Latent Spaces as Platforms for Sonic Creativity

Koray Tahiroğlu
Department of Art and Media
Aalto University

Lonce Wyse
Music Technology Group
Univeritat Pompeu Fabra

Abstract

Music and sound generation are among the many areas being disrupted by neural network (NN) technology which often play a creative partner role with humans. One of the architectural features of such neural networks that has been attracting the attention of creative audio practitioners is the latent space. In this paper, we provide a provocative review and discussion of Generative Adversarial Network (GAN) latent spaces as a platform for *active divergence* and focus on understanding the relationship between GAN latent spaces and various constructs in established musical theory and historical practices. We also discuss new musical affordances provided by latent space and its connections to computational creativity and co-creative systems. We argue that the GANs' relationship to certain musical practices is problematic for exploitation in real-time performance contexts. We discuss two alternative ways in which these challenges have been addressed in support of music making, and finally how some contemporary musicians are using latent spaces in generative systems in their own creative practices.

Introduction

The concept of a latent space is fundamental to many of the neural networks (NNs) architectures in generative modeling, where it enables the generation of new data that shares features with the training data set. This is particularly evident in generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Latent space can offer insights into the inner workings of neural networks. Creative practitioners can explore the learned representations in the latent space to gain a better understanding of what features or factors the model considers important for its generation and predictions. In this paper, we explore GAN latent spaces and their transformative potential in the practice of music. We seek to understand and unravel the relationship between the latent space and the concept of trajectories in musical contexts. Specifically, we investigate how the latent space can be viewed in terms of active divergence (Berns and Colton (2020), Broad et al. (2021)), making connections to historical musical representation, improvisation, and the exploration of novelty as well as reflecting on its potential as support for computational creativity and co-creative systems.

In this context, current large Language models in text-to-music systems (Yang et al. (2023); Kreuk et al. (2022); Liu et al. (2023a); Huang et al. (2023b,a); Agostinelli et al. (2023); Dong et al. (2023); Yuan et al. (2023); Liu et al. (2023b)) often work by mapping text and audio into a common latent space (e.g. the CLAP model, Elizalde et al. (2023)). While leveraging latent spaces in this way provides for a novel means of creating music through text, it bears little resemblance to historical music creation, nor is it useful in a real-time performance context. As the adage goes, "talking about music is like dancing about architecture." The focus of this paper is on other approaches to latent spaces which offer far more nuanced control over instrument-like sound generation (timbre and pitch) and instrument-like real-time interaction and have a closer connection to historical music practices.

Latent Space

The term "latent" in "latent space" refers to the hidden or underlying structure present in the data - the hidden factors that explain or determine the observable data. "Space" refers to the typical way the factors are represented as a set of numerical values, each interpreted as a coordinate in a different dimension. The music theoretician Leonard Meyer makes the connection between traditional music and navigable spaces noting that "cultures all over the world tend to characterize pitches in spatial terms" (Meyer (1967)). Other musical dimensions can be similarly understood. If we are interested in a set of sounds, for example saxophone notes, that can all be expressed as a combination of pitch and volume, then each point in the 2D space provides the information necessary to generate a different sound, and a sequence of points in the space would represent an expressive musical melody.

When we train GANs for generative audio, the space of sounds that will be available for navigation is largely determined by the data set and learning objective that we use for training. Typically, most sounds in the training data set can be approximately generated by some particular point in the latent space, and generated sounds that are perceptually close are close to each other in the latent space. However, a data set is finite and countable, but the latent space is made of either continuous dimensions or quantized to a set of points that can represent many more sounds than are contained in the training set. This suggests that the points be-

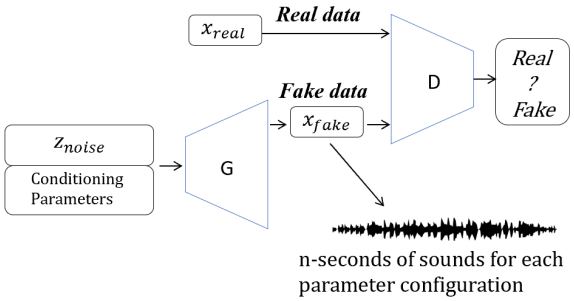

Figure 1: During training, the input to a GAN is a random vector (z) (and optionally conditioning parameters) which comprise the latent space that the generator G learns to map to chunks of audio using feedback from the discriminator D.

tween and among those that generate approximations to data set examples are "novel" in some sense sounding like they are between examples in the training data set. The learned structure in latent spaces together with the access to novelty has profound creative musical implications that will be developed and discussed throughout the paper.

