# Murzinograph: Navigating Sound Through Latent Space Visualisations

Gustavo Guzmán
gustavoguzmang@proton.me
Independent researcher
Valparaíso, Chile

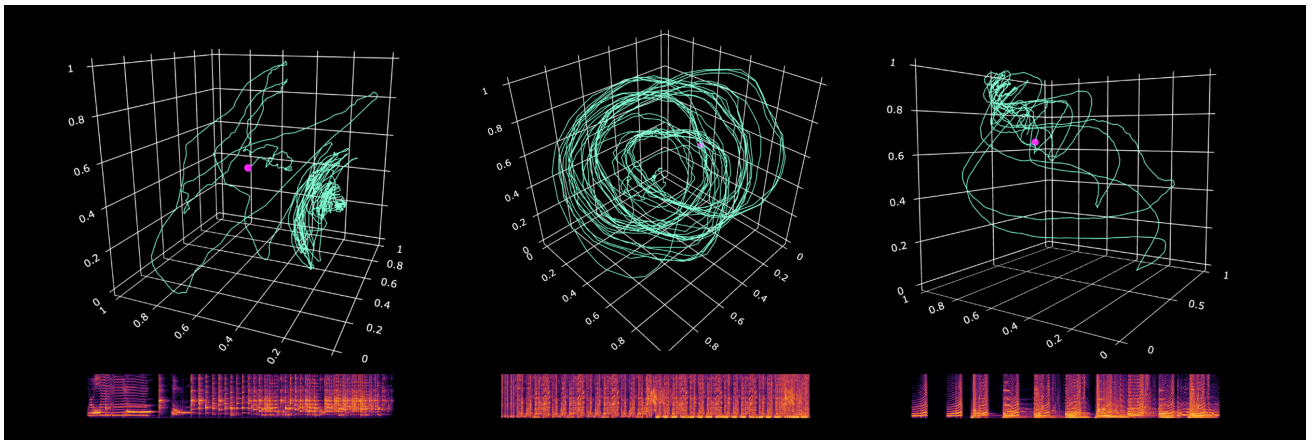

**Figure 1: Murzinograph visualisations above their respective spectrograms. The pink dots show where playback starts.**

## Abstract

This paper presents the *Murzinograph*, a proof-of-concept system that maps audio spectrograms into human-interpretable, three-dimensional visualisations through a bespoke convolutional autoencoder. By constraining latent space dimensionality, the Murzinograph privileges contour-like structures over precise spectral reconstruction, thereby revealing a 'visualisation/reconstruction trade-off' in which less reconstruction fidelity yields more semantically useful manifolds. I examine across musical and bioacoustic datasets, and present a community case study that demonstrates how collective interpretation of visualised acoustic data helps nurture ecological awareness. Furthermore, the paper discusses implications for Interactive Machine Learning and eXplainable AI in NIME contexts, and sketches out avenues toward hybrid and generative model scalings that balance interpretability and synthesis. I conclude by illustrating how this work is positioned within these fields as an accessible tool for creative endeavours, particularly in the context of the Global South.

## Keywords

Sound visualisation, Cross-modality, Interactive Machine Learning, Global South, Community-Based Learning

## 1 Introduction

In light of the recent frenzied developments in the AI industry, autoencoders (AEs) are quite 'old tech' by contemporary standards [9, 14, 15]. And whilst traditionally used for dimensionality reduction and feature learning, today they find many other useful applications. Particularly in music technology, where technical limitations in audio reconstruction quality and duration is still an issue [6], various systems have been developed with some form of autoencoding at their core. However, in this specific context, 'vanilla' AEs have been explored far less than more advanced hybrid architectures, which enable audio sampling and generation via stochastic latent representations typically modelled on Gaussian distributions (e.g. the ubiquitous *RAVE* [3], *BRAVE* [4], etc.).

