GENERATIVE LATENT SPACES FOR NEURAL SYNTHESIS OF AUDIO TEXTURES

Aaron Dees and Seán O'Leary

Dept. of Computer Science
Technological University of Dublin
Dublin, Ireland
d22127229@mytudublin.ie

ABSTRACT

This paper investigates the synthesis of audio textures and the structure of generative latent spaces using Variational Autoencoders (VAEs) within two paradigms of neural audio synthesis: DSP-inspired and data-driven approaches. For each paradigm, we propose VAE-based frameworks that allow fine-grained temporal control. We introduce datasets across three categories of environmental sounds to support our investigations. We evaluate and compare the models' reconstruction performance using objective metrics, and investigate their generative capabilities and latent space structure through latent space interpolations.

1. INTRODUCTION

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

Traditionally, the creation of audio textures has relied heavily on pre-recorded samples, obtained through field recordings or Foley design. This can be somewhat limiting in regards to the sonic and creative possibilities available to sound designers.

To expand the sonic palette available to them, many sound designers look to sound synthesis approaches, including physical modeling and digital signal processing (DSP) techniques. These methods allow for greater flexibility but often require extensive parameter tuning and domain expertise. With recent developments in deep learning, neural network based approaches have demonstrated impressive generative capabilities across a range of media, including image, text, video and audio. Specifically neural audio synthesis has achieved high-quality results in both speech and musical instrument modeling, while deep learning-based audio compression methods have enabled high-fidelity reconstruction from compressed representations.

Efforts toward deep learning based audio synthesis systems for audio texture synthesis, have been growing in recent years [1]. Despite these advancements, a significant challenge to adopting machine learning-based approaches for audio texture synthesis and sound effect generation lies in the lack of controllability and interpretability. Many deep learning models produce realistic results but offer limited control over timbre, structure, and dynamics, making them less practical for sound designers accustomed to traditional synthesis techniques that afford precise control.

In this paper, we evaluate and compare the audio texture synthesis capabilities and controllability of two popular paradigms in neural audio synthesis, using Variational Autoencoders (VAEs). We explore methods in DSP-informed neural audio synthesis and data-driven neural audio synthesis. We investigate DSP-based approaches presented in Differentiable Digital Signal Processing (DDSP) [2] and NoiseBandNet [3], alongside the data-driven latent representations presented in EnCodec [4].

We propose extensions to these works by integrating their approaches into VAE frameworks, while enabling fine grained temporal control over the generation process. We propose that VAEs, through their learned probabilistic latent representations, can enhance the expressiveness and flexibility of the chosen methods by enabling smooth traversal of the latent space during synthesis.

To achieve fine-grained temporal control, we constrain our models such that the latent space samples synthesize short sonic events, which we call, "atoms". This enables control over the synthesized audio for each atom of generated sound.

We evaluate and compare the synthesis capabilities of each of the methods presented using the objective metrics, multi-scale STFT ($mSTFT$) and Frechet Audio Distance (FAD). We investigate and compare the latent space structure of two chosen models, representing the two paradigms, through latent space interpolation.

In this paper we target three categories of audio textures inspired by environmental sounds that appear commonly in the sound and soundscape design process. These are Sea Waves, Rain and Thunder. Here we present three standardized datasets for each category, for use in machine learning models.

Through these contributions, we evaluate, compare and investigate the synthesis capabilities and latent structures of two major paradigms used in neural audio synthesis, DSP-informed and data driven approaches. We propose 4 models for neural audio synthesis of audio textures, in VAE frameworks, for evaluation and investigation.

2. RELATED WORK

A number of approach exist on audio synthesis using deep learning. Early efforts focused on data driven approaches that focused on generating waveforms directly using deep learning architectures.