Mechanisms of Latent Space

The co-creative design process starts with the choice of training data. GANs (Figure 1) learn to organize the space of the input vectors so that they represent the "hidden factors" that produce outputs that model the distribution of the collection of sounds comprising the data set. A more extensive variety of sounds in the audio data set leads to the clustering of similar timbre characteristics together with fast-changing boundaries between them. In contrast, if the data set is built on audio samples with similar timbre characteristics, the resulting latent space we be organized more smoothly with uniform or slowly changing timbral features.

The objective functions define the goals and criteria that drive weight adjustment, or "learning", during the training of neural networks. GANs often use the Wasserstein metric (Arjovsky, Chintala, and Bottou (2017)), a kind of "earth mover" comparison of distributions of training and generated data to achieve the smooth structure in the mapping from latent to output features and (at least locally) the preservation of the association of latent space direction with meaningful transformations.

A typical latent space learned by a GAN, in this case the input layer, might have 128 dimension. While 128 might be "small" compared to the "space" of the media, it is still a lot of parameters for audio generation, certainly more than an instrumentalist typically controls in real-time. How then can we know what points to choose in order to generate sounds with specific characteristics we desire for our musical purposes, or what directions to move to change sound in a particular way? Every data set will create a different latent space map, and even different training runs on the same set can create radically different spaces due to random initial conditions. Data sets can consist of any sounds, so there is no way to know ahead of time what perceptual qualities will change with the location and direction of movement in latent space (Tahiroğlu, Kastemaa, and Koli (2020)).

Novelty

Novelty is fundamental to theories of creativity (Boden (2004)), and became a driving aesthetic in art and music in the 20th century (Martindale (1990)), and in new musical instrument design in particular (Jordà (2004)). In one sense, all sounds generated from a GAN are novel because the data set is modeled (not indexed or memorized). More interestingly, trained models "generalize" so that the space of the data set is "filled in" with novel interpolated sounds. This fosters creative exploration as a search for music trajectories creating entirely new sonic experiences. The expansive latent spaces can be computationally searched or learned by other networks (e.g. Kamath et al. (2024)), or musicians can "manually" explore regions of the space that correspond to unique musical textures, creating variations on existing themes, or experimenting with new musical concepts.

Related Work

The notion of a low dimensional space where instrument timbres can be located while preserving the subjective perception of distance between pairs of instruments was formalized by Grey (1977). Wessel (1979) used the low dimensional space where instruments were located for synthesis by mapping the spatial dimensions to an additive sinusoidal synthesis model, interpolating synthesis parameters if coordinates were chosen between instrument locations.

The Latent Timbre System (Tatar, Bisig, and Pasquier (2021)) used a VAE to learn timbral frames of audio. Esling and colleagues (2018) used a VAE that learns a low-dimensional representation of timbres. The RAVE system (Caillon and Esling (2021)) incorporated those ideas in to a real-time system which Vigliensoni (2023) navigates with mapped physical interfaces. Yee-King (2022) summarizes how latent spaces can be navigated musically. Real-time navigation through VAE latent space is made possible by restricting the output for each latent point to be a very short (e.g. 25ms) sample of sound.

Engel (2017) and members of the Google Magenta team showed how a large data set (called "nsynth") of musical instrument tones could be represented as a sequence of latent vectors learned by training a neural network. Since each sound is modeled as a sequence of latent vectors, it complicates the use of the latent space for interactive control. GanSynth (Engel et al. (2019)) uses a Generative Adversarial Network (GAN) trained on the nsynth data set, but with this model, the entire note duration (4 seconds) including the attack, sustain, and decay portions of the note, is generated by a single latent vector (augmented with information specifying the pitch). The long output durations limit the real-time latent parameter update rate.

