LVNS-RAVE: Diversified audio generation with RAVE and Latent Vector Novelty Search

Jinyue Guo
RITMO, Department of Musicology
University of Oslo
Oslo, Norway
jinyue.guo@imv.uio.no

Anna-Maria Christodoulou
RITMO, Department of Musicology
University of Oslo
Oslo, Norway
a.m.christodoulou@imv.uio.no

Bálint Laczkó
RITMO, Department of Musicology
University of Oslo
Oslo, Norway
balint.laczko@imv.uio.no

Kyrre Glette
RITMO, Department of Informatics
University of Oslo
Oslo, Norway
kyrrehg@ifi.uio.no

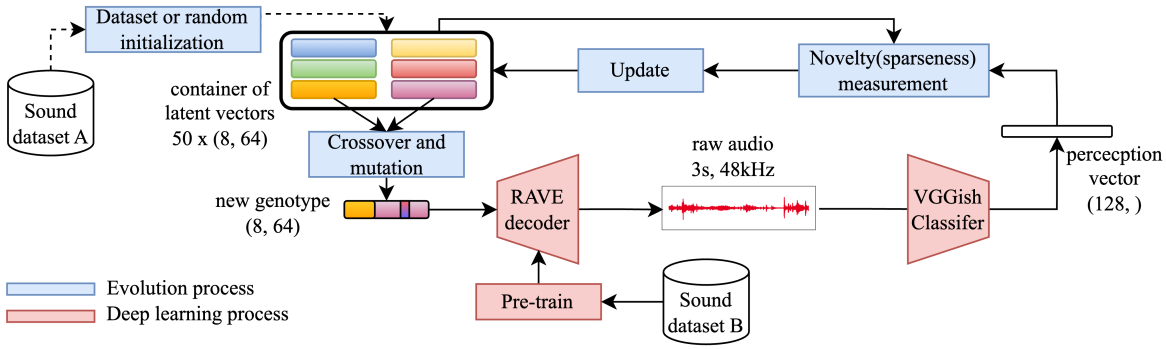

Figure 1: The overall workflow of LVNS-RAVE: the red blocks show the Deep Learning process for high-quality audio generation and perception estimation, and the blue blocks show the Novelty Search process that looks for diversified, novel samples.

ABSTRACT

Evolutionary Algorithms and Generative Deep Learning have been two of the most powerful tools for sound generation tasks. However, they have limitations: Evolutionary Algorithms require complicated designs, posing challenges in control and achieving realistic sound generation. Generative Deep Learning models often copy from the dataset and lack creativity. In this paper, we propose *LVNS-RAVE*, a method to combine Evolutionary Algorithms and Generative Deep Learning to produce realistic and novel sounds. We use the RAVE model as the sound generator and the VGGish model as a novelty evaluator in the Latent Vector Novelty Search (LVNS) algorithm. The reported experiments show that the method can successfully generate diversified, novel audio samples under different mutation setups using different pre-trained RAVE models. The characteristics of the generation process can be easily controlled with the mutation parameters. The proposed algorithm can be a creative tool for sound artists and musicians.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Information systems** → *Multimedia content creation.*

KEYWORDS

Neural audio synthesis, variational autoencoder, latent vector evolution

ACM Reference Format:
Jinyue Guo, Anna-Maria Christodoulou, Bálint Laczkó, and Kyrre Glette . 2024. LVNS-RAVE: Diversified audio generation with RAVE and Latent Vector Novelty Search. In *Genetic and Evolutionary Computation Conference (GECCO '24 Companion), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3638530.3654432

1 INTRODUCTION

Since the introduction of *wavenet* [12], generative audio Deep Learning models [4, 14] have been developing rapidly. By efficiently modeling the temporal distribution of audio samples, these models can generate realistic, high-fidelity audio outputs. The RAVE model [2], with improvements of the Variational Autoencoder (VAE), excels in balancing fidelity and computational complexity. However, the

GECCO '24 Companion, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0495-6/24/07.
https://doi.org/10.1145/3638530.3654432

nature of these probability models limits their ability to generate samples similar to their training dataset, often lacking diversity and novelty.

