MULTI-SCALE LATENT TRAJECTORY MODELLING FOR AUDIO TEXTURE SYNTHESIS

Aaron DEES[1], **Susan MCKEEVER**[1], and **Seán O'LEARY**[1]

[1]**Technological University of Dublin**, Dublin, Ireland

ABSTRACT

Audio textures exhibit complex temporal dynamics that are not well represented by single-scale models. We propose a multi-scale latent trajectory approach for audio texture synthesis using latent representations extracted from a pre-trained neural audio codec. By decomposing latent trajectories into slow structural and fast stochastic components, we enable direct modeling and interpretation of the long-term temporal structure and fine-grained detail of audio textures.

Our experiments on environmental audio textures suggest that the proposed method significantly improves audio quality over a single-scale baseline and achieves performance competitive with a latent diffusion model. Using perceptually motivated summary statistics, we demonstrate that generated continuations preserve texture identity.

Through the proposed methods, we enable continuous texture generation from a given context and expose a latent structure that may support control over long-term dynamics.

1. INTRODUCTION

Audio textures represent a diverse and expansive class of sounds that, while often unnoticed, play a fundamental role in shaping everyday sonic experiences. They are integral to sound design across media such as film, video games, interactive art installations, and musical performance, where they contribute to immersive soundscapes and sonic environments.

Traditional workflows rely on pre-recorded samples obtained through field recording or Foley techniques, constraining creative flexibility and temporal manipulation.

Despite recent advances in neural audio synthesis, modeling long-term temporal structure in audio textures remains challenging, and controllable generation is still limited.

Neural audio codecs [1,2] provide a compelling representation for audio modeling by compressing raw waveforms into compact latent sequences. In this work, we treat audio textures as trajectories with separable temporal structure, observing that they exhibit locally varying stochastic detail alongside slowly evolving global dynamics.

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Motivated by this observation, we explicitly decompose codec latent trajectories into slow and fast components, enabling targeted modeling at distinct temporal scales.

We conduct experiments on three categories of environmental audio textures, sea waves, rain, and thunder, which exhibit diverse temporal and spectral characteristics. A manually curated dataset is used to ensure high audio quality and to minimize interference from secondary sound sources.

Experimental results show that the proposed approach achieves synthesis quality competitive with a latent diffusion baseline, while significantly outperforming a single-scale model. We further demonstrate improved preservation of texture identity using perceptually motivated summary statistics.

Finally, we conduct latent interpolation experiments suggesting that the learned slow-rate representation captures meaningful variations in temporal structure of environmental textures.

Together, these contributions provide a pathway toward high-quality, continuous, and controllable audio texture synthesis, addressing key limitations in existing approaches.

A supplementary website containing audio examples for all experiments and additional materials is provided [1] .

2. RELATED WORK

Audio textures have traditionally been defined as signals exhibiting local variability with stable long-term statistics [3,4]. Traditional approaches often leverage this tiered structure for re-synthesis of audio textures [5]. The perceptual importance of temporal structure in environmental sound was established well before such statistical models, Warren and Verbrugge [6] showed that listeners categorize events on the basis of temporal patterning alone, with spectral differences held constant. A significant advance in perceptual modeling was introduced by McDermott and Simoncelli [7], who proposed a set of auditory summary statistics capable of capturing the perceptual identity of environmental sounds. These statistics have been used to synthesize textures that are perceptually similar to real recordings, despite lacking exact waveform correspondence [7, 8].

These early approaches to audio texture modeling largely assume statistical stationarity, focusing on reproducing time-averaged characteristics while neglecting the evolution of structure over longer temporal scales. Subsequent

[1] https://aaron-dees.github.io/multi-scale-latent-trajectory-model/

work extended the static definition to include the concept of dynamic textures, where long-term structure evolves over time [9, 10].

Recent advances in neural audio synthesis have enabled high-quality generation of complex sounds, including textures. Neural approaches to audio texture synthesis and re-synthesis have demonstrated the ability to reproduce perceptual characteristics of target sounds [9, 11–14]. However, many of these methods focus primarily on re-synthesis or short-duration generation, and do not explicitly approach modeling of temporal structure. As a result, they do not capture long-term dependencies or provide modeling of evolving texture dynamics. This limitation motivates approaches that explicitly model temporal structure.

