Stacco: Exploring the Embodied Perception of Latent Representations in Neural Synthesis

Nicola Privato*
Intelligent Instruments Lab
University of Iceland
nprivato@hi.is

Victor Shepardson*
Intelligent Instruments Lab
University of Iceland
victorshepardson@hi.is

Giacomo Lepri
InfoMus - Casa Paganini
University of Genoa, Italy
giacomo.lepri@edu.unige.it

Thor Magnusson
Intelligent Instruments Lab
University of Iceland
thormagnusson@hi.is

ABSTRACT

The application of neural audio synthesis methods for sound generation has grown significantly in recent years. Among such systems, streaming autoencoders such as RAVE are particularly suitable for instrument design, as they map audio to and from control signals in an abstract latent space with acceptable latency. Despite the uptake of autoencoders in NIME design, little research has been done to characterize the latent spaces of audio models, and to investigate their affordances in practical musical scenarios. In this paper we present Stacco, an instrument specifically designed for the intuitive control of neural audio synthesis latent parameters through the displacement of magnetic objects on a wooden board with four magnetic attractors. We then examine models trained on the same data with different seeds, we explore strategies for more consistent mappings from audio to latent space, and propose a method for stitching the latent space of one model to another. Finally, in a user study, we investigate whether and how these techniques are perceived through embodied practice with Stacco.

Author Keywords

Artificial Intelligence, Machine Learning, Neural Audio Synthesis, Latent Space, Latent Adaptation, RAVE, Stacco, Thales, Chowndolo, Embodied Sketching

CCS Concepts

•Applied computing → Sound and music computing; •General and reference → Empirical studies; •Human-centered computing → User interface design;

1. INTRODUCTION

In recent years, neural audio synthesis (NAS) has emerged as a novel approach to sound synthesis, considerably expanding the creative possibilities of technologists, musicians and sound artists [8]. NAS models consist of artificial neural networks trained to predict or reconstruct corpora of raw sounds, learning in the process to represent them in the network's hidden layers. The activation values of neurons encoding a sound can be thought of as a point in a multidimensional *latent space*, whose manipulation is among the most compelling features of NAS.

Historically, neural models powerful enough to handle raw audio with any degree of realism were difficult to run in real time, but the recent introduction of RAVE [8], a model capable of real-time performances and easy to integrate into existing workflows, has led to a plethora of use cases encompassing instrument design [30, 26], sound engineering[1] and music composition [23].

This diverse corpus of contributions demonstrates the rife interest in this novel approach to sound synthesis among artists and designers. But incorporating black-box neural synthesis models in compositional and design practices raises challenges for explainable artificial intelligence (XAI), as these do not expose the processes leading to the generation of their outputs and tend to distribute the sound features of the dataset in unpredictable ways [7].

Among the first applications of XAI in the artistic domain, Bryan-Kinns et al. proposes to map specific latent dimensions with meaningful musical parameters in a MeasureVAE model, providing feedback on the distribution in the latent space through a visual interface [6]. Complementing this approach, the diverse contributions from the 2023 workshop on XAI for the Arts [7] reframe the explainability problem from a broader perspective, encompassing the nature of explanation, how AI models, features, and training sets affect explanation, user-centred software and hardware design, and interaction design for explainability [5].

In line with this, Privato and Armitage [24] extend the *Explanatory Pragmatist* framework to XAI in music performance, arguing that there is no universal approach to explainability and that context (encompassing instrument, performer and audience), is crucial for effective XAI strategies.

This paper aims to address XAIxArts from this latter angle. How do algorithmic strategies for understanding latent space manifest in a NIME specifically designed for neural synthesis? What can the DMI designer expect when training neural synthesis models and mapping their latent

[*] These authors contributed equally to this work

[1] https://semilla.ai

NIME'24, 4–6 September, Utrecht, The Netherlands.

spaces? And what novel compositional and performative strategies emerge around latent spaces' peculiar affordances?

In the following sections, we overview RAVE's features and frame our contribution within diverse XAI approaches. We then introduce Stacco, a Digital Musical Instrument (DMI) specifically designed for the intuitive and playful navigation of latent spaces, and explore strategies for adapting latent distributions from one model to another. We finally use Stacco to explore how users perceive and understand the adapted latent spaces through embodied interaction.

2. BACKGROUND

2.1 RAVE

The Real-time Audio Variational Encoder (RAVE) is a neural audio synthesis method introduced in 2021 by Caillon and Esling [8]. Its relatively high-fidelity sound and low latency have drastically facilitated the applications of neural synthesis in interactive contexts.

RAVE learns by a two-phase procedure, consisting of a representation learning phase as a variational autoencoder (VAE) [15] followed by an adversarial fine-tuning phase which improves sound quality. Training requires many hours of GPU computation, but once trained, models can run in real time on laptop CPUs, through Pure Data, Max/MSP [1] and SuperCollider [11] plugins, or direct implementation in Python or C++ programs.

