Soundwalking Deep Latent Spaces

Hugo Scurto, Ludmila Postel

Soundwalking Deep Latent Spaces

Hugo Scurto
EUR ArTeC / Paris 8 / EnsadLab
Paris, France
hugo.scurto@ensad.fr

Ludmila Postel
Locus Sonus Vitæ / ESAAIX
Aix-en-Provence, France
ludmilapostel@gmail.com

ABSTRACT

This paper relates an early art-research collaboration between two practitioners in machine learning and virtual worlds toward new embodied musical experiences of Artificial Intelligence (AI). Instead of a digital music instrument or a music-generating agent, we propose to craft a soundwalk experience where a human person moves through a three-dimensional virtual world to explore a latent sound space generated by deep learning. We report on the diffractive prototyping and iterative crafting of three such soundwalks through/out deep latent spaces, using *nn∼* and *New Atlantis* as computational platforms for AI audio processing and virtual world experimentation. We share critical perspectives emerging from our latent soundwalking practice, with the hope that they contribute to ongoing community-wide reflections toward new AI for musical expression.

Author Keywords

AI, Soundwalk, Generative Deep Learning, Virtual World.

CCS Concepts

•**Applied computing** → **Sound and music computing;**

1. INTRODUCTION

Deep generative models are high-dimensional parametric representations of large datasets computed by deep learning. In the sound and music domain, they typically enable to generate unheard-of sounds, by exploring a so-called latent parametric space, which expressively interpolates sounds contained in the original dataset. The versatility of deep generative models enabled the creation of a wide range of new interfaces for musical expression, including gestural music systems [7, 5], autonomous generative agents [17], or stream-based sound installations [10]. Such musical interfaces in turn contributed to the modern hype surrounding music Artificial Intelligence (AI) [6], despite more nuanced reflections from NIME practitioners [16, 9].

NIME'23, 31 May–2 June, 2023, Mexico City, Mexico.

In response to this modern hype, several academics expressed the need to develop critical practices and representations of AI, both within and beyond the NIME community [8, 2]. Sound and music researchers elicited ethical issues raised by corporate scraping of music datasets used for deep generative models, which are often hidden behind commercial discourses on creative potentials of music AI [15]. Scientists shed light on the vast energy consumption and greenhouse gas emission of deep generative models, whose unprecedented performances often imply high computational costs [4]. Writers have revealed the planetary computational infrastructures that support modern AI applications, and how they perpetuate an extractivist agenda that transcends bodies, spaces and societies [3].

As a duo of artists-researchers, we are interested in exploring soundwalking as a joint musical interface for deep generative models and critical practice of AI. Soundwalking consists in an individual or collective walk with a focus on listening to the environment [18]. It has joint pedagogical, political and artistic implications, including educating persons to listen to uncanny sounds, giving voice to places and bodies that have been muted throughout history, or articulating musical performances throughout such situated soundscapes [13]. We believe that soundwalking could help reveal social, political and planetary issues raised by modern AI applications, while suggesting novel practices of deep generative models within and beyond NIME [12].

In this paper, we relate early conceptual and technical work toward soundwalking deep latent spaces. Specifically, we leveraged our respective expertise in machine learning and virtual worlds to craft a soundwalk experience where a human person moves through a three-dimensional virtual world to explore a deep latent sound space. Rather than engineering deep learning to map walking movement with sound [9], we adopted a diffractive approach to machine learning [11], using existing computational platforms to prototype and iteratively craft such musical experiences while probing them together. Our wish is that such latent soundwalks may help reveal novel representations for AI that better attend to socio-political aspects of music practices.

The next sections successively describe the computational platforms and our musical experiences. We end by sharing critical perspectives raised by our latent soundwalks, with the hope that they contribute to ongoing community-wide reflections toward new AI for musical expression.

2. COMPUTATIONAL PLATFORMS

This section describes *nn∼* and *New Atlantis*, the two computational platforms that we used to iteratively craft our musical experiences, along with our current working prototype, whose workflow is shown in Figure 1.