An early break through in data driven approach to speech synthesis was presented in WaveNet, [5], where an auto-regressive convolutional neural network (CNN) is used to generate waveforms sample by sample. This achieved high quality speech synthesis but again at a high computational price. High quality speech and music synthesis was exhibited in SampleRNN [6], which uses

multiple recurrent neural network (RNN) layers at different time scales showed impressive synthesis capabilites, but at a high computational cost. Approaching the problem of efficiency WaveRNN [7] presents a compact and efficient RNN for real-time waveform generation, while still maintaining high quality synthesis. As well as audio synthesis, efforts have been made toward deep learning based audio codec, that aim at learning a compressed discrete representation of audio signals seen in the training data. EnCodec [4] is one such approach that has demonstrated high-fidelity audio reconstruction capabilities, from compressed latent representations. EnCodec uses a streaming encoder-decoder architecture with a quantized latent space. The encoder transforms input audio into a compact latent representation, which is then compressed using residual vector quantization (RVQ). The decoder reconstructs the original time-domain signal from the compressed representation. Training stability is ensured through multiscale spectrogram adversary and a novel loss balancer mechanism, which optimizes the trade-off between perceptual quality and reconstruction accuracy. On top of working as an audio codec, EnCodec is used as a foundational component in models used for audio generation, such as AudioGen [8], a transformer based model used for text-to-audio generation.

As well as purely data driven, and machine learning based approaches to audio synthesis, some have looked to prior knowledge on audio representations and digital signal processing techniques to aid learning capabilities and present more interpretable architectures.

GANSynth [9] showed early progress with this approach, it proposed a Generative Adversarial Network (GAN) that models log magnitude and instantaneous frequency spectrograms, outperforming WaveNet [5] in efficiency and quality. In [10] investigates a variety of audio representations, including raw waveforms and time-frequency representations, its results showed that complex-valued spectrograms and magnitude with instantaneous frequency spectrograms achieve the best results in terms of generation quality and efficiency.

DDSP [2] introduces a framework that combines deep learning with differentiable DSP components. Specifically DDSP proposes integrating an additive harmonic synthesizer, subtractive noise filtering synthesizer and reverberation module into a deep learning framework for synthesizing musical instruments. By doing so they create a more interpretable and controllable framework for audio synthesis. Many approaches have used the methods from DDSP within their own deep learning frameworks, such as RAVE [11].

Control is an important attribute of all audio synthesis systems, much research has been undertaken to expose different level of controllability of neural audio synthesis systems.

The work proposed in [12] introduces generative latent spaces that enable interpolation and extrapolation between the timbres of musical instrument sounds through latent space exploration. Building on this concept [13] uses a VAE architecture that disentangles salient musical features (descriptors) from the latent space and reintroduces them as conditioning variables during generation. Other approaches focus on enabling high level control, such as text, AudioGen [8] introduces a text-to-audio synthesis framework, it uses a transformer based architecture with EnCodec as it's audio synthesis component and has shown impressive results in synthesizing environmental sounds, sound effects and musical instruments.

As research has progressed in other areas of sound synthesis, such as speech and musical instruments, interest in neural audio synthesis of audio textures has been growing [1]. In [14] authors explore the use of GANs for audio texture synthesis, focusing on the impact of different audio representations. Authors propose training GANs on single-channel magnitude spectrograms, using the Phase Gradient Heap Integration (PGHI) inversion algorithm for phase reconstruction. As with other audio classes, controllability is an important factor in audio texture synthesis, and is often a key concern in proposed methods. MTCRNN [15] introduces recurrent neural network (RNN) based model a differing timescales, with a conditioning strategy that allows for user-directed synthesis. Again concepts from DDSP have been used in audio texture synthesis, DDSP-SFX [16] proposes transient modeling techniques synthesizing impulsive sounds like footsteps and gunshots, it also adds a learned weight to the DDSP harmonic synthesizer so it can more appropriately synthesize inharmonic sounds. Approaching the time-frequency trade-off exhibited in DDSPs noise filtering synthesizer NoiseBandNet [3] offers an alternative to the original noise filtering synthesizer. This approach is limited in that the model is trained on a limited dataset with a small number of specified controls. [17] takes a data-driven approach for audio texture morphing using a GAN conditioned on "soft-labels" derived from an audio classifier's penultimate layer that facilitates smooth interpolation between audio textures.

Other methods in audio texture synthesis have taken data-driven approaches to enabling control, offering control through transformations applied through example audio texture being provided, as seen in [18].

3. DATASET

In this work, we aim to investigate the capability of our proposed method to model three categories of audio textures. We focus specifically on environmental sounds, identifying them as our target group of audio textures. Within this group, we define three distinct sub-categories: Sea Waves, Rain, and Thunder.

Curated datasets of clean environmental sounds are limited in both availability and scale. Most publicly accessible datasets contain relatively few high-quality samples, which constrains deep learning research in audio texture and sound effect synthesis to data-scarce environments. In contrast, our goal is to construct larger, high-quality datasets to support the training of more robust models.