Modeling temporal structure across multiple scales has been explored in various domains of audio generation, such as speech synthesis and music generation. Hierarchical recurrent models and multi-scale sequence models have been proposed to capture long-range dependencies while maintaining local detail [10, 15, 16]. However, the application of explicit multi-scale decomposition to audio texture synthesis remains underexplored, particularly in conjunction with compressed and powerful latent representations.

Neural audio codecs have emerged as a powerful representation for audio, learning compact and perceptually meaningful latent spaces through end-to-end compression [1, 2, 17]. These representations enable efficient downstream modeling by reducing the dimensionality and complexity of raw waveforms.

Generative models operating in these latent spaces have demonstrated strong performance in audio synthesis tasks, such as latent diffusion approaches [18–20]. These approaches have achieved high-quality synthesis but typically require computationally intensive sampling procedures and offer limited direct control over the evolving temporal structure.

Control in generative audio models has been explored through conditioning mechanisms, latent space manipulation, and representation learning [21–23]. While these approaches enable control over attributes such as pitch, timbre, or style, controlling the temporal evolution of audio textures still requires explicit temporal modeling.

In contrast to prior approaches, which either focus on perceptual matching, short-term synthesis, or implicit temporal modeling, our method explicitly decomposes and models texture dynamics across temporal scales within a learned latent representation. This enables both high-quality synthesis and supports control over long-term structure.

3. PROPOSED METHOD

Our method enables perceptually coherent, controllable texture synthesis by modelling a neural audio codec's latent trajectories at multiple temporal scales. The framework consists of a slow-rate variational autoencoder, which captures long-term structural evolution and learns a compact latent representation for interactive manipulation, and a fast-rate conditional autoregressive model, which generates local stochastic detail. Together, these components enable structured yet diverse texture generation.

3.1 Latent Trajectory Decomposition

Let $z_T \in \mathbb{R}^D$ denote a latent trajectory extracted from a pre-trained neural audio codec, in our case, we use Music2Latent [2], that compresses waveforms sampled at 44.1kHz into time-varying latent representations with a time compression ratio of 4096. We use the official implementation and weights provided [2] .

We observe that the latent trajectories extracted from environmental audio textures exhibit structured temporal dynamics. Statistical analysis shows that these trajectories comprise both slowly varying components associated with long-term structural evolution and rapidly fluctuating components reflecting local stochastic variation.

To illustrate this, we apply Gaussian low-pass filtering to the latent trajectories. This allows the trajectory to be decomposed into low- and high-frequency components, resulting in the approximate additive split:

$$z_T = z_T^{\text{slow}} + z_T^{\text{fast}}, \tag{1}$$

where z_T^{slow} is obtained by low-pass filtering z_T and z_T^{fast} is defined as the residual component. Here slow and fast are relative to the latent frame rate of approximately 10.8 Hz, which limits the trajectory to variation below approximately 5.4 Hz. The fast component therefore models stochastic modulation of the texture rather than audio-rate detail, which is reconstructed by the codec decoder. We note that the codec latent space is non-linear, so this split is a working approximation rather than an exact structural decomposition. To analyze the effectiveness of this approximate decomposition, we compute several metrics designed to empirically measure the slow and fast components correlation.

These include low-frequency energy ratio, spectral flatness, and decorrelation lag. Decorrelation lag is the lag at which a component's autocorrelation first falls below a fixed threshold.

We compute these metrics across a range of filter kernel sizes in order to determine which parameters produce the most effective separation between slow and fast components. The results are shown in Figure 1.

The results suggest that, for kernel sizes in the range of 9–25, latent trajectories can be effectively decomposed into slow and fast components that are weakly correlated and exhibit distinct statistical properties. Across our dataset, the slow component consistently shows high low-frequency energy concentration, long decorrelation lags, and low spectral flatness, indicating structured long-term dynamics. In contrast, the residual fast component exhibits short decorrelation lags and higher spectral flatness, suggesting locally varying dynamics with shorter temporal memory.

Based on these observations, we select a kernel size of 15 as it provides a good balance between preserving long-

[2] https://github.com/SonyCSLParis/music2latent

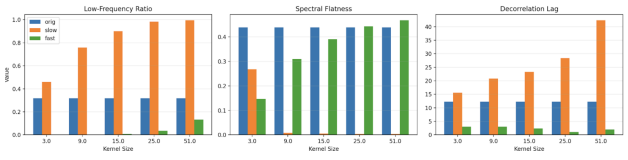


Figure 1. Analysis of trajectory decomposition across Gaussian filter kernel sizes. For each kernel size we report low-frequency energy ratio, spectral flatness, and decorrelation lag for the full, slow, and fast components. Kernel sizes of 9–25 yield slow and fast components with distinct statistical properties; we select 15.

term structure in the slow component and isolating stochastic variation in the fast component.