On the other hand, like Rodrigues [24] mentions, machine learning (ML) "has been engaged in numerous NIME studies, often focused on generating, recognising and manipulating control or gestural data". Particularly in recognition tasks, there seems to be exploratory potential with the more simple AEs. Furthermore, latent spaces bear interesting analogies to conceptualisations of gesture [12, 17, 31], and, as other authors have noted [5, 27], they hold significant pedagogical potential for teaching ML in the context of sound and media arts. It is in this spirit that I present the Murzinograph: a prototype system that not only provides coherent visualisations of audio material, but also seizes opportunities at pattern recognition for data that goes beyond traditional sonic analysis and dimensionality reduction.

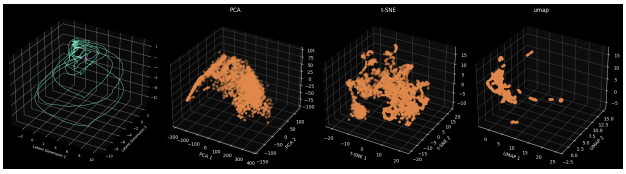


**Figure 2: Murzinograph vis-à-vis PCA, t-SNE and UMAP (notice the lack of temporality in these methods).**

In the following pages, I present the Murzinograph concept, data pipeline and model properties; then proceed to characterise


*NIME '26, June 23–26, 2026, London, UK*

the *visualisation/reconstruction (V/R) trade-off* as an empirical framing that links latent space usability for visualisation with reconstruction fidelity. Afterward, I demonstrate results across diverse sonic materials alongside a case study mapping traffic noise in a peri-urban ecosystem, showing its applied value for socio-ecological and creative practices. Finally, I discuss design prospects for Interactive ML (IML) and Explainable AI for the Arts (XAIXxArts), and provide suggestions for hybrid and generative extensions.

## 2 Murzinograph

### 2.1 A reappraisal of transduction

The Murzinograph is a prototype system for encoding acoustic data into human-interpretable spatiovisual depictions. Named in honour of Soviet audio engineer and inventor of the *ANS* synthesizer (c. 1950-57), Yevgeny Murzin, the Murzinograph has at its core a bespoke convolutional AE designed to provide meaningful representations in a way that is intuitive for many people: three-dimensional space plus time.

More specifically, the model collapses spectrographic information from a wide range of samples (i.e. music, soundscape recordings, animal vocalisations, etc.) into a sequence of data points within an arbitrary *n*-dimensional space. This space—known as the *latent space*—automatically compresses input information by calculating reconstruction accuracy at the opposite end of the network (the *decoder*), thereby producing a consistent depiction of its most relevant components. Because the encoding is deterministic, it yields reliable 3D sketches that can be explored spatially, replayed alongside the original sonic material, and used to delineate salient acoustic structures that are suitable for higher-level analysis.

Arguably, these qualities bode well with developments born from NIME—namely, in terms of providing a structural grasp of aural phenomena by mapping the informational content of acoustic energy contours onto the visual domain. This cross-modal exchange could, in turn, allow for sonic reconstructions derived from visual, somatic, or other types of differentiated source material, as discussed in the following sections.

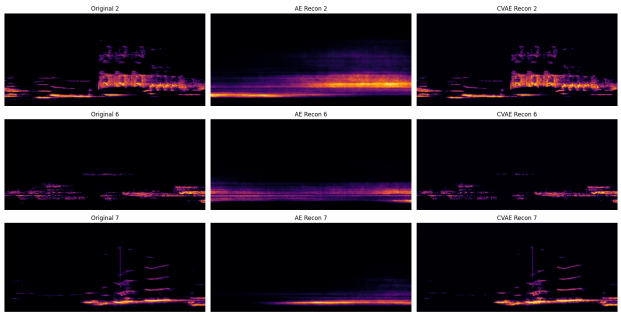


**Figure 3: (left to right) Birdsong: original, low quality recon AE, and high quality recon CVAE (skips).**

Of course, the notion of three-dimensional sound visualisation is neither new nor is there a lack of existing sound visualisers. Lucio Arese's superb experiments in the generative parametrisation of bird vocalisations[1] is one such case that does not rely on neural networks at all, but on the direct examination of features resulting from music information retrieval. However, the novelty in this implementation lies in its practicality: contour structures are automatically generated by extracting salient information that may be overlooked by other analytical or empirical approaches. This allows for the ready visualisation of any type of time-series transforms that one wishes to input to the model, allowing for the sequential embedding of expressive data point trajectories that are consistent with human perception, rather than black-box abstractions of machine representation.