As an autoencoder, RAVE consists of two main functions that the user can call separately: an *encoding* phase compresses a 48 KHz stream of audio to a stream of latent vectors, typically with a sampling rate of about 23Hz and 4 to 32 channels, and a *decoding* phase that synthesizes audio from latent vectors.

A trained RAVE model is typically used by feeding a sound through both encoder and decoder to reinterpret it through the model's training data, or by using control-rate signals to manipulate directly the latent space before decoding. These methods may also be creatively combined in various ways, for instance by directly controlling one latent variable while taking the others from a sound fed into the encoder.

RAVE has appeared in diverse contributions to recent NIME conferences and beyond: it is part of the synthesis engine in the Living Looper [30], in Thales [26] and Semilla [31]. Pelinsky includes RAVE in a pipeline for embedded synthesis [22], and others have explored its embodied navigation through spatial metaphors [29, 3].

Yet, incorporating RAVE and other audio autoencoders in compositional and design practices raises unique XAI challenges. Among these are the model's arbitrary distribution of the sound features in the latent space and entanglement of the latent dimensions, with one latent responding differently as the state of the other is changed, making the navigation less predictable [25].

2.2 Understanding Latent Spaces

One approach to making latent spaces more understandable is to align them with known features of the audio, either by pulling out certain aspects (e.g. pitch) to be controlled explicitly, or attempting to factor the latent space into independent parts dealing with distinct aspects of sound. Devis at al. [10] show how to use explicit audio descriptors together with a learned latent space while preventing redundancy between explicit descriptors and learned latents, so that descriptors can later be modified to control the sound. Relatedly, Nercessian [21] separates pitch and a phonetic encoding out of the RAVE latent space, targeting singing voice synthesis.

A different line of research focuses on post-hoc explanations of how sound is represented in learned latent spaces. Hawley and Steinmetz [13] visualize the latent space of audio autoencoders, finding that as audio effects are applied, the movement of sounds in latent space is visible but not easy to describe.

A third approach is to understand latent spaces in terms of one another, "stitching" neural networks together so that e.g. the encoder of one model and the decoder of another can communicate. This was first studied for computer vision models [16][9][4]. In these methods, a linear transformation is fit such that it maps a hidden layer of one model to the same layer of another, and the resulting hybrid model is found to retain most of the performance of the originals.

Moschella et al. [20] introduce a related method, using similarity to anchor points rather than a linear transformation. Though this allows them to stitch two models without an extra training step, it assumes that the decoder model has already been trained using their method. The modified latent space is also semantically different from either original model, which poses difficulties in a context where latent space is meant to be manipulated directly. For these reasons, we eschew relative representations and return to the technique of model stitching in Section 4.2.

Finally, we identify a fourth approach in the embodied understanding of latent spaces. Scurto and Postel's latent soundwalks [29] explore a literal equivalence between RAVE latent space and 3D space as mediated by a virtual avatar, Valenzuela's Semilla [31] maps the positions of seeds cast across a table by hand, and Armitage and Privato investigate the three-dimensional projection of latent spaces through sound spatialisation [3]. This line of investigation is particularly relevant for a community focused on instrument design and music creativity such as NIME, in that it recognizes the role played by our instruments and compositional practices in interpreting the algorithm's workings, and explores how the principles of XAI apply to such contexts.

3. STACCO

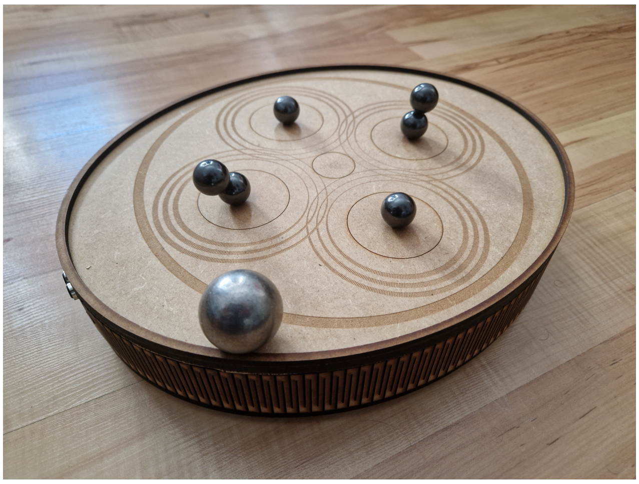

Figure 1: Stacco.

In line with this approach, we developed Stacco, an interface that facilitates the interaction and composition with RAVE whilst providing a rich and playful musical experience. We used Stacco to investigate the application of XAI strategies in a NIME, focusing on how these affect the performative and compositional experience.

