AI: Practices, Aesthetics and Structures* (pp. 137–154). Springer International Publishing. https://doi.org/10.1007/978-3-031-10960-7_8

Zappi, V., & Tatar, K. (2025). Neural audio instruments: Epistemological and phenomenological perspectives on musical embodiment of deep learning. *Frontiers in Computer Science*, *7*, 1575168. https://doi.org/10.3389/fcomp.2025.1575168

Zheng, S. J., Xambó Sedó, A., & Bryan-Kinns, N. (2025). Exploring gestural affordances in audio latent space navigation. *Frontiers in Computer Science*, *7*. https://doi.org/10.3389/fcomp.2025.1575202

Zheng, S., Sedó, A. X., & Bryan-Kinns, N. (2024). *A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials* (arXiv:2407.04379). arXiv. https://doi.org/10.48550/arXiv.2407.04379