The present research has found that contour mappings derived from AEs systematically perform well under extremely constrained dimensionalities. Notably, that 3 dimensions suffice for both spectrogram reconstruction and latent representation, giving way to experimentation with the latent space itself as a tool for cross-modal investigation. As Schwarz and others have noted [2, 20, 25, 33], these topological representations of data offer valuable insights into the learning aspect of ML itself. Moreover, they can become stand-alone tools for analysing sonic material (i.e. trends, patterns and structure) and facilitate downstream processing. Notable examples of the latter are IRCAM'S *CataRT* and *AudioStellar*, which while not using encoders, do rely on descriptor spaces spread and adjusted over 2- or 3-dimensional manifolds.

## 3 Visualisation and the V/R trade-off

At this point, it should be stressed that, for the purposes of this paper, *visualisation is not the same as reconstruction*, nor is it analogous to the much broader topic of representation. Rather, visualisation pertains here to the compression of the fundamental characteristics of a high-dimensional vector representation of the input for the purpose of human readability. Interestingly, AEs excel at efficient reconstruction [14], particularly when overfitting input data. This is partly due to the fixed, deterministic mapping from the input data to the latent space, which translates to similar inputs yielding similar encodings provided that the model, its parameters and the data pipeline remain unchanged.

Remarkably, a significant compromise emerges in what I term the *visualisation/reconstruction (V/R) trade-off*, which describes the inverse relationship between an AE's reconstruction fidelity and the semantic/structural usefulness of its latent manifold for visualisation. In other words, models pushed toward optimal reconstruction produce compressed encodings that preserve fine detail but yield dense, uninformative visual clusters, whereas relaxing reconstruction accuracy affords a latent space that better exposes meaningful, discriminative structures for visualisation (e.g. trajectorial shifts related to amplitude and harmonic content).

This parallels the *time-frequency uncertainty principle* in classical signal processing, which highlights a fundamental limit on simultaneously localizing a signal in the time and frequency domains—i.e., improving detail in one area limits clarity in the other. This could be partly explained by the fact that the input data is in itself spawned from such signal transforms (i.e. spectrograms). That said, it is also likely that the model architecture itself hits an information bottleneck, akin to Shannon's rate-distortion theory, suggesting that any given channel has a maximum amount of information it can carry. In this case, since the information we are interested in is not the reconstruction fidelity but the latent manifold as a sound visualiser, we consequently deal with this balancing act as a feature in itself.

[1] https://www.lucioarese.net/portfolio_page/visual-birds

## 4 Data processing and model architecture

Generally speaking, AEs do not address data distribution variance and make it challenging to derive meaningful interpolations *between* data points. Therefore, the Murzinograph is fundamentally driven by processes that enhance its ability to learn structural patterns across non-adjacent segments of the source audio, thereby reflecting the temporal unfolding of acoustic features. In order to minimise the *V/R trade-off*, regularisation functions are combined with loss functions—specifically, spectral convergence loss paired with MSE reconstruction—to help compare the original data with its reconstruction in terms that are perceptually meaningful.

Roughly speaking, the system works as follows: spectrograms are windowed into overlapping snapshots stacked along the temporal dimension, frequency ranges are selected and masked, and the dataset shuffled via batched loading. This forces the attention-regularised AE to learn a temporally coherent representation of the data points through convolutional encoding and batch normalisation, which are subsequently sorted for contour visualisation. For reconstruction purposes, this can be notably efficient, especially with properly scheduled U-Net-style skip connections. This efficiency *may* prove significant for future downstream generative tasks, e.g. compared to RAVE, which requires 128-dimensional latent vectors sampled every 20 ms, whilst the Murzinograph uses 3-dimensional latent vectors sampled every 50 frames (~2 ms).