Stacco is a DMI embedding magnetic sensors underneath an engraved surface. It builds on the notion of *instrument-score* [27] to underlay the fact that “computer systems used in musical performance carry as much the notion of an instrument as that of a score” [28], and inherits the design features of other instruments based on permanent magnets, such as the Chowndolo [19], a magnetic pendulum whose trajectories are affected by permanent magnets placed over a metal board, and Thales [26], a pair of handheld, disc-shaped controllers that interact with ferromagnetic objects through permanent magnets. Similarly to these instruments, Stacco exploits the liveness of ferromagnetic objects to manipulate sound, engaging the performer with the entanglements of its magnetic forces.

3.1 Design

Stacco consists of an oval laser-cut case containing a Bela [18] and four attractors, each combining one magnetometer with a stack of magnets in a 3D-printed enclosure. The attractors are placed in four symmetrical points below an engraved board placed on the top of the case, which features a raised edge and is enclosed via a living hinge structure. Each magnetometer performs two-dimensional readings of nearby magnetic fields, providing 8 continuous streams of data reflecting the position of nearby ferromagnetic objects.

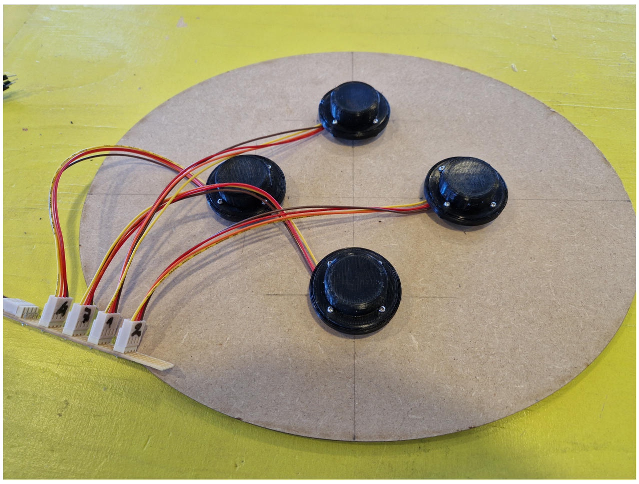

Figure 2: Stacco’s attractors.

The sensors are connected to Bela for embedded synthesis and/or OSC data forwarding to a connected laptop, and are coupled with permanent magnets to interact with nearby magnetic and ferromagnetic objects actively. Circular engravings on the upper face of the board hint at the position of the four attractors.

The performer interacts with Stacco by throwing and displacing on the board a series of magnetic spheres of variable dimensions (Figure 1), engaging in a playful dance of agencies with the four attractors, whose intertwined magnetic fields metaphorically reflect the entanglements in the latent space. Indeed, with RAVE it is often impossible to univocally map one input with one particular sonic feature, and the manipulation of one latent dimension drastically affects the navigation of the others (Section 4.1). This is mirrored in Stacco’s magnetic materiality, through the overlapping of the attractors’ sensing fields, and the complex interactions of the magnetic spheres with each other and with the instrument itself.

To help with charting and recalling gestures in the latent space, we designed tailored oval sheets to be placed on Stacco’s top, which the performer can customise with different inscriptions (Figure 3). As we discuss in 6, this feature proved useful in designing our study, as we could easily compare the different performances with one another, and was particularly appreciated by the participants in that it allowed them to notate and reproduce their compositional ideas while exploring the latent space.

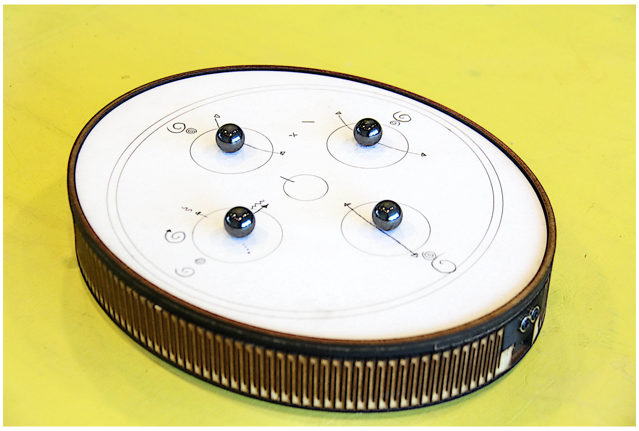

Figure 3: Embodied Sketching with Stacco. P4 Score.

As it reads the changes in the magnetic fields around its attractors, Stacco forwards 8 continuous values. These may be individually mapped to each of the first 8 parameters of a RAVE decoder, which, as described in Section 4.1, are usually the most significant and clearly perceivable when manipulating the latent space.

However, as we discuss in the next section, in mapping Stacco to RAVE we slightly deviated from this strategy.

3.2 Mapping

Because in many RAVE models the first latent dimension mostly determines loudness (Section 4.1), tailoring a mapping for the first latent is often a first, effective “macro-scale” [2] step. Once this is addressed, the distribution of sound features in the remaining latent dimensions may be drawn out through embodied exploration.