Evolutionary Algorithms provide an alternative for audio generation. They can generate novel audio samples without the limitation of pre-defined datasets. Parameter values for synthesizer models are evolved in [3] and [11], and compositional pattern-producing networks (CPPNs) are used for synthesis with interactive evolution in [8]. The Sound Innovation Engines approach [7] employs Quality-Diversity search and automatic fitness evaluation. However, in all of these approaches the quantitative quality measurement of samples remains a challenge.

To combine the fidelity of Deep Learning models and the diversity of Evolutionary Algorithms, Bontrager et al. [1] introduced Latent Variable Evolution (LVE). Using the VAE model, they generated *MasterPrints*, a set of natural and synthetic fingerprints. The VAE learned a latent representation of the original fingerprint images, ensuring any latent vector could be decoded into a meaningful fingerprint image. The LM-MA-ES algorithm then evolved latent vectors to match most fingerprint images. LVE was also applied to other fields such as game level generation [13, 15], illustrating LVE's potential for generating diversified and high-quality samples.

In this paper, we make an initial exploration of applying LVE to audio generation. Using Novelty Search to evolve the latent vectors of the RAVE model and a deep learning-based audio distance measure, we propose the LVNS-RAVE [1] method that can generate realistic and diversified audio samples.

2 METHODS

The RAVE model has an autoencoder structure with two parts of networks: the encoder and the decoder. The encoder can be seen as a compression network that extracts meaningful information from input audio waveforms to form the latent vectors $\mathbf{s}_i \in \mathbb{R}^{d \times l}$, where d represents the RAVE latent dimension, usually between 8 and 20 according to the pre-trained model, and l represents the sequence length that is fixed at 64. In our experiment, we utilize the latent vectors as genotypes and use the RAVE decoder network to map them back to audio waveforms or phenotypes. The variational modeling and generative adversarial fine-tuning of RAVE create a smooth latent space with semantic meanings so that all vectors in the latent space are mapped to meaningful phenotypes and that vectors with smaller distances are more similar. These properties make it ideal for the evolutionary algorithm to explore the latent space and find diversified phenotypes.

We chose the Novelty Search algorithm [9, 10] to evolve the latent vectors of RAVE. This decision stems from our primary objective to generate diversified and realistic audio samples. Common Evolutionary Algorithms (CMA-ES or LM-MA-ES) geared towards optimizing a single metric, do not align with our diversity-centric goal. On the other hand, although proven powerful in generating realistic audio samples from a given training set, the RAVE model often lacks diversity and novelty in their output. Since defining a qualitative measurement for sound is difficult, we want to use the realistic RAVE models to guarantee quality. At the same time, the evolutionary part can focus on improving diversity. We modified the Novelty Search algorithm's original crossover and selection process to adapt for the RAVE latent vector sequences. The LVNS-RAVE process, depicted in Figure 1, includes the following four stages:

Initialization. We define the container as a set $\mathbf{C} = \{\mathbf{s}_1, \mathbf{s}_2, \dots \mathbf{s}_n\}$, with n denoting the sample count. Container initialization involves two methods: drawing samples from a multivariate normal distribution or extracting latent vectors from a dataset. The latter involves randomly selecting samples and using the RAVE encoder to convert audio waveforms into latent embedding sequences.

Crossover and Mutation. We define genotypes as matrices in our experiments; $\mathbf{s}_i \in \mathbb{R}^{d \times l}$. Using columns as the basic elements for crossover and mutation processes, each column represents a latent vector of an audio frame. To generate a new genotype, we randomly select two samples $\mathbf{s}_i$ and $\mathbf{s}_j$ from the container $\mathbf{C}$ as parents, with all samples having equal possibilities. Then, we pick a random breakpoint k and assign the columns before or at m from $\mathbf{s}_i$, and columns after m from $\mathbf{s}_j$, to the child $\mathbf{s}^*$. After the crossover, we select one column in $\mathbf{s}^*$ to add noise from a normal distribution as the mutation. We keep generating new breeds and add them to the set $\mathbf{C}_{new}$, until the set has N elements.