Lastly, although the latent space is constrained to three dimensions, visualisations in practice operate in four: three spatial axes corresponding to the latent dimensions, plus time as an implicit fourth dimension introduced by the sliding window. As a consequence, audio grain features of any duration can be mapped onto sketch-like trajectories that run from beginning to end of the input file during inference. Additional dimensions can be incorporated via analytical pre-processing, alternative signal transforms, or *unsupervised* feature extraction, but the present system yields sufficiently eloquent representations to be considered valuable in themselves.

## 5 Case study: Ravine visualisations

During the prototyping phase, we were invited to partake in a series of workshops with a community[2] living across a ravine facing Chile's central Pacific coast, which is continuously affected by the traffic noise of a major freight highway running from the port city of Valparaíso toward the interior (hereafter the *Ravine Project*[3]). This is therefore an ecotone—a transitional zone where distinct ecosystems and communities intersect, clash, and commingle. In this case, endemic flora and fauna, a working-class neighbourhood, and the aforementioned highway colossus. The idea was to activate a range of methodologies drawn from soundscape research (e.g. soundwalking, deep listening) and alternative technological approaches to convey material entanglements both sonically and visually.

In this setting—four three-hour sessions with a rotating pool of ~12 participants, including children—field recordings from various locations around the ravine were processed with the Murzinograph to produce noise and species distributions across time and frequency. The resulting visualisations offered insights for distinguishing the acoustic signatures of different areas by tracking recurring noise sources, temporal intensity peaks, and

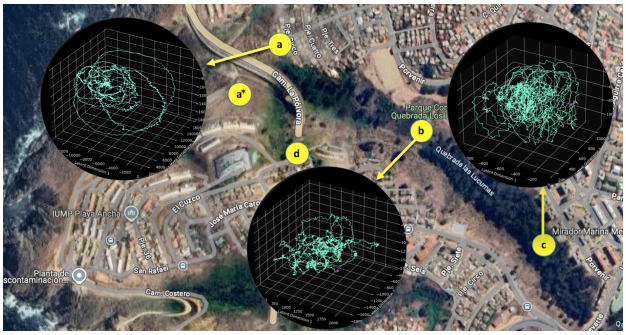


**Figure 4: Map showing the locations where field recordings were made during the workshop (~2 hours total).**

spatially linked events (e.g. passing vehicles, intermittent construction bursts, and transitory bird choruses). This approach emphasised the temporal clustering of high-energy events and uncovered spatially consistent noise patterns across the recording sites.

These shared and solo listening sessions were complemented with free-form sketches and group discussions to ideate mitigation strategies—such as temporally rescheduling noisy activities, installing targeted vegetation buffers, and establishing citizen-driven monitoring checkpoints—showcasing the methodology's capacity to translate complex acoustic information into actionable, easily understood insights for peri-urban resistance.

In this instance, collectively generated interpretations of the data, along with post hoc reflections after each session, were key for cultivating a more refined understanding of our aural entanglements and the sometimes ambiguous roles that sound and noise play in such complex ecotonal scenarios. Likewise, the experience proved successful insofar as it enabled an exploration of "slowness, community, and the old" [11] as valid approaches to technological experimentation, conveying how sound carries the subtle imprints of the environment.

## 6 Embodying the latent

For NIME purposes, visualisation belongs to the wider category of cross-modal systems, for example, of acoustic information (animal vocalisations [22], ecoacoustic data [24], and music [20, 29]) as well as the converse sonification of graphic or pictorial information (e.g. tracing movement contours in 3D and feeding this data into a generative model for sonic output [18]).

In the case of the Murzinograph, its deterministic mechanics and lack of generalisation to unseen data refer to an explicit design decision. That is, the model is not designed to optimize reconstruction (nor compression), but rather to serve as a tool for probing the latent space, in the manner of a digital data transducer. By eschewing expectations of the model's output or any other downstream purpose and instead tethering the *V/R trade-off*, one can ultimately engage explicitly with the latent space as its own goal.