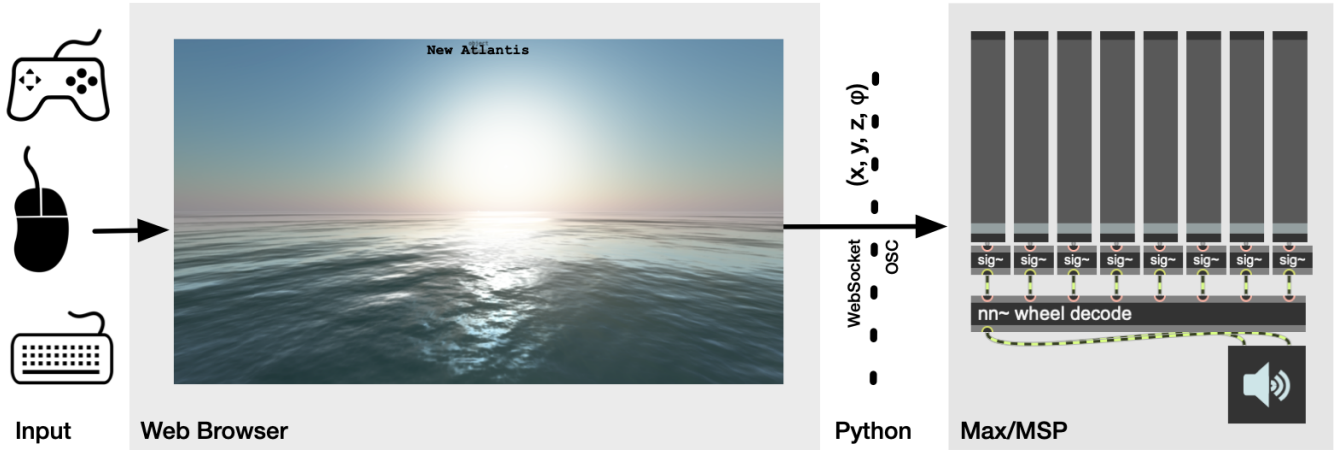

Figure 1: Our current prototype workflow for soundwalking deep latent spaces.

2.1 nn~

nn~ is a Max/MSP and PureData external for AI audio processing[1]. It enables to easily explore deep latent spaces and play with audio streams they generate. The current version works in real-time over standard CPUs, which makes it more affordable than other GPU-based deep learning platforms for music [16], while keeping energy costs standard.

Different interaction modalities can be used as input with *nn~*. The first is latent parameters, which consists in generating sound in a way similar to standard sound synthesisers, *i.e.,* by exploring different parameter combinations to produce different sounds. The second is live audio: it consists in reconstructing audio coming either from a live stream or an offline audio file by transferring timbres that are contained in the deep generative model. The third one is unconditional generation, which consists in autonomously producing an audio stream based on temporal probabilistic modeling of the deep generative model.

The versatility of *nn~* enabled the creation of diverse music AI experiences, including human-machine live improvisations, infinite audio livestreams over the Web, or sample-based offline music composition. Yet, to the best of our knowledge, it has never been used to create musical experiences based on soundwalking, nor have been other AI platforms or deep generative models for sound and music.

2.2 New Atlantis

New Atlantis is a web-based virtual platform dedicated to sound art experimentation[2]. It enables to compose sound spaces through a three-dimensional virtual world, and to explore them through individual or collective walking. The latest version consists of a web-based environment, which makes it usable on both standard CPUs and smartphones, without requiring specific commercial software or hardware.

Different interaction modalities can be used to compose and explore soundwalks in New Atlantis. Composing consists in positioning sound samples in the three-dimensional space, and adding virtual boxes to create reverberation and filtering effects, following acoustic rules of the physical world. Exploring typically consists in freely navigating through sound spaces by walking and flying with varying speed, using either keyboard and mouse or a USB gamepad. Virtual objects can also be added as spatial landmarks suggesting specific soundwalks and atmospheres to the audience.

New Atlantis has been used in artistic and pedagogical contexts, ranging from sound art installations and experimental music performances to workshops on computational and sound art. Yet, to the best of our knowledge, it has never been used to soundwalk deep latent spaces, nor have been other three-dimensional platforms or virtual worlds dedicated to artistic creation.

2.3 Current Prototype

Our current prototype sought to link New Atlantis with *nn~* to craft soundwalks through/out deep latent spaces. Within New Atlantis, we implemented a Javascript script that get the camera's position (x, y, z) and azimuthal orientation (φ) in the virtual world every 100 ms, and sends it locally through a WebSocket connection. Within our CPU-based machine, we implemented a Python script that receives data from the WebSocket connection, and sends it locally to Max/MSP using the OSC protocol. Future implementation may rely on Node for Max and the jweb object.