3.2 Modeling Overview

Having established the slow–fast decomposition, we now describe how the two components are combined at synthesis time; Figure 2 gives a high-level overview. Given an encoded context, the slow-rate model first generates a plausible slow trajectory. The fast-rate model then produces the corresponding residual trajectory conditioned on the generated slow component, and the final codec latent trajectory is obtained by summing the two. This formulation enables stable continuation from context while allowing diverse realizations of the same texture through stochastic fast-rate generation, explicitly separating structure from variation.

Continuations of arbitrary length can be achieved by appending the generated slow trajectory to the context trajectory. By applying a sliding context window, the model can recursively extend the trajectory, allowing textures of arbitrary duration to be synthesized.

3.3 Slow-Rate Modeling

The objective of the slow-rate model is to generate a latent trajectory of length T_g conditioned on a context trajectory of length T_{ctx}.

The slow component is modeled using a variational autoencoder (VAE) designed to capture long-term temporal structure. The encoder consists of a bidirectional GRU followed by linear projection layers that produce the parameters of a Gaussian latent distribution, μ_e and $\log \sigma_e^2$. This maps the input context trajectory $z^{\text{slow}}_{1:T_{\text{ctx}}}$ to a latent representation e, sampled via the reparameterization trick:

$$e = \mu_e + \sigma_e \cdot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1),$$

where $\sigma_e = \exp(0.5 \log \sigma_e^2)$.

The decoder generates a continuation of the slow trajectory, $\hat{z}^{\text{slow}}_{1:T_g}$, conditioned on the sampled latent vector e. Specifically, e is combined with a learned time embedding representing generation steps and passed through a recurrent neural network to produce the output sequence.

This formulation encourages the latent variable e to capture the long-term structure of the trajectory, providing a compact and continuous representation that may be explored for control and interpolation in future works.

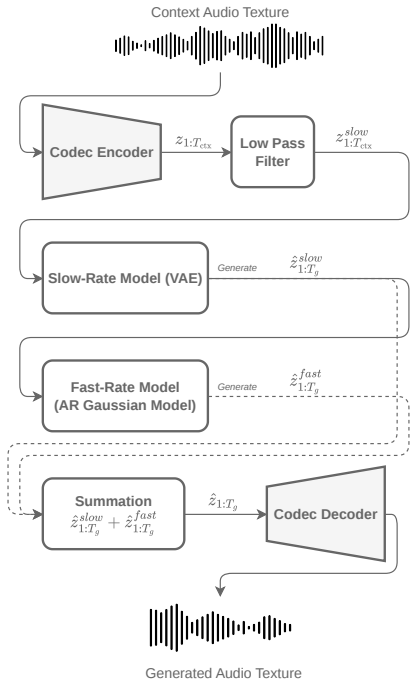


Figure 2. Overview of our proposed modeling approach

A detailed diagram of the slow-rate model can be found on the supplementary website.

3.4 Fast-Rate Modeling

The objective of the fast-rate model is to generate the high-frequency residual component corresponding to the generated slow trajectory $\hat{z}^{\text{slow}}_{1:T_g}$. This component captures the stochastic local variation of the latent trajectory.

We model the residual sequence autoregressively. The conditional distribution is factorized as

$$p(z^{\text{fast}}_{1:T_g} \mid \hat{z}^{\text{slow}}_{1:T_g}) = \prod_{i=1}^{T_g} p(z^{\text{fast}}_i \mid z^{\text{fast}}_{i-1}, \hat{z}^{\text{slow}}_i).$$

At each time step i, the model receives as input the slow trajectory sample $\hat{z}^{\text{slow}}_i$, the previously generated residual $\hat{z}^{\text{fast}}_{i-1}$, and a learned time embedding. These inputs are concatenated and passed through an RNN, which outputs the parameters of a diagonal Gaussian distribution, μ_i and $\log \sigma_i^2$.

The residual sample is then obtained via

$$\hat{z}^{\text{fast}}_i = \mu_i + \sigma_i \cdot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1),$$

where $\sigma_i = \exp(0.5 \log \sigma_i^2)$.