Interestingly, the ongoing refinement of encoder architectures largely depends on their engagement with dimensional structures that remain, to a large extent, intelligible to humans. As several authors have noted [23, 28, 33], as long as these latent spaces encode the information required for their intended purposes, they can also become valuable instruments for IML and XAIXxArts. Within these realms, the Murzinograph approach

[2]The *Quebrada Los Lúcumos* community in the Porvenir neighbourhood.
[3]https://tsonamiediciones.cl/el-ruido-en-la-quebrada

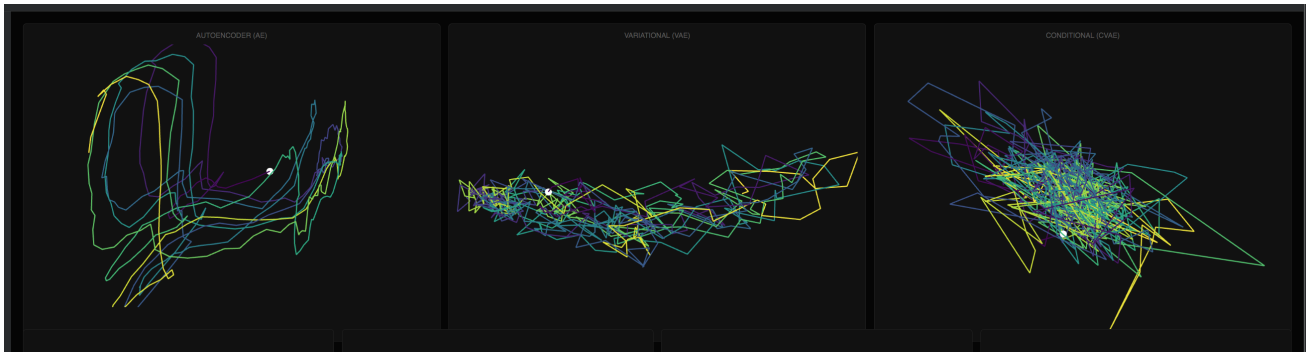


**Figure 5: (left to right) Depth-compressed, time color-mapped latent reconstructions of whale song for AE, VAE, and CVAE.**

can be particularly revealing insofar as they delve into the explanation, uses, and practical deployments of AI that departs from hegemonic corporate logics. In addition, the study of latent spaces may enrich ML-focused pedagogies in the arts and open up stimulating creative possibilities amidst the conceptual terrain of V/R trade-offs, perhaps most notably through the design of 'steering' mechanisms that enable users to traverse these manifolds via gestures or interface-based controls, as other authors have already demonstrated [20].

For example, Vigliensoni & Fiebrink [29] have steered the high-dimensional latent space of the *RAVE* neural audio model by leveraging simple regressive models learned from a set of demonstrative actions. Zheng et al's approach [33], on the other hand, involves using *unsupervised learning* features to encode a human control space based on sketches charted to an audio synthesis model's latent space. Both attempts target the classical 'mapping problem' [10] in NIME literature, this time derived from high-dimensional sensor data automated onto gestural control mechanisms [16].

## 6.1 Designing for IML and XAIXxArts

It is important to remember that not only can the dimensionality of latent spaces be arbitrarily large, but their axes may not necessarily correlate to perceptual labels. In the case of the most popular, generative flavour of autoencoders—variational autoencoders (VAEs)—these representations are also not usually fixed, given their reliance on stochastic modelling and sampling. Moreover, not all dimensions of their latent space will be equally relevant for their compressed representation. These challenges reflect broader issues related to world representation in information theory [19, 32].

Such open questions are particularly promising for IML, as originally proposed by Fails and Olsen Jr. [7], since it relates to processes and methods whereby people interact in an iterative fashion with the model during training to fine-tune, steward or otherwise bias its behaviour. O'Flaherty et al. [21] add that the role of IML "is to find unexpected relationships" between distinct, correlated phenomena. Vigliensoni & Fiebrink [29], for their part, complement that "art- and music-making are non-teleological and purposeless activities in nature, not problems to be optimized," and it is in this spirit that the Murzinograph has been conceived. In other words, to serve specific artisanal and situated purposes, as seen with the *Ravine Project*.