The Freedom of real-time systems

To exploit the GAN's ability to generate complexity and novelty, but to manage the real-time limitations the come from extended output sample durations, several systems have been proposed. Tahiroğlu et al. (2021) developed Al-terity, a physical interface for navigating a GAN space. Physical configurations measured with sensors are mapped

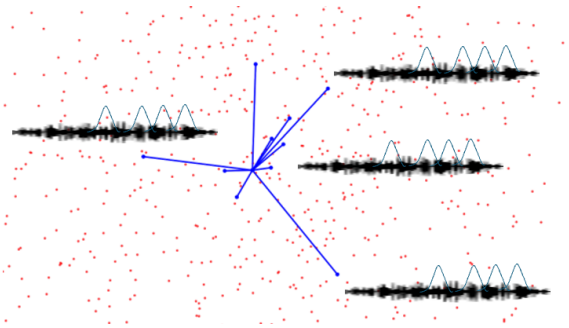

Figure 2: Blue lines are showing the principle components and red points are the vector points in the reduced dimensional space. Each vector point generates 4 second of audio samples feed into a custom granular synthesis module, as shown with audio waves and time windows.

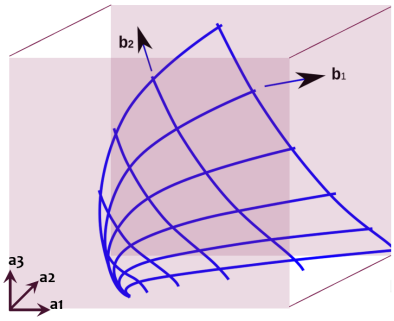

Figure 3: From the high dimensional GAN latent space ('a'), we can take lower dimensional surfaces ('b'), and then use indexes in the low-dimensional space as parameters for their associated sounds, to train a new synthesizer,

to latent vectors, but generate audio for several latent vectors in a neighborhood of the index from which grains (short windows of samples) are chosen and overlapped and added using granular synthesis techniques. While Al-terity serves as the physical interface for exploiting the latent space, Principal Component Analysis (PCA) has been used to find latent directions as it is shown in Figure 2. PCA is applied to the GANSpaceSynth's activation space. These directions are utilised for a more deliberate sampling of the latent space. In order to navigate to a specific point in the activation space suitable for synthesis, linear combination of the PCA directions is calculated. In this structure, adjustable coefficients determine the distance to traverse along each direction, originating from a specified starting point (Tahiroğlu, Kastemaa, and Koli (2021)). Multiple samples are generated within a small neighborhood in latent space, and short windows ("grains") of samples are extracted from the output segments. Overlap and adding of grains comprises the final synthesis output. The precomputation of audio segments combined with granular synthesis enable real-time continuous transformation through Al-terity instrument. This transformation occurs when the musician discovers a new point in the instrument's sound space, initiating a shift in the performance and leading to another transition. This continuous process is how the music consistently changes, shifts, and transforms.

A different approach was pursued by Wyse et al. (2022) to take advantage of the GAN's ability to organize sounds spatially and produce novelty, but to overcome the GAN's inherent obstacle to real-time control. They explore the trained GAN through random latent point generation and navigation. Four points are then chosen for their musical potential that then serve as the four corners of a two-dimensional submanifold embedded in the high-dimensional latent space spanning other musically interesting sounds (Figure 3). The low-dimensional coordinates of the embedded submanifold are then grid-sampled and paired with the GAN-generated sounds to create a synthetic data set to conditionally train a Recurrent Neural Networks (RNN) which generates samples sequentially so that during inference it responds immediately and continuously to changes in the conditioned parameters. The sounds are statistically similar to the corresponding GAN sounds, but don't have the fixed temporal duration and structure, and can respond to parameter changes at the audio sample rate.

Humans in the Co-creative Trajectory Loop

Sound artists imagine, explore, create, and perform relationships between sounds. The process of designing spaces that embody relationships between sounds is as old as instrument building and composition, and as new as DJ rigs and live coding. Smooth transformations between timbres and textures also have a history that extends back over 100 years. Morphing becomes an important concept in 20th century music with "tape techniques" (e.g. Pierre Schaeffer (1948) *Etude aux chemins de fer*) and synthesis (e.g. Karlheinz Stockhausen, *Gesang der Junglinge*) which used carefully constructed transformations between bell-like synthetic tones and the timbre of children's voices. Trevor Wishart's music heavily draws on morph-like transformations (e.g. 1973, *Redbird: A Political Prisoner's Dream*). Other examples from 20th century electroacoustic composers that organize sounds into navigable spaces include Michael McNabb (1978) *Dreamsong* and Jonathan Harvey's (1980) *Mortuos Plango, Vivos Voco*. The transformations we hear in these works as paths through neighboring sounds were created with almost unfathomable manual labor.