On the generative deep learning side, we relied on RAVE[3], a state-of-the-art deep generative model for raw audio waveforms [1]. Two advantages of RAVE over other deep generative models are that audio waveforms are modeled at a 48 kHz sampling frequency, *i.e.,* standard audio quality, and that they are generated in real-time with low latency, *i.e.,* less than one second on a standard CPU. One disadvantage is that audio generation can be perceived as choppy due to RAVE's architecture based on waveform grains [1]. Future work may explore deep generative models for raw spectrogram generation to achieve smoother sound output.

3. MUSICAL EXPERIENCES

In this section, we describe three musical experiences that we iteratively crafted based on our current prototype. The first and second authors leveraged their respective expertise in machine learning and virtual soundwalks to implement and test these experiences. In each of these experiences, the second author was not told the type of latent sound space at stake: rather, the first author only told them to explore the deep latent space by freely soundwalking the three-dimensional space, using a keyboard and mouse, as long as they wanted to. Extracts of these three musical experiences are shown on a supplementary video material created afterwards by the first author[4].

[1]https://github.com/acids-ircam/nn_tilde
[2]https://jonlab.github.io/NewAtlantisWeb/index.html
[3]https://github.com/acids-ircam/RAVE
[4]https://vimeo.com/814921401

3.1 Walking Into Latent Sounds

The first experience sought to create unexpected encounters with voices and sonic bodies of a deep latent space. The first author used a RAVE deep generative model pre-trained on the VCTK dataset—monophonic English voice data [1], which expressively interpolates between latent noises and human voices. They crafted a one-to-one mapping between the camera's three-dimensional position in New Atlantis to the three first latent parameters of *nn~*, leaving remaining parameters at a static null value. The sole visual landmark was one small sphere placed at the origin of the space.

After a few minutes of careful exploration, the second author commented that the experience felt like *"a new game for [them]"*. Specifically, the fact that the camera's movement was mapped to latent parameters created the feeling to be *"within a mass of sound"*. This differs from previous experiences with New Atlantis, which often reproduce acoustic properties of the physical space using located sound sources. Also and interestingly, they did not perceive our current prototype's low latency during the soundwalk.

Despite enthusiasm, the experience destabilised the second author, who felt a bit lost in such an abstract latent space. As a consequence, they sought to link each three dimensions of the virtual space to timbre variations of sound, based on linear displacements relative to the spherical visual landmark. Such an approach was suggested by how space was designed in New Atlantis: *"From the moment that there is only one point of reference, I turn around it"*.

Still, listening alone did seemed to provide the second author with a few landmarks within the latent space. First, they sought to find uncanny voices contained in the latent space: *"In fact, there is silence and there are voices, and since we have no other visual clues, we try to find such sounds by ear."* Then, they discovered extreme values of the latent space through their corresponding noisy continuums: *"when I get to white noise, if I continue to move forward there's no more evolution, so that means it's a bit of an invisible barrier to the world you've built."*

3.2 Embodying Listening With 3D Landmarks

Based on this feedback, the first author crafted a second experience by adding visual landmarks to the same latent sound space using New Atlantis. Specifically, they added ellipsoid objects of different sizes, colors and relative distances, to create a three-dimensional space that encourages both slow- and fast-paced movements. Their positioning related to latent sound space was deliberately incidental: landmarks could just as easily indicate timbre variations, as be placed in the middle of a static noise continuum.

While remaining abstract, such visual landmarks suggested a way of moving within the latent sound space, which seemed to have cultivated embodied listening along the soundwalk experience: *"With the spheres, I understand space very well, so it really allows me to listen much more, and to try to move in relation to what I hear, which is really cool"*. Interestingly, the second author spent much more time soundwalking this experience, even if the latent sound space and mapping were the same as in the first experience.

From an aesthetic perspective, the experience seemed to relate to an uncanny soundscape that entangle recognisable sonic bodies with noisy continuums: *"There are moments of transition which are very beautiful [...]. There are times when you slowly come close to one or several spheres, and you first recognise voice, then it slowly evolves into something else... There's something quite soaring, a little bit like ambient music, something that makes you want to let yourself be carried in the evolution of sound"*.