# 7 Discussion

Historically, technocentric scholarship has prioritised complex interfaces and sensor-input systems, often at the expense of usability and the performer's relationship with the medium [8]. This gap remains partly unaddressed, and in the AI era it is common to encounter statements confronting "the well-known low-dimensional human-performance space versus the high-dimensional latent space of a generative audio model."[29] In reality—and quite evidently—human performance is far from being 'low-dimensional.' It appears so only when it is artificially constrained to a limited set of parametric descriptors framed within a specific technical jargon. The persistent 'mapping problem' can therefore be understood as a manifestation of this underlying conceptual misunderstanding.

Moving forward in this domain, the most promising strategies might be hybrid systems that combine data-driven methods with models of the perceptual, embodied, and cognitive mechanisms underlying the human auditory system [13]. As it pertains to the Murzinograph, such hybridisations present a compelling direction to explore. This can lead, however, to latent spaces that do not reflect the broader variability or overall distribution of data. The challenge, then, is to design architectures that yield human-interpretable latent space representations while still being suitable for integration into generative interactive systems.

Consequently, hybrid methods that combine AEs and VAEs might show potential, as they can leverage the strengths of each architecture while minimizing their respective drawbacks. Likewise, Conditional VAEs (CVAEs) could enable the incorporation of supplementary inputs (e.g. user gestures) that modulate the latent space in myriad ways. This would allow for more controlled generation and navigation of the manifold, facilitating the articulation of user preferences contingent on a coherent representation space that is also apt for visualisation. Such an approach may ultimately yield multimodal representations derived from a broader, combined range of input sources (motion sensors, acoustic features, computer vision, etc.).

Alternatively, using an ensemble of models, where one is used primarily for reconstruction and data visualisation while the other dedicates to generative tasks, may also be a fruitful direction for inquiry (RAVE and Eco-Sonic Interfaces [1] are such examples). For bioacoustic or geophysical applications that rely on signal transforms, adding navigation or visualisation capabilities can also shed light on how a model encodes data. This could, in turn, contribute to high-level models of system interactions and improvements on the model's communicative effectiveness.

And this can work both ways: users may input their own choreographies and motion captures into the latent space to investigate other forms of sound-vision relationships [18, 31]—thus, keeping with the legacy of pioneering 'multimodal' systems such as the *ANS* synthesizer.

## 8 Conclusion

For this particular experiment, acoustic information was fed to an AE network via spectrograms, resulting in compelling visual structures between machine representation and reconstruction. These experiments demonstrated that AEs have merit in creative computing by deliberately treating the latent space as a locus for cross-modal auscultation, while hinting at some interesting compromises between representation, visualisation and reconstruction—the *V/R trade-off*. These dynamics are not solely technical, but also point to the limits of our own engagement with the machine, insofar as effective visualisations are inherently an externalisation of embodiment—in this case, through the data collection dynamics and soundscape research methods that shaped the datasets. In this regard, the Murzinograph personifies a contemporary ethos of sustainability through *small data*-derived insights [26, 30] and lo-tech ML approaches aimed at social betterment.

It is typically understood that the success of a ML model hinges on its ability to generalise beyond its training inputs, rather than simply memorising them. But as Fiebrink has noted [8], artistic ML is more or less based upon the transgression of functionality and the non-generalisation of models. Moreover, by recycling an old ML paradigm such as the AE, the Murzinogaph is firmly framed in Gaster et al.'s [11] motto of "pushing against the driving force of economic structures, which emphasises a continuous cycle of replacement, musicians and instrument designers".

Its focus on latent-space exploration inherently sparks gesture-driven interaction workflows familiar to NIME practitioners, while its community case work with the Ravine Project demonstrates how sonic visualisation can inform situated and socially engaged practices. Thus, the Murzinograph contributes to NIME discourse by providing concise, easily understandable correspondences between sound and three-dimensional visualisations, which may be directly applied in aural pedagogy, instrument design, and—with further developments—music composition.