This process is repeated autoregressively until all T_g residual samples are generated, producing a stochastic trajectory conditioned on the slow component.

A detailed diagram of the fast-rate model can be found on the supplementary website.

3.5 Training Strategy

The slow- and fast-rate models are trained in a staged manner to stabilize learning.

First, the slow-rate VAE is trained independently using a reconstruction objective with KL regularization and KL annealing. The reconstruction loss consists of two components: a mean squared error (MSE) between the predicted and ground-truth trajectories, and a velocity loss that penalizes differences between temporal derivatives. The combined reconstruction loss is defined as:

$$L_{\text{recon}} = \lambda_{\text{pos}} \,\text{MSE}(z, \hat{z}) + \lambda_{\text{vel}} \,\text{MSE}(\Delta z, \Delta \hat{z})$$

where, $\Delta z = z_t - z_{t-1}$.

The regularization term is given by the Kullback–Leibler divergence between the approximate posterior and a unit Gaussian prior:

$$L_{\text{KL}} = D_{\text{KL}}(q(e|z) \,||\, \mathcal{N}(0, I)).$$

The overall objective is:

$$L = L_{\text{recon}} + \beta L_{\text{KL}}.$$

Once the slow-rate model has converged, the fast-rate model is trained to generate the residual trajectories conditioned on the slow component. The fast-rate model is trained using a Gaussian negative log-likelihood (NLL) objective and scheduled teacher forcing to improve stability.

The training objective is defined as:

$$L_{\text{NLL}} = -\log p(z^{\text{fast}} | z^{\text{slow}})$$

This objective encourages the model to assign high probability to the observed residual trajectories while learning an appropriate uncertainty for each latent dimension.

4. EXPERIMENTS

We evaluate the proposed method along two key axes: perceptual audio quality and preservation of texture identity under continuation. Models are trained and evaluated both on *individual* texture categories and on a *combined* dataset containing all three categories.

We compare our proposed method against two baselines designed to evaluate both the effectiveness of the hierarchical trajectory decomposition and the quality of the generated audio against a state-of-the-art generative approach.

4.1 Datasets

4.1.1 Curated Freesound Dataset

We construct a curated dataset of environmental audio textures spanning three categories: *Sea Waves*, *Rain*, and *Thunder*. These categories were selected to capture a range of spectral and temporal characteristics commonly found in environmental textures.

All audio samples were sourced from Freesound [3]. The dataset was manually curated to ensure accurate category representation while minimizing interference from secondary sound sources.

Each category contains 450 unique recordings, each with a duration of 30 seconds. In total, the dataset contains 1,350 recordings, resulting in 11.25 hours of audio.

[3] www.freesound.org

4.1.2 Codec Latent Representation

Our proposed method models trajectories in the latent space of a neural audio codec rather than directly operating on raw audio waveforms. All audio recordings are therefore encoded into latent trajectories, z_T, using the pre-trained neural audio codec Music2Latent [2].

Music2Latent compresses audio by a factor of 4096. Consequently, a 30-second audio recording sampled at 44.1 kHz is encoded as a latent sequence of length 321.

Training samples are obtained by applying a sliding window of size $T_{\text{ctx}}+T_g$ over each recording's latent trajectory, with a stride of 15 latent steps, producing 14,850 sequences in total.

Latent trajectories are decomposed into slow and fast components using a Gaussian low-pass filter (with kernel size 15), as motivated by the analysis in Section 3.1.

4.2 Implementation Details

In total, eight models are trained: four slow-rate models and four fast-rate models, one for each individual sound category and the combined dataset. This enables us to assess the model's ability to generalize across texture classes when trained jointly on all categories.

During training, we use a context length of 10 seconds corresponding to a latent sequence length of $T_{\text{ctx}} = 107$. The model then generates a continuation of 5 seconds, corresponding to $T_g = 53$ latent steps.

The dataset is split into 80% for training, 10% for validation and 10% for testing. All models are trained for 200,000 optimization steps on an NVIDIA RTX A5500 GPU.

4.3 Evaluation Metrics

To evaluate the proposed method, we assess both the audio quality of generated samples and the extent to which they preserve the texture identity of the reference signal.