3.3 Sensing Deep Learning Spatialities

The first author crafted a third experience based on this embodied listening feedback. They used a stereophonic RAVE deep generative model [12], trained on a custom soundscape dataset collected over the planet [14]. In addition to previous position mapping, they implemented a one-to-one mapping between the camera's azimuthal orientation and the stereophonic speaker position, using the Spat library. They removed all visual landmarks but the spherical origin to experiment with this stereophonic latent sound space.

The stereophonic aspect of this deep latent space was subtle, yet perceivable by the second author: *"There are much more things that take place in the sound space itself, before I even move. And there are moments when I arrive at places where I can really hear a right-left"*. This sonic depth helped the second author locate themself within the latent space: *"My feeling is that there are hinge points that are not visual like the spheres, but that are sonorous, which gives the impression to pass from one place to another, contrary to previous experiences, which felt like one single place"*.

Interestingly, spatial properties of this latent sound space seemed to differ from those of the physical space: *"When you move towards a sound, you don't actually get closer [...] but it does change the stereo"*. While very preliminary, this uncanny attribute of stereophonic latent spaces may pave the way for music AI practices that focus on spatialities, as the second author suggested: *"there are really moments, or places—it's funny that I just said moments to express how very written it feels, in fact—there are places where something particular happens"*.

4. CRITICAL PERSPECTIVES

In this section, we share critical perspectives emerging from our early art-research collaboration, looking at latent soundwalk against music AI, and at deep latent spaces against virtual worlds, with the hope that it contributes to community-wide reflections toward new AI for musical expression.

4.1 Music AI as Latent Soundwalk

The one-to-one mappings crafted for our musical experiences might seem quite limited from a technical perspective. Yet, we believe they enabled to reveal subtle differences between soundwalking and musical gestures as movement-sound relations. While musical gestures typically seek to expressively control sound within relatively short time scales of music performance, soundwalking seeks to reveal sounds of our socio-material environments within larger scales of time and space, thus implying different embodied listenings.

While these two practices are not mutually exclusive, we suggest that future work might delve more into soundwalk-inspired mappings between movement and latent sound spaces. For example, latent parameters mapped to camera position could change over different time scales to simulate soundscape evolution over climatic cycles of the planet. Or, displacements in specific parts of the space could trigger abrupt changes in latent parameter mappings to simulate alterations of the sonic environment based on human actions.

On a conceptual level, we suggest that latent soundwalks could inspire how we may practice music AI. Rather than engineering AI to autonomously perform human musical tasks [15], practitioners might design interfaces that support encountering uncanny sounds through embodied exploration of deep latent spaces [16]. We argue that such embodied interfaces might foster new AI practices that do not impinge on existing musical labor [8], similar to how soundwalk fosters ethical awareness of our environments [18].

4.2 Deep Latent Spaces as Virtual Worlds

The computational platforms used to craft our latent soundwalks enabled to experiment with spatial landmarks as guides for music AI. Specifically, we found that visual landmarks seemed to deepen embodied listening throughout latent sound spaces in relation to walking in a three-dimensional space. While such insights may be common for virtual world and soundwalk practitioners, we believe they may be novel for deep generative models practitioners.

Future work may craft three-dimensional worlds in greater detail to create narratives through latent soundwalks, for example using concrete objects, spaces and textures. Moreover, collective modes of soundwalking could also give shape to such latent virtual worlds, relying on avatars to walk through space, and providing social landmarks along soundwalks. We will build on New Atlantis' web standards to craft such collective musical experiences, by implementing online, live audio streaming from deep generative models.

More conceptually, we suggest that virtual worlds could be used as a metaphor for deep latent spaces. Indeed, deep latent spaces have deeper social and aesthetic implications than standard synthesisers or timbre spaces, through the culturally-situated datasets and human work they build on [3]. Yet, commercial discourses that anthropomorphise AI tend to hide this complexity [15], along with its ecological implications [4]. We believe that describing deep latent spaces as virtual worlds may better highlight the spatiality of AI—its planetary corporate infrastructures and sociomaterial resources [12]—, and how these may transform the spaces where musicking and listening take place.

4.3 Toward New AI for Musical Expression

Our art-research collaboration had us find compromises between technical engineering of deep generative models and musical experience design. For example, the first author did not perform dimensionality reduction of deep latent spaces [16], which tends to produce parameters that are interpretable from a sound perception perspective—*e.g.,* spectral centroid evolving linearly along one parameter. Rather, they opted for elementary mappings between latent parameters and New Atlantis to let the second author experience the raw sonic materiality of deep latent spaces through their virtual soundwalk practice, and interpret latent parameters from a bodily, sensitive perspective.