Lastly, the Ravine Project approach can be extended to additional use cases where one aims to leverage the expressiveness of deep learning models to systematically assess which data properties carry semantic significance. Where soundscape studies and long-term bioacoustic monitoring are concerned, the Murzinograph demonstrates consistent ability to reveal relationships within the data more succinctly than other options, such as false-colour spectrograms, which can be constrained by temporal resolution and scale.

## 9 Ethical Standards

This work complies with the NIME code of ethics and does not include any experiments involving human or animal subjects. The datasets used to train the ML models in this study were obtained either from open-access repositories, from private music archives or from proprietary collections gathered during workshops, with the informed consent of all participants.

## Acknowledgments

I would like to thank Camila, Teo and Patricia for their ongoing support, and the entire LATAM NIME community for their commitment to amplifying Latin American voices in this beautiful 'dawn chorus.' I am also very grateful to Ableton for granting me the opportunity to partake and contribute to this gathering.

## References

[1] Sabina Hyoju Ahn, Ryan Millett, and Seyeon Park. 2025. Eco-Sonic Interfaces for Embodied AI Sound Exploration. (2025).
[2] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N. Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. 2023. Exploring XAI for the Arts: Explaining Latent Space in Generative Music. https://arxiv.org/abs/2308.05496 _eprint: 2308.05496.
[3] 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 Version Number: 2.
[4] Franco Caspe, Andrew McPherson, and Mark Sandler. 2025. Waveform Autoencoding at the Edge of Perceivable Latency. (June 2025). https://doi.org/10.5281/ZENODO.15698791
[5] Xinran Chen, Iurii Kuzmin, Mela Bettega, and Raul Masu. 2025. Embodying Sustainability: Paving Opportunities for NIME Research. (June 2025). https://doi.org/10.5281/ZENODO.15698845
[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. https://doi.org/10.48550/arXiv.1704.01279 arXiv:1704.01279 [cs].
[7] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, Miami Florida USA, 39–45. https://doi.org/10.1145/604045.604056
[8] Rebecca Fiebrink and Laetitia Sonami. [n. d.]. Reflections on Eight Years of Instrument Creation with Machine Learning. ([n. d.]).
[9] P. Gallinari, Yann Lecun, Sylvie Thiria, and Francoise Soulie Fogelman. 1987. Mémoires associatives distribuées: une comparaison (distributed associative memories: a comparison).
[10] Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. 2025. Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations. https://arxiv.org/abs/2412.08550 _eprint: 2412.08550.
[11] Benedict Gaster, Nathan Renney, and Jasmine Butt. 2025. Looping slowly: Diffraction through the lens of nostalgia. (June 2025). https://doi.org/10.5281/ZENODO.15698774
[12] Gustavo Guzmán. 2023. *El gesto transmedial: el fenómeno gestual a través de distintos medios de expresión artística.* Maestría en Música (Tecnología Musical). Universidad Nacional Autónoma de México, Programa de Maestría y Doctorado en Música. https://hdl.handle.net/20.500.14330/TES01000848366
[13] Laurie M. Heller, Benjamin Elizalde, Bhiksha Raj, and Soham Deshmukh. 2023. Synergy between human and machine approaches to sound/scene recognition and processing: An overview of ICASSP special session. https://doi.org/10.48550/arXiv.2302.09719 arXiv:2302.09719 [eess].
[14] G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (July 2006), 504–507. https://doi.org/10.1126/science.1127647
[15] Yann Lecun and Francoise Soulie Fogelman. 1987. Modeles connexionnistes de l'apprentissage. *Intellectica, special issue apprentissage et machine* 2 (Jan. 1987). https://doi.org/10.3406/intel.1987.1804
[16] Sebastian Löbbers, Louise Thorpe, and György Fazekas. 2023. SketchSynth: Cross-Modal Control of Sound Synthesis. In *Artificial Intelligence in Music, Sound, Art and Design*, Colin Johnson, Nereida Rodríguez-Fernández, and Sérgio M. Rebelo (Eds.). Vol. 13988. Springer Nature Switzerland, Cham, 164–179. https://doi.org/10.1007/978-3-031-29956-8_11 Series Title: Lecture Notes in Computer Science.
[17] Guerino Mazzola and Moreno Andreatta. 2007. Diagrams, gestures and formulae in music. *Journal of Mathematics and Music* 1, 1 (2007), 23–46. https://hal.science/hal-01161060 cote interne IRCAM: Mazzola07a.
[18] Sarah Nabi, Philippe Esling, Geoffroy Peeters, and Frédéric Bevilacqua. 2024. Embodied exploration of deep latent spaces in interactive dance-music performance. In *Proceedings of the 9th International Conference on Movement and Computing*. ACM, Utrecht Netherlands, 1–9. https://doi.org/10.1145/3658852.3659072
[19] Ivars Namatēvs, Artūrs Ņikuļins, Anda Slaidiņa, Laura Neimane, Oskars Radziņš, and Kaspars Sudars. 2023. Towards Explainability of the Latent Space by Disentangled Representation Learning. *Information Technology and Management Science* 26 (Nov. 2023), 41–48. https://doi.org/10.7250/itms-2023-0006
[20] Ashley Noel-Hirst, Charalampos Saitis, and Nick Bryan-Kinns. [n. d.]. Sampling the Latent Space: Exploring the Creative Potential of Generative AI Through the Lens of Sample-Based Music Making. ([n. d.]).
[21] Tug F O'Flaherty and Luigi Marino. 2025. Sonicolour: Exploring Colour Control of Sound Synthesis with Interactive Machine Learning. (2025).
[22] Orr Paradise, Pranav Muralikrishnan, Liangyuan Chen, Hugo Flores García, Bryan Pardo, Roee Diamant, David F. Gruber, Shane Gero, and Shafi