For our user study (Section 5), we devised a mapping for the first latent accommodating both drone-oriented and percussive models, combining the total rate of change of sensor readings with their total distance from a neutral position (no spheres present). This allowed us to keep the instrument silent at rest, to activate it through the sustained tension of magnets against the field, and to trigger transient sounds with quick motions of the spheres.

Individual readings from the sensors were then scaled and applied to latents 2 through 9.

4. INVESTIGATING RAVE LATENTS

As we performed with Stacco, RAVE’s arbitrary distribution of the sound features in the latent space became strikingly apparent, with similar gestures often producing very different sonic outcomes when performed on different models.

In the following sections, we investigate how the model distributes sound features in latent space, whether it is possible to map one latent to another, and how this adaptation is perceived and understood as we perform with a DMI.

We began by studying the simple case of what happens when two RAVE models are trained on the same data under similar conditions. We immediately found that the latent spaces of such models could very different, even opposite to one another; to understand why requires some elaboration of how RAVE reduces dimensionality of the latent space.

4.1 Sign Normalization

A VAE such as RAVE balances reconstruction error with precision in the latent representation. The training process includes an incentive to add noise[2] to the latent representation, leaving only as much precision as needed to reconstruct the input. Typically, some latent dimensions become pure noise, carrying no information about the input;others retain varying amounts of information, limited by some amount of noise. To make the latent space easier to work with directly, RAVE applies a principal components analysis (PCA) transformation to identify meaningful dimensions buried in the haystack.

Essentially, this takes the many raw latent dimensions of a trained RAVE model, discards pure noise dimensions, and arranges the remaining significant dimensions in order of descending importance[3]. This transformation gets baked into the RAVE encoder, and the inverse transformation into the RAVE decoder, where the discarded dimensions are filled back in with noise.

For example, in RAVE models trained on audio datasets with any significant dynamics, the first and most important latent tends to encode loudness. The second might encode something like pitch or brightness. Another one might represent the presence of transients or a particular spectral band. However, more often each dimension corresponds to some entangled combination of such features.

There is one difficulty with the PCA method used by RAVE: it is agnostic to the sign of each latent dimension, due to a symmetry in the definition of PCA. For example, positive values of the first latent might correspond to greater loudness, but it could just as well be that negative values correspond to greater loudness. The only difference would be a negation of the first column of the PCA projection, and such an alternate PCA projection is equally valid.

So, if there are N significant latent dimensions, there are 2^N equivalent latent spaces where some subset of dimensions are flipped; the training process chooses the signs arbitrarily. With Stacco, the effects of this became obvious when switching between models; when the first latent dimension changed direction, gestures which had produced loud sound became silent and vice versa.

To make any more nuanced comparisons between different RAVE models, we first needed to normalize the signs of the PCA transformation. To achieve this, we randomly sampled sequences of latents from the prior distribution, mapped them to audio via the RAVE decoder, and then extracted an audio descriptor, so that each latent time-step had an associated descriptor value. We then measured the correlation of the audio descriptor with each latent dimension and took the sign (±1). Finally, we modified the PCA transform to incorporate those signs. In other words, the polarity of each latent was now flipped so it correlated positively with an audio descriptor.

Notably, this method does not cause the latents to become more strongly correlated with each other; indeed the purpose of a VAE is that they be uncorrelated. The only effect is that for any latents which do correlate with the descriptor, the direction of correlation is positive.

The most significant dimensions of RAVE models often relate to loudness and brightness, qualities conventionally associated with an increasing parameter. We chose a descriptor representing louder and brighter sounds, so that RAVE latents would tend to respect the convention. Specifically, the RMS amplitude of the discrete time difference (i.e., envelope of high-pass filtered audio) was used.

This *sign normalization* is implemented as an option when exporting models in our fork of RAVE[4]. We use sign normalization as a baseline modification to all models described below.

4.2 RAVE Model Stitching

At this point, we could be confident that the most superficial differences between RAVE models were mitigated, and investigate how similar the latent spaces actually are. Do two models trained in the same way on the same dataset discover the same latent factors of variation, or not?

As it turns out, mostly not. Depending on the dataset, only the first (albeit most important) 1-3 dimensions are similar, with the remaining dimensions apparently quite different. For example, consider the pairwise correlations between latent dimensions in different RAVE models (Figure 4). Here, we trained trios of models with the same data and hyperparameters but different random seeds affecting model initialization, noise, and sampling order of data. We can see that where the correlation between latents is strong in the upper left corner of each plot (between the first few latents) it is also positive, reflecting the sign normalization. However, the remaining correlations are only vaguely clustered around the diagonal, implying that the two models don't use similar latent dimensions in a similar order, but rather have learned different representations of the sound.