Along with other artists-researchers [11, 16, 2], we believe that diffracting AI throughout musical expression—*e.g.,* attending to one's subjectivity and responding to more-than-human materialities of musical interfaces—may produce situated knowledge on deep generative models that complements that produced by technical engineering [12]. This is in line with ongoing community-wide reflections toward ethical and epistemological turns for AI [8]. In future work, we will further entangle NIMEs, soundwalk and AI, by leading latent soundwalks within physical public spaces, using smartphones as joint localisation and microphone devices for a responsive, hybrid listening experience [13]. We believe that such collective practices of deep generative models will be key to reveal AI actions over our environments and ethically reconfigure our representations of AI.

5. CONCLUSION

This paper related early art-research toward soundwalking deep latent spaces. We look forward to critical perspectives from other NIME practitioners to reconfigure deep generative models in more depth and entangle a plurality of subjectivities within music AI.

6. ACKNOWLEDGMENTS

We thank Axel Chemla–Romeu-Santos, Antoine Caillon, Peter Sinclair, Jonathan Tanant, Emanuele Quinz, Samuel Bianchini, as well as reviewers for their enriching feedback.

7. ETHICAL STANDARDS

The speech and soundscape datasets used for training are open-source [1, 14] and were fully anonymised in terms of person and place identities. We know of no way to regenerate one's voice or soundscape from the trained models.

8. REFERENCES

[1] A. Caillon and P. Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

[2] B. Caramiaux and S. Fdili Alaoui. " explorers of unknown planets" practices and politics of artificial intelligence in visual arts. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.

[3] K. Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[4] C. Douwes, P. Esling, and J.-P. Briot. Energy consumption of deep generative audio models. *arXiv preprint arXiv:2107.02621*, 2021.

[5] C. Erdem, B. Wallace, and A. R. Jensenius. Cavi: A coadaptive audiovisual instrument–composition. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2022.

[6] S. Knotts and N. Collins. A survey on the uptake of music ai software. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 499–504, 2020.

[7] C. P. Martin, K. Glette, T. F. Nygaard, and J. Torresen. Understanding musical predictions with an embodied interface for musical machine learning. *Frontiers in Artificial Intelligence*, 3:6, 2020.

[8] F. Morreale. Where does the buck stop? ethical and political issues with ai in music creation. *Transactions of the International Society for Music Information Retrieval*, 4(1), 2021.

[9] T. Murray-Browne and P. Tigas. Latent mappings: Generating open-ended expressive mappings using variational autoencoders. *arXiv preprint arXiv:2106.08867*, 2021.

[10] H. Scurto. ciel. In *Proceedings of the 20th International Conference on New Interfaces for Musical Expression (NIME 2020)*, 2020.

[11] H. Scurto, B. Caramiaux, and F. Bevilacqua. Prototyping machine learning through diffractive art practice. In *Designing Interactive Systems Conference 2021*, pages 2013–2025, 2021.

[12] H. Scurto and A. Chemla-Romeu-Santos. Deeply listening through/out the deepscape. In *International Symposium on Electronic Art (ISEA 2023)*, 2023.

[13] T. Shaw and J. Bowers. Ambulation: Exploring listening technologies for an extended sound walking practice. In *New Interfaces for Musical Expression*, pages 23–28, 2020.

[14] P. Sinclair. Locus stream open microphone project. In *Proceedings of the 2018 International Computer Music Conference (ICMC 2018)*, 2018.

[15] J. Sterne and E. Razlogova. Tuning sound for infrastructures: artificial intelligence, automation, and the cultural politics of audio mastering. *Cultural Studies*, 35(4-5):750–770, 2021.

[16] K. Tahiroğlu, M. Kastemaa, O. Koli, et al. Ai-terity 2.0: An autonomous nime featuring ganspacesynth deep learning model. In *International Conference on New Interfaces for Musical Expression*, 2021.

[17] G. Vigliensoni, L. McCallum, E. Maestre, R. Fiebrink, et al. R-vae: Live latent space drum rhythm generation from minimal-size datasets. *Journal of Creative Music Systems*, 1(1), 2022.

[18] H. Westerkamp. Soundwalking. *Sound Heritage*, 1974.