Novelty Evaluation. To measure the novelty of newly generated breeds, we use the VGGish model[6] that was pre-trained on AudioSet[5] as a general auditory measurement model. The 128-dimension vector before the classification head serves as the *perception vector*. Since the VGGish model was trained on general audio samples, unlike the specifically trained RAVE models, we expect it to output the "perceptual understanding" of the evolved samples. We use the Euclidean distance between the VGGish embeddings of the two audio samples as their similarity score, with lower distances meaning more similar samples. The distance between genotype $\mathbf{s_i}$ and $\mathbf{s_j}$ is given by:

$$\text{dist}(\mathbf{s}_i, \mathbf{s}_j) = ||\text{VGGish}(\text{RDec}(\mathbf{s}_i)) - \text{VGGish}(\text{RDec}(\mathbf{s}_j))|| \quad (1)$$

where $RDec()$ represents the pre-trained RAVE decoder.

Then, we measure the novelty by calculating the sparseness ρ of each sample, which is the average distance of each sample to its k-nearest neighbors with the container $\mathbf{C}$ and the set of newly generated breeds $\mathbf{C}_{new}$:

$$\begin{gathered} \rho(\mathbf{s}_i) = \frac{1}{k} \sum_{j \in \mathbf{U}} \text{dist}(\mathbf{s}_i, \mathbf{s}_j), \\ \text{where } \mathbf{U} = \text{knn}(\mathbf{s}_i, \mathbf{C} \bigcup \mathbf{C}_{new}) \end{gathered} \quad (2)$$

$\text{knn}(\text{s}, \text{C})$ represents the k-nearest neighbors of point $\mathbf{s}$ in the data collection set C. We use $\mathbf{k = 50}$ in all following experiments.

Selection and Update. After calculating the sparseness of every sample in $\mathbf{C}_{new}$, we define the new container $\mathbf{C}^*$ with samples that has the highest sparseness values:

$$\mathbf{C}^* = \text{argmax}_{\mathbf{C}' \subset \mathbf{C} \bigcup \mathbf{C}_{\text{new}}, |\mathbf{C}'| = N} \sum_{\mathbf{s}_i \in \mathbf{C}'} \rho(\mathbf{s}_i) \quad (3)$$

where N is the size of the container.

[1] Code and audio samples can be found at https://github.com/fisheggg/LVNS-RAVE.

3 EXPERIMENTS AND EVALUATION

We conducted the evolutionary experiments on three different pre-trained RAVE models[2]: *vintage*, *darbouka_onnx*, and *VCTK*. The *vintage* model was trained on 80 hours of vintage music, the *darbouka_onnx* model was trained on 8 hours of darbouka recordings, and the *VCTK* model was trained on CSTR's VCTK Corpus. We used the pytorch implementation[3] of the VGGish model. The evolution process was computed using a server with 8 CPUs and no GPU. Table 1 shows the setups used in the experiments.

Table 1: Experiment setups of Latent Vector Novelty Search.

Setup No.	Number of Generations	Container Size	New Breed Size	Container Initialization Method
1	100	50	30	random
2	100	50	30	dataset
3	300	50	30	dataset
4	100	500	50	dataset

We conducted two groups of experiments: Group 1 used different pre-trained RAVE models under the same setup; Group 2 used the same *VCTK* RAVE model under different setups. Figure 2 and 3 show the sparseness of the containers of each group along the evolutionary generations. Figure 4 plots the distribution of embedding vectors during the evolution process in group 2, Setup 3, and highlights some of the trajectories. Due to time limitations, the results are from single runs to demonstrate insights, but more repetitions will be needed to ascertain statistically significant differences.

Figure 2 shows the sparseness plot using different pre-trained RAVE models. Since the containers were randomly initialized, all experiments had an initial sparseness of around 1.8. Over generations, the mean sparseness of the containers grew, while their standard deviations staid in the same range. This indicates that the VGGish evaluation model can create environmental pressure, leading to more diversified species. The growth rate varies but is present in different models. Lower growth rates do not necessarily correspond to less diversified samples since the space of the evaluation vectors is different for each model. For example, the *vintage* model creates audio samples with similar frequency ranges, which the VGGish model perceived as less sparse. That does not imply that the evolution results of the *vintage* experiment are less diversified than the other experiments. It is only meaningful to compare the relative levels of diversity between the initial and the final container of the same experiment.