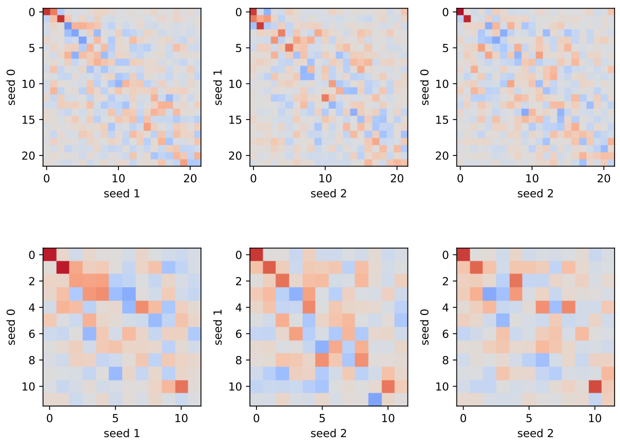

Figure 4: Pearson correlation between latent dimensions in pairs of RAVE models trained on the same dataset with different random seeds. Top: electric guitar models with 22 latents. Bottom: soprano sax models with 12 latents. Red is positive, blue negative.

[2] Here 'noise' refers to randomness, not specifically to noisy sounds.

[3] In the VAE terminology, if a latent dimension has collapsed to the (standard normal) prior, a sample from the posterior contains no information about the input, while its mean is always zero. Note that RAVE's PCA step is fit to the *means* of the posterior, not to samples, so noise dimensions appear to PCA with low variance across the dataset, while significant dimensions have varying means. Thus PCA finds that the low-variance noise dimensions can be truncated without loss of information.

[4] `https://github.com/victor-shepardson/RAVE`

But just how different are these latent spaces? Might the apparent differences be superficial and the latent space of one model be linearly related to another?

To find out, we fit a linear adapter between parallel latents obtained by encoding the original training data with a pair of RAVE models. This adapter is simply the least-squares solution to an equation relating the encodings:[5]

$$\min_{W} \|Z_A W - Z_B\|$$

Where Z_A and Z_B are $T \times N$ and $T \times M$ matrices of latent encodings of the same audio data by two different RAVE models A and B. Here, T is the number of time steps and N the number of significant latent dimensions after PCA. W is the resulting $N \times M$ adapter matrix, allowing the encoder of A to communicate with the decoder of B.

Csiszárik et al. call this the "direct matching" method of model stitching [9]. It doesn't require backpropagation through the model, and can be fit quickly on a CPU[6]. We did not fit a bias parameter, since RAVE latents already have zero mean.

Figure 5 shows correlations between latents with adapters (compare to Figure 4).

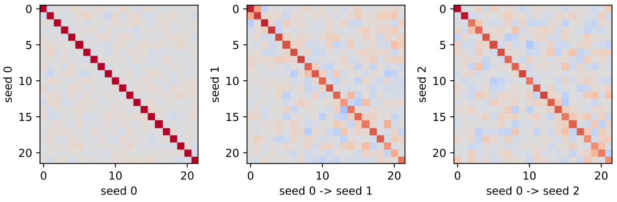

Figure 5: Left: Pearson correlation between latent dimensions in one of the guitar models. Center: between latents encoded by model 1 and latents encoded by model 0 but adapted to model 1. Right: likewise for models 2 and 0.

As expected, the leftmost plot in Figure 5 was nearly the identity matrix, with the latents of model 0 perfectly correlated with themselves along the diagonal and nearly uncorrelated with each other. With adapters, this structure was closely recovered in the cross-correlations between models.

It appears that RAVE models do learn similar latent spaces under similar conditions, but only up to an arbitrary linear transformation. This is in line with the observation of Locatello et al. that VAEs with a factorized prior cannot learn disentangled representations in a purely unsupervised setting [17].

Even though neither the RAVE method of applying PCA nor our supervised sign normalization circumvent this, using a linear adapter as described above we can encode audio through one RAVE model and decode through another, obtaining a reconstruction close in fidelity to either of the individual RAVE models.

We found that this method of latent adaptation can also stitch RAVE models trained on different datasets to some degree. In this case, at least one of the models is being fed out-of-domain data, i.e. being used for timbre transfer, when we collect the dataset of parallel latents. This can lead to extreme values of latents which then show up in the adapter matrix. To mitigate this, we clipped latents to five standard deviations before fitting the adapter and when adapting new latents. In this setting, the latents are not zero-mean, so we do fit a bias parameter.

Latent adaptation is implemented as an additional script in our RAVE fork, exporting an `nn~` compatible adapter which stitches the encoder of one RAVE model to the decoder of another given a reference audio dataset.

5. USER STUDY

At this point, we wanted to know how the latent adaptation developed in Section 4.2 was perceived in practice as part of a DMI. We designed a task-centric user study with Stacco to facilitate direct comparisons between RAVE models by performers, and collect qualitative observations from these.