Goldwasser. 2025. WhAM: Towards A Translative Model of Sperm Whale Vocalization. https://doi.org/10.48550/ARXIV.2512.02206 Version Number: 1.
[23] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From Explainable to Interactive AI: A Literature Review on Current Trends in Human-AI Interaction. https://doi.org/10.48550/ARXIV.2405.15051 Version Number: 1.
[24] Frederick Rodrigues. 2025. Synthetic Ornithology: Machine learning, simulations and hyper-real soundscapes. *Machine learning* (2025).
[25] Diemo Schwarz. 2012. The Sound Space As Musical Instrument: Playing Corpus-Based Concatenative Synthesis. (June 2012). https://doi.org/10.5281/ZENODO.1180593
[26] Yann Seznec. 2025. The Memory Cloud: Personal media libraries as affordance and constraint. (2025).
[27] Notto J. W. Thelle and Bernt Isak Wærstad. 2023. Co-Creatives Spaces: The machine as a collaborator. (May 2023). https://doi.org/10.5281/ZENODO.11189170
[28] Laia Turmo Vidal, Ana Tajadura-Jiménez, and Judith Ley-Flores. 2025. Temporal Trajectories: Characterizing Somatic Experiences that Unfold Over Time. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. ACM, Madeira Portugal, 2931–2949. https://doi.org/10.1145/3715336.3735777
[29] Gabriel Vigliensoni and Rebecca Fiebrink. 2023. Steering latent audio models through interactive machine learning. (June 2023). https://doi.org/10.5281/ZENODO.8087978
[30] Gabriel Vigliensoni, Phoenix Perry, and Rebecca Fiebrink. 2022. A small-data mindset for generative AI creative work. (May 2022). https://doi.org/10.5281/ZENODO.7086327 Version Number: V2.
[31] Federico Ghelli Visi and Atau Tanaka. 2020. Interactive Machine Learning of Musical Gesture. https://doi.org/10.48550/ARXIV.2011.13487 Version Number: 1.
[32] Karn N. Watcharasupat and Alexander Lerch. 2021. Evaluation of Latent Space Disentanglement in the Presence of Interdependent Attributes. https://doi.org/10.48550/ARXIV.2110.05587 Version Number: 1.
[33] Shuoyang Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2024. A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials. https://doi.org/10.48550/arXiv.2407.04379 arXiv:2407.04379 [cs].