Figure 3 shows the sparseness of the four different setups using the same *VCTK* model. In Setup 3, the growth of sparseness slowed down after the first 100 generations, but we did not observe a limit to the growth. Setup 4 used a larger container size of 500 instead of 50, with the number of new breeds being increased accordingly. The initial sparseness of Setup 4 is higher than the other setups, but there is no apparent difference after 100 generations. We observe a more linear growth of Setup 3 than the others. Compared to the random initialization of Setup 1, the other setups that use samples from the original dataset have larger initial sparseness. However, after 100 generations, there is no significant difference between the two initialization methods.

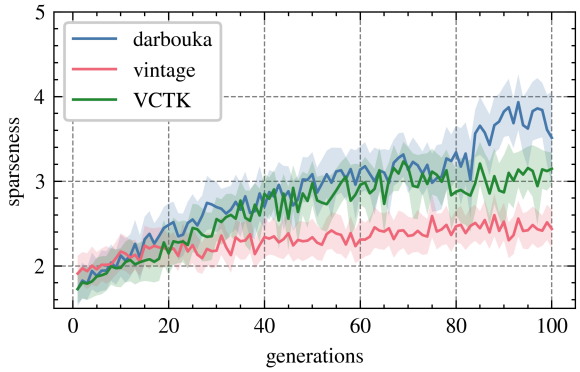

Figure 2: Experiment group 1: Sparseness of the containers during the evolution process of three different pre-trained models, all using experiment Setup 1. Lines show the mean sparseness; shadow areas show the standard deviation.

Figure 4 plots the distribution of embedding vectors during evolution and highlights some trajectories. Over generations, the range of the embedding vectors expanded beyond the original dataset's distribution, while the directions and shapes of the embedding sequences have also become more diversified. We also observed much sparser embedding distributions since the first 40 generations.

4 CONCLUSION AND FUTURE WORK

This paper introduced our initial effort to apply latent variable evolution to deep audio generation using Novelty Search on RAVE latent sequences. We used an external evaluation network to measure the similarity between samples and calculated the sparsenesses of each phenotype. Our exploratory experiments indicate that Latent Variable Evolution selecting for Novelty using a VGGish distance measure can lead to a notable increase in the diversity of the generated results. We tested the method over three pre-trained RAVE models and four experiment setups. The algorithm is flexible to produce high-quality sound samples with various characteristics, which is preferable for artistic purposes.

In future work, we will collect more data from multiple evolutionary runs and explore a broader range of hyperparameters. Larger containers with thousands of samples could lead to different behaviors. Since the RAVE models use a sequence of latent vectors as the embeddings, we have adjusted the original crossover and mutation process. However, there might be better mutation methods that can further increase the diversity of phenotypes. Finally, quality evaluation has been a consistent problem for audio generation. Human assessment could be involved to provide better guidance to the generator.

ACKNOWLEDGMENTS

This project is supported by the Research Council of Norway through projects 262762 (RITMO) and 324003 (AMBIENT).

[2]https://acids-ircam.github.io/rave_models_download
[3]https://github.com/harritaylor/torchvggish/tree/master

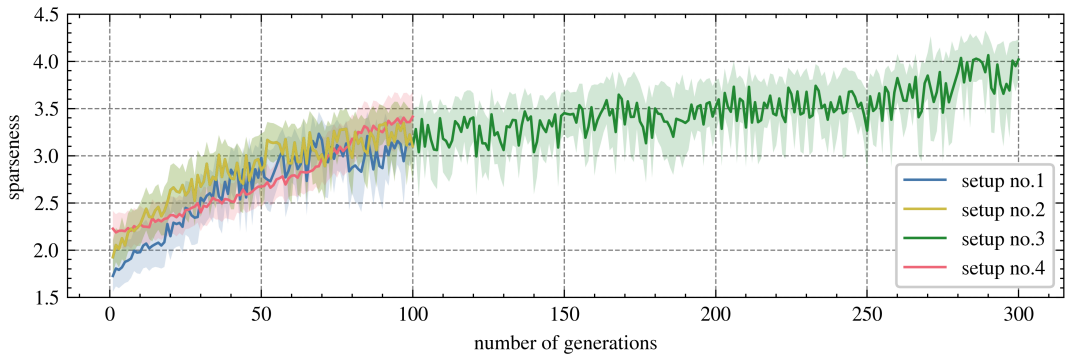

Figure 3: Experiment group 2: Sparseness of the containers during the evolution process using VCTK pre-train model, under 4 different experiment setups. Lines show the mean sparseness and shadow areas show the standard deviation.