With Stacco, only RAVE decoders were used, to audibilize the state of the magnetic sensors, which is mapped to latent space. Rather than using adapters to stitch one RAVE encoder to a different decoder, we studied them as a way of maintaining a stable mapping from sensor to sound when swapping RAVE models.

5.1 Methodology

We recruited five participants, P1-P5, from the artistic community of the area. Four of them are trained musicians, each with a different musical expertise: P1 is a trained classical pianist; P2 is a bass player; P3 and P5 are electronic musicians. Of the five, only one (P4) is not a trained musician, being instead a practising installation artist.

We planned five individual sessions, preparing Stacco with a blank oval cardboard for embodied sketching (Figure 3), a pencil and an eraser on the side, and four magnetic spheres placed each at the centre of one attractor.

We began each session by introducing Stacco as a new musical interface of our design, that we are experimenting with to investigate a novel synthesis technique called "neural synthesis" and invited the participant to explore and play the instrument. In this phase, the participants were free to ask questions on the instrument's workings.

At the end of this exploratory phase, we invited each participant to work for 10 minutes on one or more small pieces, indicating a total length of 5 to 30 seconds. We also clarified that the pieces should be repeatable and that to help memorisation it was optionally possible to write notes on the cardboard.

At the end of the ten minutes, the participant was invited to perform the composition three times, first on *setting* A, that is, the same model used to compose, then on two different settings (B and C).

After each performance, we asked the participant to freely comment, and after playing with B and C we asked what, if anything, from the original composition was preserved in playing on the new settings.

Finally, we asked two more questions:

1. Which of the two settings (B and C) preserved most of the original character of the performance?

2. Which setting did the participant enjoy most, and why?

The whole process was repeated once more in a second *round* for each participant, with a different series of settings A, B and C. We write P3-2, for example, to denote 'P3, round 2'.

[5] As in RAVE's PCA step, we use the latent mean $\mu_z(x)$ rather than posterior samples $z \sim Q(z|x)$.

[6] While prior work reported that direct matching was less effective than end-to-end optimization, the efficiency of direct matching makes it more useful to DMI designers, allowing existing RAVE models to be stitched quickly without special hardware.

5.2 Settings

In each of the ten total rounds, setting A consisted of one RAVE model, while settings B and C both used a second RAVE model, with either B or C including an adapter (see 4.2) to make the latent space more similar to that of A. The order in which the adapted model was presented (as B or C) was randomized to reduce the influence of order effects.

Each participant encountered settings derived from RAVE models trained on the same data in one round, and different data (between A and B/C) in the other round. The order of these same/different rounds was also varied between participants.

Setting A always used a RAVE model trained on guitar or saxophone sounds. When B and C used models trained on different data, it was the organ music. By using the three random seeded versions 4 of each guitar and sax model, we were able to use a unique configuration of A, B and C in every round, to make our observations less dependent on the particularities of any pair of RAVE models.

	Setting A	Setting B	Setting C
P1 - 1	G0	G1	G0 → G1
P1 - 2	S0	S0→OR	OR
P2 - 1	S1	S1→S2	S2
P2 - 2	G2	G2→OR	OR
P3 - 1	S2	S2→OR	OR
P3 - 2	G1	G1→G0	G0
P4 - 1	G2	OR	G2→OR
P4 - 2	S0	S1	S0→S1
P5 - 1	S2	S0	S2→S0
P5 - 2	G0	OR	G0→OR

Figure 6: Study Settings. P1 to P5 round A and B, with guitar (G), saxophone (S), organ (OR) trained on random seeds 0 to 2 and latent adaptation.

5.3 Results

In a few cases, the effect of adapters was striking, seemingly working as expected. This was the case with P3-2, where the adapted setting (B) was perceived as fun and "more trustworthy" than the unadapted one, even responding with a "neater" feeling than setting A in navigating the latent space. Similarly, P5-2 described the unadapted setting (B) as very different from A, and the adapted one as a "blend between A and B." In both cases, when asked which setting they enjoyed the most, the performers described the adapted model as more fun than the unadapted one.

In other cases, participants perceived no particular similarity between setting A and the adapted setting. In both rounds, P1 dismissed any tonal or expressive similarity of settings B and C with setting A, arguing that what had remained consistent was the dynamic response of the spheres to the gestures, which was mapped to the first latent. This was also the case with P5-1, where settings B and C were perceived as similar to each other and quite different from setting A. P3-1 perceived the unadapted setting (C) as more similar to setting A in that it provided higher responsiveness to their gestures, thus making the experience more enjoyable. At the same time, the adapted setting (B) was described as disappointing in that it compelled P3 to change the gestures.

P2 and P4 developed the most idiosyncratic approaches to playing with Stacco, reporting the unadapted setting to better preserve their composition at least once.