To train and evaluate our models, we construct three separate datasets, each corresponding to one of the defined sub-categories. Additionally, we create a fourth dataset by combining samples from all three sub-categories. All audio samples were sourced from Freesound[1], and each dataset was manually curated to ensure accurate sub-category representation and to minimize interference from secondary or extraneous sound sources.

Sea Waves, Rain and Thunder Each sub-category contains 2.5 hours of recordings of its respective sound source, yielding 9,000 one-second audio samples per category.

Combination This dataset merges the Sea Waves, Rain, and Thunder categories into a single collection, totaling approximately 7.5 hours of audio and 27,000 one-second samples.

All audio samples are monophonic, resampled to have the same sampling rate at 24kHz and stored with a 16-bit bit depth.

[1] www.freesound.org

4. PROPOSED METHODS

In this work, we aim to investigate the generative capabilities of two paradigms in neural audio synthesis, within VAE frameworks, these are;

1. Integrating traditional Digital Signal Processing (DSP) techniques into deep learning frameworks.
2. Developing purely data-driven deep learning architectures for audio synthesis.

We choose three state of the art implementations of the above approaches, two based on noise filtering techniques presented in DDSP [2] and NoiseBandNet [3], and one based on the data-driven approach in EnCodec [4]. We integrate these into VAE frameworks and propose that VAEs and their learned latent representations offer expressive and flexible synthesis capabilities.

4.1. Variational Autoencdoer

Before detailing our specific methods, we first provide an overview of Variational Autoencoders (VAEs), as they form the core architecture for our generative deep learning models.

VAEs are generative models designed to learn a probabilistic latent space, which can be used both for encoding data and generating new samples. They consist of:

- An encoder, which maps input data into a structured latent distribution.
- A decoder, which reconstructs data from sampled latent representations.

The VAE objective function includes two key components:

1. Reconstruction Loss (L_r) - Ensures that generated outputs resemble the original input during training.
2. Regularization Loss (L_{reg}) – Regularizes the latent space by minimizing the Kullback-Leibler (KL) divergence between the learned distribution and a predefined prior.

The overall objective function can be formulated as;

$$L = L_r + \beta * L_{reg} \tag{1}$$

Where β is used to control the strength of the regularization component.

The reconstruction loss differs based on the two paradigms used, but the regularization loss remains the same. The regularization loss, L_{reg} minimizes the KL divergence between the unknown conditional latent distribution, and a predefined prior distribution. By carefully choosing the prior, VAEs allow for smooth transitions and interpolations between latent space points. We choose the prior to be an isotropic gaussian of unit variance, $N(0, 1)$. We propose that such a choice should facilitate smooth interpolation through the latent space.

One of our objectives is to provide fine-grained temporal control over the latent space during synthesis, meaning that each decoded latent space vector will contribute a predefined number of samples to the final output audio. By training a VAE, we aim to learn a structured latent space that enables artifact-free interpolation, this is a desirable attribute when we wish to sample from the latent space sequentially whilst maintaining coherent and artifact free outputs.

Furthermore, a well-regularized latent space should allow for the synthesis of novel sounds that interpolate meaningfully between observed training data points.

4.2. Temporal Control

A common feature across all models we present is their ability to operate at a granular level. Each model learns a generative latent space, where individual latent vectors generate a fixed number of time-domain waveform samples, that we shall call 'atoms', meaning that the latent space operates at an atom based temporal scale. Audio texture can thus be generated by;

- Sampling trajectories through the latent space to generate complete audio textures.
- Appending newly generated atoms to extend audio textures in a continuous manner.

Depending on the implementation, generated atoms can be concatenated sequentially or overlap-added.

4.3. DSP-Informed Synthesis

The first paradigm we wish to explore is the integration of traditional DSP methods into deep learning frameworks. Given the noisy and inharmonic nature of the audio texture categories we aim to model we begin by exploring state-of-the-art noise filtering techniques. Specifically, we integrate the methods presented in DDSP [2] and NoiseBandNet [3] into VAE frameworks.

4.3.1. DDSP Noise Filtering

The original DDSP model [2] assumes a harmonic structure of the modeled audio by including an additive harmonic synthesizer. However, given the nature of the audio textures we are interested in, this assumption is unsuitable and can lead to artifacts in synthesized audio [16]. To account for this, we remove the harmonic synthesizer, and focus solely on the filtered noise synthesizer from the DDSP.