4.3.1 Audio Quality

To measure audio quality, we use the Fréchet Audio Distance (**FAD**), [24]. The FAD metric evaluates the quality of generated audio by measuring the distributional similarity between real and synthesized audio features. Inspired by the Fréchet Inception Distance (FID), FAD computes distribution statistics of embeddings extracted from a pre-trained neural network. The Fréchet distance between these distributions is used to approximate perceptual similarity, where lower FAD values indicate higher fidelity. We calculate the FAD score, using embeddings from a VGGish model to directly compare the distributions of generated audio with target audio [4].

4.3.2 Texture Identity

The second evaluation task measures whether generated continuations preserve the perceptual identity of the input texture. Specifically, we assess whether the generated audio is consistent with the same underlying process as the provided context.

[4] https://github.com/gudgud96/frechet-audio-distance

To evaluate this, we compare time-averaged summary statistics of the context and continuation signals, following the framework of McDermott and Simoncelli [7, 8]. These statistics are computed on an auditory representation based on a cochlear filterbank and have been shown to capture perceptually relevant properties of environmental sound textures.

We compute four classes of statistics:

- Cochlear-envelope marginal moments (mean, variance, skewness, kurtosis)
- Cochlear-envelope correlations across frequency channels
- Modulation power within each cochlear channel
- Modulation correlations, including (i) correlations across cochlear channels within each modulation band (C1), and (ii) correlations across modulation bands within each cochlear channel (C2)

For each audio segment, the summary statistics are computed independently, and distances are evaluated for each statistic class separately using Euclidean distance.

We report five distances for each sound category: (1) context versus generated audio from the proposed model, (2) context versus the ground-truth continuation from the same recording, (3-4) context versus audio generated by baseline models, and (5) context versus white noise.

Statistics are computed over 10-second segments of context audio and 15-second segments of comparison audio, consistent with [7].

A model is considered to preserve texture identity if the distance between context and generated audio is comparable to the distance between the context and the ground-truth continuation, and substantially smaller than that of an unrelated signal, such as white noise.

4.4 Baselines

Latent Diffusion Model As a strong generative baseline, we train a latent diffusion model, (**LDM**), on the codec latent trajectories from the combined dataset. Our implementation follows the architecture described in Mousai [18] and is based on the publicly available implementation [5] .

To perform trajectory continuation, we employ a diffusion inpainting strategy. Given a context trajectory of length T_{ctx}, the remaining portion of length T_g is masked and the diffusion model iteratively denoises the masked region conditioned on the provided context to generate a plausible continuation.

Single-Scale Model To evaluate the impact of the proposed multi-scale decomposition, we train a single-scale variant of our model that directly predicts the latent trajectory without separating slow and fast components. This model uses the same architecture as the slow-rate model but is trained to predict the full latent trajectory values $z_{1:T_g}$ conditioned on the context trajectory $z_{1:T_{\text{ctx}}}$.

[5] https://github.com/archinetai/audio-diffusion-pytorch

5. RESULTS

5.1 Audio Quality

First we evaluate the audio quality of generated continuations using Fréchet Audio Distance (FAD). Lower FAD values indicate greater similarity between the distributions of generated and reference audio. Results are presented in Table 1.

	Waves ***FAD***↓	*Thunder* ***FAD***↓	*Rain* ***FAD***↓	*Avg* ***FAD***↓
Single-Scale	2.825	3.2678	4.623	3.572
Multi-Scale	1.396	1.5298	1.364	1.430
Multi-Scale*	0.882	**0.852**	1.080	0.939
LDM	**0.754**	1.271	**0.728**	**0.918**

Table 1. FAD results across sound categories for different model configurations (↓ lower is better). * indicates models trained on individual categories; others are trained on the combined dataset.

The results suggest that the proposed multi-scale approach significantly improves audio quality compared to the single-scale baseline. The single-scale model achieves relatively high FAD values across all categories, indicating that directly modeling the latent trajectory without decomposition leads to poor continuation quality. In contrast, the proposed multi-scale model substantially reduces FAD, demonstrating the effectiveness of explicitly modeling temporal structure through multi-scale decomposition.

When trained on individual sound categories, the multi-scale model achieves an average FAD of 0.94, competitive to the latent diffusion baseline (0.92). This suggests that the proposed approach is competitive with state-of-the-art generative models despite using a substantially more efficient and interpretable generation procedure. We note that the proposed method generates trajectories in a single forward pass (for the slow component) followed by autoregressive refinement (for the fast component), resulting in significantly reduced computational overhead, when compared to the iterative sampling procedure required by diffusion techniques.