In order to better understand how today's musicians interpret, understand and apply sound - space - creativity - transformation relationships to their own music creation process using latent space, we contacted practicing musicians making significant contributions to music with AI methods and technologies. We asked them three main questions: a) what are the most important feature of latent spaces and AI models they use that stand out from other musical instruments, b) what DAWs or interfaces they use to navigate latent space and how they influence their approach to creating or performing music, and c) what do latent space and AI models inspire them to do differently in music that was difficult or not possible before? Below we elaborate the answers we received from *Farzaneh Nouri*[1]; musician, researcher & sound artist, *Sam Pluta*[2]; composer, laptop improviser, electron-

[1]Farzaneh Nouri `https://farzanehnouri.com`
[2]Sam Pluta `http://www.sampluta.com`

ics performer & sound artist, *Erik Nyström*[3]; composer of electroacoustic works, live computer music & sound installations, and *Mat Dryhurst*[4]; artist, musician, & technological researcher. Their responses clustered around three themes:

- **Real-time Interaction and Instrumentation:** Nouri integrates AI models into live music improvisation scenarios where latent spaces are the foundation for artificial improvisers. These spaces listen to other performers and respond in real-time, contributing to the development of compositions. Similarly Pluta's music-making practice is centered around creating a varied latent space for live electronic performance. Pluta has developed a large software instrument for live electronic performance over the past 15 years. The goal is to create an environment where a latent space is available to the performer. Nyström's approach as well involves using latent space for exploration and improvisation. Nyström uses UMAP dimension reduction to map a corpus of samples in 2D space, providing an intuitive and improvisation-friendly sound interface. This enables exploration during performances using an x/y MIDI controller. Dryhurst has created music using vocal timbre transfer (Holly+), tone transfer (voice to instrument, such as the DDSP plugin they worked on with Google Magenta), vocal generation, and more recently toyed with IRCAM rave models and stable audio for music generation from custom data sets.
- **Accessibility and Recall:** Nouri's recent models use Self-Organizing Maps for dimensionality reduction and Neural Networks for generating output values based on incoming sounds. The AI agent's outputs influence sound synthesis through various methods, including digital and hybrid sound synthesis. The crucial aspect for Pluta is making each sonic identity within the latent space accessible quickly and musically. AI models make it easier to recall musically meaningful settings in high dimensional spaces. Alternatively, Nyström employs regression to reduce a large number of synth parameters to a 3D space, offering a rich territory of possible sonic textures and timbres that wouldn't be easily discovered through manual adjustments. Dryhurst emphasises the ability to customise a sonic playspace for exploration and finding unique corners. Additionally, the capability to share models enables "Identity play," allowing performers to embody the sonic characteristics of others, as shown in the Holly+ project.
- **Humans in the Loop:** Nouri mentions that the use of latent space allows for alternative approaches to human-computer interaction, exploring machine behaviors, and studying the possibility of liberation from control processes in improvisational settings. Latent spaces and AI models allow Pluta to explore complex and chaotic timbre spaces, such as feedback systems. He can harness these systems for musical expression, finding settings that work and storing them for later recall. Nyström frequently uses timbre classification within generative works. An MLP classifier algorithm is trained to classify certain aspects of the complex system's output, generating new sounds that are subjected to the same listening and classification. The evolving landscape, as highlighted by Dryhurst, introduces transformative possibilities. The ability to create and share a model for others to navigate or use in performance is fundamentally new. The feature space of a model presents a different terrain to navigate than a synthesizer or sample pack. Additionally, the ability to transform one's voice to feature another, or an instrument, is a unique opportunity that deep learning models now provide. Timbre transfer techniques, while previously employed, now offer exciting possibilities for mutating input sounds. As prompted models improve and grow in complexity, new sound become achievable and push the evolution of interfaces for invoking sounds semantically or developing new subjective vocabularies.

In the musicians' responses, we can observe how they emphasise the fundamental building blocks of the latent space as being a configurable and divergence exploratory space that enables new modes of human-machine creativity and sonic possibilities that go beyond traditional approaches in music-making. The transformative potential of the latent space mechanisms is also acknowledged by the musicians as they configure and explore trajectories of possible sonic timbres that would not have been be possible to discover otherwise. This highlights the configurability of the latent space as a creative space that is tailored to the musician's needs and compositional visions. Their utilisation of the latent spaces for exploring complex timbre spaces, designing and probing for out-of-domain novelty, and reconfiguring architectures for real time exploration, supports alternative approaches to explore machine behaviours in musician-AI interaction, reflecting a symbiotic relationship between human creativity and AI models in shaping new musical expression.