Here, we propose a VAE based architecture that generates the filter coefficient amplitudes required by the DDSP noise filtering synthesizer.

As input representation we use Mel-Frequency Cepstral Coefficients (MFCCs), computed from target audio. The spectral and temporal resolution of the MFCCs is determined by the STFT Frame Size and STFT Hop Size. MFCCs provide de-correlated spectral features and by taking the lower order coefficients, effectively describe the spectral shape in a perceptually informed manner.

With MFCCs as input, the VAE encoder is tasked with learning a probabilistic latent space and a compressed representation of spectral distribution.

Sampled latent vectors are decoded into the amplitudes of time-varying filter coefficients of a Finite Impulse Response (FIR) filter. White noise is convolved atom-by-atom with the learned FIR filter to synthesize audio.

Each point in the latent space decodes to an atom. Latent trajectories generate sequences of atoms, these are overlap-added to produce synthesized audio textures.

The number of samples present in an atom is equivalent to the STFT Frame Size. The hop size used in the overlap-add operation is determined by the STFT Hop Size.

The temporal and frequency resolution of the model is dictated by the choice of STFT analysis window length. The frequency resolution is also dependent on the number of Mel bands and MFCC coefficients - i.e. the parameters used to describe the spectral-shape of the STFT analysis. In MFCC computation, higher STFT

Frames Sizes improve frequency resolution but reduce time resolution, while smaller STFT Frame Sizes enhance time resolution but blur frequencies. This introduces a time-frequency trade-off to the DDSP Noise Filter;

- Higher filter resolution → Larger STFT frames → Lower temporal resolution.
- Higher temporal resolution → Smaller STFT frames → Lower filter resolution.

This tradeoff affects both synthesis control and reconstruction fidelity, requiring careful parameter choice during MFCC computation.

4.3.2. Noise Band Synthesis

The Noise Band synthesis method presented in NoiseBandNet [3] is a noise filtering approach that aims to address the time-frequency trade of issue exhibited in the DDSP Noise Filtering technique. We integrate the Noise Band Synthesizer into a VAE framework, removing the time-frequency trade off and enhancing its generative capabilities.

MFCCs are again the chosen input representation and are encoded to a probabilistic latent space. The sampled latent vectors are decoded to generate amplitudes for a learned filterbank, used to filter white noise. The number of bands present in the filterbank are no longer coupled to the MFCC decomposition and thus does not assume a predefined filter shape. This differs from the DDSP Noise Filter whose filter shape is defined by MFCC decomposition, offering increased flexibility with high-resolution per band control, as well as high resolution temporal control. With this method the number of samples present in an atom is now set to the STFT Hop Size.

Sequences of atoms are concatenated to form fixed length audio textures, or single generated atoms can be concatenated to a signal to form continuous audio textures, as they are produced.

4.3.3. VAE-Based Architecture

We train a VAE-based architecture that consists of the following;

- Preprocessing: Extracts Mel-Frequency Cepstral Coefficients (MFCCs) from input signal.
- Encoder: The encoder consists of a GRU-based recurrent network followed by linear layers that generate distribution parameters, mean and log-variance.
- Latent Space Sampling: The distribution parameters are use to generate the latent vector, z, computed as;

$$z = \mu + \epsilon \cdot \sigma^2$$

  where, $\epsilon \sim N(0, 1)$, μ is the mean, and $\sigma = e^{0.5 * logvar}$.
- Decoder: Sampled latent vectors act as direct input to the decoder. The decoder consists of a multi-layer perception (MLP) followed by a GRU layer, another MLP layer and a final linear layer.

4.3.4. Reconstruction Loss

For the reconstruction loss, L_r we use a similar multi-scale STFT loss ($mSTFT$) as used in [3][2][16]. The $mSTFT$ loss measures the difference between the magnitude spectrograms of generated and target audio across multiple time-frequency resolutions, capturing both fine details and long-term spectral structure.

We extend the $mSTFT$ loss function to include both the magnitude spectrograms and mel spectrograms, in order to align closer with humans perception of sound. We use STFT window lengths of [2048, 1024, 512, 256, 128] with a hop size, $window\_length/4$.