We also observe that training on individual categories consistently improves performance compared to training on the combined dataset, suggesting that category-specific models better capture the statistical structure of each texture class, and that more work is needed to improve audio quality of a generalized model.

5.2 Texture Identity

Figure 3 shows normalized distances between the summary statistics of context audio and five comparison signals. All distances are normalized by the context-to-reference distance, such that values ≈ 1 indicate similarity to the reference continuation, values < 1 indicate closer agreement with the context statistics, and values > 1 indicate greater dissimilarity. Note that the Reference values thus all $\equiv 1$.

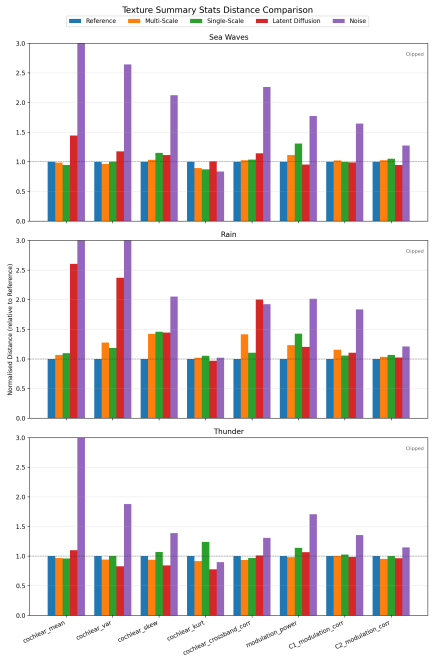


Figure 3. Texture summary statistics distance comparison (normalized). Values near 1 indicate similarity to the reference continuation. White-noise bars exceeding the axis limit are clipped for readability.

The proposed multi-scale model achieves distances close to unity across all statistic classes for sea waves and thunder, indicating that its continuations are statistically consistent with the ground-truth continuation. Rain proves the most challenging category for all models; here the multi-scale model shows moderate deviations, notably in cochlear skewness and cross-band correlation.

In comparison, the latent diffusion model exhibits substantially larger deviations from the reference, particularly for the rain category, where its marginal moments and cross-band correlations deviate substantially from the reference statistics. This indicates that explicitly modeling slow and fast structural dynamics separately improves the preservation of perceptually relevant texture statistics.

The single-scale model achieves distances close to the reference for some statistics, however, it exhibits greater variability across statistics and categories, with noticeable deviations in higher-order and modulation statistics. As expected, white noise produces substantially larger distances across all metrics.

6. TOWARD CONTROL

A key motivation of this work is to move toward controllable texture synthesis for interactive and creative applications. While the proposed model is not explicitly trained with control objectives, it exposes a structured and continuous latent representation, e, through the slow-rate model. This latent space captures long-term structural dynamics and provides a potential interface for controlling the evolution of generated textures.

6.1 Latent Interpolation

To probe the structure of the learned latent space, we perform latent interpolation experiments. While this does not constitute a full control interface, it provides insight into whether the latent representation encodes meaningful and manipulable dimensions of variation.

Given two context audio segments, we encode their corresponding slow trajectories to obtain latent representations e_1 (source) and e_2 (target). We then generate a series of intermediate latent codes via linear interpolation:

$$e_\alpha = (1-\alpha)e_1 + \alpha e_2, \quad \alpha \in [0,1].$$

For each interpolated latent, we decode a fixed-length trajectory segment (5 seconds), producing a sequence of independent generations corresponding to increasing values of α.

Results for these interpolations can be found on the supplementary webpage. The generated samples suggest smooth and consistent changes in spectral energy distribution and temporal dynamics across interpolation steps.

While further analysis is required to fully characterize the structure of the latent space, these results indicate that it is organized along perceptually meaningful dimensions of texture variation.

In contrast to prior approaches, which often rely on explicit conditioning and large amounts of labeled data, the proposed latent representation provides a flexible pathway toward controllable and interactive texture synthesis by enabling structured manipulation of long-term dynamics.

7. CONCLUSION

In this work, we introduced a multi-scale latent trajectory modeling approach for audio texture synthesis using neural audio codec representations. By decomposing latent trajectories into slow structural and fast stochastic components, the proposed method improves long-term coherence and short-term detail compared to single-scale approaches.

Experimental results show that the method achieves competitive audio quality with a latent diffusion baseline while significantly outperforming a single-scale model. Additionally, evaluation using perceptually motivated summary statistics indicates improved preservation of texture identity across both spectral and temporal characteristics.