5.3.1 P2

P2 developed a unique method of playing, holding the spheres in the air and rotating them against the tug of Stacco's magnets. P2 approached the composition task by developing a vocabulary of gestures, and performed in a semi-improvisational manner, continuing to experiment long past the suggested two minutes. P2 was also the only participant who did not notate the instrument with the pencil.

P2's first round was particularly noteworthy, because the similarity in the sound between A and B (which had been adapted) was striking to the researchers and also commented on by P2. However, P2 identified the unadapted C as "paradoxically" better at preserving their original composition.

In the adapted setting (B), the transients were preserved together with the overall character of the composition, but the unadapted setting (C) provided a darker, "sombre change of mood." Despite the tonal similarity of the adapted model and the precision it endowed in reproducing the composition, with the unadapted one P2 felt more capable of going "where [the] inner ear was leading to."

P2's experience appears to illustrate both what was preserved by adapting the latent space, but also what was lost. Similarly to the first round, in P2-2 the participant struggled to reproduce the original compositional idea with the adapted model but enjoyed distancing from it and exploring the uncharted territory offered by the unadapted one.

5.3.2 P4

In contrast to P2, P4 made immediate and extensive use of the pencils (Figure 3), drawing an articulated and fine score while systematically exploring the attractors in sequence.

In both rounds, P4 chose the unadapted model and gave specific reasons why it was more similar to setting A, describing specific gestures and how the sound they produced had been preserved. P4 described the unadapted setting as more sensitive and therefore fun in round 1, and as more similar in character to A, in that the sounds it produced were longer than A. Interestingly, P4 added that they improvised more on B and had more fun because the sounds "were more out of control."

P4's experience seems to illustrate how, as we interact with the model through a DMI, the frictions from gesture to latent space could be more relevant and interesting to the user than the similarities in the mappings between different sound engines.

6. CONCLUSION

From this user study, we draw the conclusion that once it enters embodied experience through a DMI such as Stacco, the algorithmic adaptation of latent spaces becomes less easy to perceive. Furthermore, an adapted latent space is not always valuable to the performer, who might feel more attuned to a different yet more sensitive setting, to a richer sound, or value surprise and instability over control and predictability.

These results confirm the importance of adopting pragmatic approaches to XAI in music and instrument design, where explanations depend on the interests of the stakeholders within the musical ecosystem rather than on the a-priori choices of the designer.

Nevertheless, it is likely, and we intend to investigate this in future works, that more experienced users of Stacco (as well as of other DMIs) might still benefit from a consistent mapping of the latent space, which would allow them to transfer the embodied knowledge of a model's latent distribution, matured in many hours of practice, to a new model.

From this study, a series of secondary yet meaningful findings emerged regarding Stacco. First, all of the participants enjoyed the experience of playing with the instrument, they intuitively understood its workings without the need for further clarifications, and developed diverse and original performative techniques; most participants provided valuable insights on how to further enrich Stacco's ergodynamics, for instance, by marking the polarities of the spheres, by adding pressure sensitivity on the board and by raising the board's edge to bounce the spheres around. Second, it confirmed the utility of embodied sketching as a practice for notating gestures with neural synthesis. This method allowed the participants to easily memorise and repeat the trajectories on the instrument's board, and facilitated our analysis of their compositional strategies.

Through sustained practice with Stacco, we also intend to explore whether embodied sketching might facilitate the customization of the instrument, and investigate the novel compositional strategies that artists might develop as the score and the instrument become whole.

Beyond our user study, this research adds to the XAI literature in two ways: first, by reporting on how RAVE and, more generally, audio autoencoders tend to organise sound features extracted from the dataset into the latent space; second, by offering a practical method for a consistent mapping of RAVE's latent spaces between different models, that users can access on our RAVE fork.[7]

Importantly, by embracing a wider, pragmatic scope as advocated by recent XAI contributions within and beyond the arts, our work underlines how the affordances of the interface, the performer's practice, the artistic intent and other, often unpredictable context-dependent factors redefine the understanding and perception of the model as it enters embodied experience.

7. ACKNOWLEDGMENTS

The Intelligent Instruments Lab is supported by the European Research Council (ERC) as part of the Intelligent Instruments project (INTENT), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101001848)

This research was also supported by an NVIDIA hardware grant.

8. ETHICAL STANDARDS

All the participants in the study were informed of the nature of the research before the interviews and consented to the use and analysis of the anonymised data for the purposes of this study. As per NIME principles and code of practice on ethical research [12], this user study complies with gender, culture and language variety requirements in the selection of the participants.

The datasets used to train the models have been collected and deployed with the consent of their legitimate author and/or from open archives [14].

9. REFERENCES

[1] ACIDS. acids-ircam/nn_tilde, Jan. 2023. original-date: 2022-01-07T11:06:57Z.