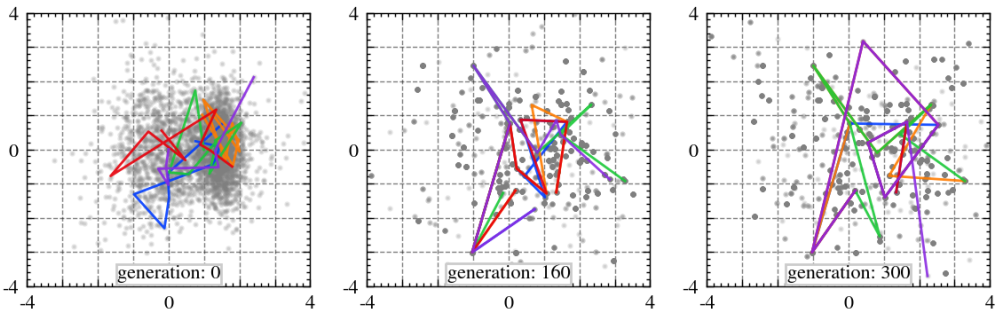

Figure 4: Plot of the first two dimensions of the container embedding sequences, generated using Setup 3. Colored lines show the trajectories of five sequences. Gray dots show the distribution of all other embedding vectors in the container.

REFERENCES

[1] Philip Bontrager, Aditi Roy, Julian Togelius, Nasir Memon, and Arun Ross. 2018. DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution. In *IEEE 9th Intl. Conf. on Biometrics Theory, Applications and Systems (BTAS)*. 1–9. https://doi.org/10.1109/BTAS.2018.8698539 ISSN: 2474-9699.

[2] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. https://doi.org/10.48550/arXiv.2111.05011 arXiv:2111.05011 [cs, eess].

[3] Palle Dahlstedt. 2007. Evolution in Creative Sound Design. In *Evolutionary Computer Music*, Eduardo Reck Miranda and John Al Biles (Eds.). Springer London, London, 79–99. https://doi.org/10.1007/978-1-84628-600-1_4

[4] Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. http://arxiv.org/abs/2307.04686 arXiv:2307.04686 [cs, eess].

[5] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430

[7] Björn Þór Jónsson, Çağrı Erdem, Stefano Fasciani, and Kyrre Glette. 2024. Towards Sound Innovation Engines Using Pattern-Producing Networks and Audio Graphs. In *Artificial Intelligence in Music, Sound, Art and Design*, Colin Johnson, Sérgio M. Rebelo, and Iria Santos (Eds.). Springer Nature Switzerland, Cham, 211–227.

[8] Björn Þór Jónsson, Amy K. Hoover, and Sebastian Risi. 2015. Interactively Evolving Compositional Sound Synthesis Networks. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation (GECCO '15)*. Association for Computing Machinery, New York, NY, USA, 321–328. https://doi.org/10.1145/2739480.2754796

[9] Joel Lehman and Kenneth O. Stanley. 2011. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19, 2 (June 2011), 189–223. https://doi.org/10.1162/EVCO_a_00025

[10] Joel Lehman and Kenneth O. Stanley. 2011. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation (GECCO '11)*. Association for Computing Machinery, New York, NY, USA, 211–218. https://doi.org/10.1145/2001576.2001606

[11] James McDermott, Niall J. L. Griffith, and Michael O'Neill. 2008. Evolutionary Computation Applied to Sound Synthesis. In *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, Juan Romero and Penousal Machado (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 81–101. https://doi.org/10.1007/978-3-540-72877-1_4

[12] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. https://doi.org/10.48550/arXiv.1609.03499 arXiv:1609.03499 [cs].

[13] Anurag Sarkar and Seth Cooper. 2021. Generating and Blending Game Levels via Quality-Diversity in the Latent Space of a Variational Autoencoder. In *Proceedings of the 16th International Conference on the Foundations of Digital Games (FDG '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3472538.3472545

[14] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion. https://doi.org/10.48550/arXiv.2301.11757 arXiv:2301.11757 [cs, eess].

[15] Takumi Tanabe, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. 2021. Level generation for angry birds with sequential VAE and latent variable evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '21)*. Association for Computing Machinery, New York, NY, USA, 1052–1060. https://doi.org/10.1145/3449639.3459290