Beyond synthesis quality, the model exposes a learned latent representation that captures temporal structure of audio textures. Latent interpolation experiments suggest that this representation captures meaningful variations in texture structure and may provide a foundation for future control of generated textures.

The proposed framework supports continuous and arbitrarily long generation through recursive context extension, enabling the synthesis of extended audio textures from a short context. Compared to diffusion-based approaches, the method offers a more informed and efficient framework for modeling temporal structure. We do not analyse error accumulation under recursive continuation; characterising long-term drift over extended horizons is left to future work.

While objective metrics provide useful proxies for evaluation, a comprehensive assessment of perceptual quality requires human listening studies. Future work will include controlled listening experiments, exploration of richer control mechanisms, and evaluation on more diverse texture datasets to assess generalization.

Overall, this work demonstrates the effectiveness of combining neural audio codecs with multi-scale temporal modeling, providing a promising direction for high-quality and potentially controllable audio texture synthesis.

Acknowledgments

This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

8. REFERENCES

[1] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Machine Learning Research*, vol. 2023, pp. —, 2023.

[2] M. Pasini, S. Lattner, and G. Fazekas, "Music2latent: Consistency autoencoders for latent audio compression," in *Proc. Int. Soc. Music Information Retrieval Conf. (ISMIR)*, San Francisco, USA, 2024, pp. —.

[3] N. Saint-Arnaud, "Classification of sound textures," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, Sep. 1995.

[4] D. F. Rosenthal, H. G. Okuno, H. Okuno, and D. Rosenthal, Eds., *Computational Auditory Scene Analysis: Proceedings of the IJCAI-95 Workshop*. CRC Press, 1995.

[5] S. O'Leary and A. Röbel, "A montage approach to sound texture synthesis," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2014.

[6] W. H. Warren and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: A case study in ecological acoustics," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 5, pp. 704–712, 1984.

[7] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," in *Neuron*, 2011, pp. 926–940.

[8] E. Gutiérrez, F. Font, X. Serra, and L. Wyse, "A statistics-driven differentiable approach for sound texture synthesis and analysis," in *Proc. Int. Conf. Digital Audio Effects (DAFx25)*, 2025, pp. 457–464.

[9] H. Caracalla and A. Roebel, "Sound texture synthesis using ri spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 416–420.

[10] M. Huzaifah and L. Wyse, "Mtcrnn: A multi-scale rnn for directed audio texture synthesis," *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2011.12596

[11] C. Gupta, P. Kamath, and L. Wyse, "Signal representations for synthesizing audio textures with generative adversarial networks," in *Proc. Sound and Music Computing Conf. (SMC)*, 2021.

[12] Y. Liu, C. Jin, and D. Gunawan, "Ddsp-sfx: Acoustically-guided sound effects generation with differentiable digital signal processing," in *Proc. Int. Conf. Digital Audio Effects (DAFx24)*, 2024, pp. 216–221.

[13] A. Dees and S. O'Leary, "Generative latent spaces for neural synthesis of audio textures," in *Proc. Int. Conf. Digital Audio Effects (DAFx25)*, Ancona, Italy, 2025, pp. 419–426.

[14] A. Barahona-Ríos and T. Collins, "Noisebandnet: Controllable time-varying neural synthesis of sound effects using filterbanks," in *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2024, pp. 1573–1585.

[15] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[16] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[17] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," in *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2022, pp. 495–507.

[18] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Efficient text-to-music diffusion models," in *Proc. Annual Meeting Assoc. Computational Linguistics (ACL)*, 2024, pp. 8050–8068.

[19] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 21 450–21 474.

[20] J. Nistal, M. Pasini, C. Aouameur, M. Grachten, and S. Lattner, "Diff-a-riff: Musical accompaniment co-creation via latent diffusion models," in *Proc. Int. Soc. Music Information Retrieval Conf. (ISMIR)*, 2024, pp. 272–280.

[21] J. Engel, K. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Gansynth: Adversarial neural audio synthesis," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[22] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Generative timbre spaces: Regularizing variational auto-encoders with perceptual metrics," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2018.

[23] S. Nabi, N. Demerlé, G. Peeters, F. Bevilacqua, and P. Esling, "Adding temporal musical controls on top of pretrained generative models," in *Proc. Int. Soc. Music Information Retrieval Conf. (ISMIR)*, 2025, pp. 793–800.

[24] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019, pp. 2350–2354.