[2] J. Armitage, T. Magnusson, and A. McPherson. A Scale-Based Ontology of Musical Instrument Design. *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 339–349, May 2023. Place: Mexico City, Mexico.

[3] J. Armitage, N. Privato, V. Shepardson, and C. B. Gutierrez. Explainable AI in music performance: Case studies from live coding and sound spatialisation. In *XAI in Action: Past, Present, and Future Applications*, 2023.

[4] Y. Bansal, P. Nakkiran, and B. Barak. Revisiting Model Stitching to Compare Neural Representations, June 2021. arXiv:2106.07682 [cs, stat].

[5] N. Bryan-Kinns, B. Banar, C. Ford, C. Reed, Y. Zhang, S. Colton, and J. Armitage. Explainable AI and Music. In *TBC (Forthcoming)*. Taylor & Francis, Online, 2024.

[6] N. Bryan-Kinns, B. Banar, C. Ford, C. N. Reed, Y. Zhang, S. Colton, and J. Armitage. Exploring xai for the arts: Explaining latent space in generative music, 2023.

[7] N. Bryan-Kinns, C. Ford, A. Chamberlain, S. Benford, H. Kennedy, Z. Li, W. Qiong, G. Xia, and J. Rezwana. Explainable AI for the Arts - XAIxArts. https://xaixarts.github.io/, 2023.

[8] A. Caillon and P. Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis, 2021.

[9] A. Csiszárik, P. Kőrösi-Szabó, A. K. Matszangosz, G. Papp, and D. Varga. Similarity and Matching of Neural Network Representations, Oct. 2021. arXiv:2110.14633 [cs].

[10] N. Devis, N. Demerlé, S. Nabi, D. Genova, and P. Esling. Continuous descriptor-based control for deep audio synthesis, Feb. 2023. arXiv:2302.13542 [cs, eess].

[11] G. Elia. elgiano/nn.ar, Feb. 2024. original-date: 2023-06-04T00:40:17Z.

[12] M. Fabio, Gold, Nicolas, Chevalier, Cécile, and R. Masu. Nime principles and code of practice on ethical research, 2023.

[13] S. H. Hawley and C. J. Steinmetz. Leveraging Neural Representations for Audio Manipulation, Apr. 2023. arXiv:2304.04394 [cs, eess].

[14] Intelligent Instruments Lab. rave-models (revision ad15daf), 2023.

[15] D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. arXiv: 1906.02691.

[16] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, June 2015. ISSN: 1063-6919.

[17] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv:1811.12359 [cs, stat]*, Nov. 2018. arXiv: 1811.12359.

[7] https://github.com/victor-shepardson/RAVE

[18] A. McPherson, R. Jack, and G. Moro. Action-sound latency: Are our tools fast enough? In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 20–25, Brisbane, Australia, 2016. Queensland Conservatorium Griffith University.

[19] N. Merendino, G. Lepri, A. Rodà, and R. Masu. Redesigning the chowndolo: a reflection-on-action analysis to identify sustainable strategies for nimes design. May 2023.

[20] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà. Relative representations enable zero-shot latent space communication, Mar. 2023. arXiv:2209.15430 [cs].

[21] S. Nercessian. P-RAVE: Improving RAVE through pitch conditioning and more with application to singing voice conversion. 2023.

[22] T. Pelinski, R. Diaz, A. L. B. Temprano, and A. McPherson. Pipeline for recording datasets and running neural networks on the bela embedded hardware platform. May 2023.

[23] N. Privato. Mouja: Experiencing ai through magnetic interactions. In *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[24] N. Privato and J. Armitage. A Context-Sensitive Approach to XAI in Music Performance. In *The 1st International Workshop on Explainable AI for the Arts (XAIxArts), ACM Creativity and Cognition (C&C) 2023*, New York, NY, USA, Sept. 2023. arXiv.

[25] N. Privato, G. Lepri, T. Magnusson, and E. T. Einarsson. Sketching embodied interactions for neural synthesis. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'2024*, Zurich, CH, 2024. To be published.

[26] N. Privato, T. Magnusson, and E. T. Einarsson. Magnetic interactions as a somatosensory interface. May 2023.

[27] N. Privato, T. Magnusson, and E. T. Einarsson. The magnetic score: Somatosensory inscriptions and relational design in the instrument-score. In A. P. D. Ritis, V. Zappi, J. V. Buskirk, and J. Mallia, editors, *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'2023*, pages 36 – 44, Boston, Massachusetts, USA, 2023. Northeastern University.

[28] N. Schnell and M. Battier. Introducing composed instruments, technical and musicological implications. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 156–160, Dublin, Ireland, 2002.

[29] H. Scurto and L. Postel. Soundwalking deep latent spaces. May 2023.

[30] V. Shepardson and T. Magnusson. The living looper: Rethinking the musical loop as a machine action-perception loop. May 2023.

[31] M. H. Valenzuela. semilla.ai.