Conclusions

In this paper we have presented an approach to understanding latent spaces as a platform for exploring new sound spaces and their intricate relationship with established music theory and historical practices. We believe that latent spaces should be considered as co-creative platforms as these neural networks permit us to create explicit spatial embodiments of sonic relationships with our creative practices. Interestingly, this perspective aligns closely with historical practices and with the experiences shared by contemporary musicians.

As we have shown in this paper, latent space can be considered as a co-creative space with rich musical potential in the work of the practicioners we have discussed. With that in mind, future work includes making training neural networks more accessible and available across diverse economic, social, and geographic circumstances, increasing computational efficiency in the interest of both real-time musical performance and the environment, and discovering new musical possibilities (instruments, compositions, genres) that must exist given capabilities of new emerging computational tools.

Author Contributions

The authors participated equally in writing this manuscript.

[3]Erik Nyström `https://www.eriknystrom.com`

[4]Mat Dryhurst `https://herndondryhurst.studio`

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Berns, S., and Colton, S. 2020. Bridging generative deep learning and computational creativity. In *ICCC*, 406–409.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Broad, T.; Berns, S.; Colton, S.; and Grierson, M. 2021. Active divergence with generative deep learning–a survey and taxonomy. *arXiv preprint arXiv:2107.05599*.

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

Dong, H.-W.; Liu, X.; Pons, J.; Bhattacharya, G.; Pascual, S.; Serrà, J.; Berg-Kirkpatrick, T.; and McAuley, J. 2023. Clipsonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. *arXiv preprint arXiv:2306.09635*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Norouzi, M.; Eck, D.; and Simonyan, K. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *Int. Conf. on Machine Learning*, 1068–1077. PMLR.

Engel, J.; Agrawal, K. K.; Chen, S.; Gulrajani, I.; Donahue, C.; and Roberts, A. 2019. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.

Esling, P.; Bitton, A.; et al. 2018. Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics. *arXiv preprint arXiv:1805.08501*.

Grey, J. M. 1977. Multidimensional perceptual scaling of musical timbres. *JASA* 61(5):1270–1277.

Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.

Jordà, S. 2004. Instruments and players: Some thoughts on digital lutherie. *J. of New Music Research* 33(3):321–341.

Kamath, P.; Gupta, C.; Wyse, L.; and Nanayakkara, S. 2024. Example-based framework for perceptually guided audio texture generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Liu, H.; Tian, Q.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2023b. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*.

Martindale, C. 1990. *The clockwork muse: The predictability of artistic change.* Basic Books.

Meyer, L. B. 1967. Music, the arts, and ideas: patterns and predictions in twentieth-century culture. *(No Title)*.

Tahiroğlu, K.; Kastemaa, M.; and Koli, O. 2020. Al-terity: Non-rigid musical instrument with artificial intelligence applied to real-time audio synthesis. In *Int. Conf. on New Interfaces for Musical Expression*. NIME.

Tahiroğlu, K.; Kastemaa, M.; and Koli, O. 2021. Ai-terity 2.0: An autonomous nime featuring ganspacesynth deep learning model. In *International Conference on New Interfaces for Musical Expression*, 1001–1004. NIME.

Tahiroğlu, K.; Kastemaa, M.; and Koli, O. 2021. Ganspacesynth: A hybrid generative adversarial network architecture for organising the latent space using a dimensionality reduction for real-time audio synthesis. In *2nd Joint Conf. on AI Music Creativity.* Aalto University.

Tatar, K.; Bisig, D.; and Pasquier, P. 2021. Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications. *Neural Computing and Applications* 33:67–84.

Vigliensoni, G.; Fiebrink, R.; et al. 2023. Steering latent audio models through interactive machine learning. In *International Conference on Computational Creativity*.

Wessel, D. L. 1979. Timbre space as a musical control structure. *Computer music journal* 45–52.

Wyse, L.; Kamath, P.; and Gupta, C. 2022. Sound model factory: An integrated system architecture for generative audio modelling. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 308–322. Springer.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yee-King, M. 2022. Latent spaces: A creative approach. In *The Language of Creative AI: Practices, Aesthetics and Structures*. Springer. 137–154.

Yuan, Y.; Liu, H.; Liu, X.; Huang, Q.; Plumbley, M. D.; and Wang, W. 2023. Retrieval-augmented text-to-audio generation. *arXiv preprint arXiv:2309.08051*.