4.4. Data Driven Synthesis

The second paradigm we wish to investigate is data driven approaches to neural audio synthesis. Due to its high-fidelity audio synthesis capabilities, observed in [8], we adapt the methods presented in EnCodec [4] into a VAE framework.

4.4.1. EnCodec

In our proposed method we modify the original EnCodec architecture by removing the RVQ component and thus the discrete quantized latent space. We train a Variational Autoencoder variant, that aims to learn a smooth, continuous latent space.

Unlike the DSP-based VAE models, which encode extracted spectral features to the latent space, the VAE-EnCodec architecture encodes time-domain waveform segments, and thus relies much more on the encoder to find meaningful and compressed representations from the input waveform.

Each latent vector is decoded into a predefined fixed number of waveform samples, sequences of decoded audio samples are concatenated into one second segments. Larger outputs are generated by overlap-adding segments.

4.4.2. VAE-Based Architecture

In our proposed method we modify the original EnCodec architecture by removing the RVQ component and training a Variational Autoencoder variant.

The architecture consists of;

- Encoder: The encoder consists of a 1D convolutional layer, followed by multiple convolutional blocks. Each convolutional block comprises of a single residual unit with two convolutional layers and a skip connection, followed by a strided convolution for down-sampling. The convolutional block is followed by a two layer LSTM for temporal modeling and a 1-D convolutional layer. A final linear layer generates distribution parameters, mean and log-variance.
- Latent Space Sampling: As with the previous VAE architecture, the mean and log-variance values are used to generate the sampled latent vector, z.
- Decoder: The decoder mirrors the encoder, using transposed convolutions to progressively up-sample the latent representation back to time-domain waveforms. Strided values are reversed to ensure a symmetric reconstruction process.

The internal sampling rate, determined by the down-sampling and up-sampling of the encoder and decoder, means a single latent vector generates a single atom of audio. Given that we constrain the number of sample to be relatively small we provide fine-grained temporal control over the latent space and generated audio.

4.4.3. Reconstruction Loss

The reconstruction loss is composed of a combination of loss terms used to achieve the same high-fidelity audio reconstruction observed in the EnCodec.

First we use both time and frequency domain loss terms. The time domain term minimizes the L1 distance between the target and reconstructed audio over the time domain;

$$L_t(x, \hat{x}) = ||x - \hat{x}||_1 \tag{2}$$

where x is the input signal, and $\hat{x}$ the reconstructed signal. For the frequency domain we again use the multi-scale STFT loss, ($mSTFT$), as previously presented.

In addition to the reconstructions and regularization terms, we include a discriminative loss term, L_{dis}, as in [4]. The discriminative loss term is part of an adversarial training framework and aims to improve the audio quality of generated samples. It uses a multi-scale STFT-based discriminator, that consists of multiple discriminators operating at varying time resolutions. The multi-scale STFT discriminator consists of identically structured networks operating on multi-scaled complex-valued STFTs with the real and imaginary parts concatenated, five scales with STFT window lengths of [2048, 1024, 512, 256, 128] are used. Thus an adversarial loss, L_g for our objective function can be constructed as follows;

$$L_g(\hat{x}) = \frac{1}{K} \sum_k \max(0, 1 - D_k(\hat{x})) \tag{3}$$

where, K is the number of discriminators.

As in [4] we include a relative feature matching loss, L_{ft} in our overall objective function, whose formulation can be found in [4].

Combining these terms our final objective function looks like;

$$L_r = \lambda_t \cdot L_t(x, \hat{x}) + \lambda_f \cdot L_f(x, \hat{x}) + \lambda_g \cdot L_g(\hat{x}) + \lambda_{ft} \cdot L_{ft}(x, \hat{x}) \tag{4}$$

where, $\lambda_t, \lambda_f, \lambda_g$ and λ_{ft} are scalar coefficients used to balance between the terms.

To stabilize training the loss balancer as described in [4] is used.

5. EXPERIMENTS

5.1. Training

We propose a total of four models for evaluation and comparison, these are;

- VAE-DDSP-NF: DSP-informed method with VAE architecture and noise filtering synthesizer, introduced in [2].
- VAE-NoiseBand: DSP-informed method with VAE architecture and noise band synthesizer, introduced in [3].
- VAE-EnCodec$_{16}$ Data driven method with VAE implementation of EnCodec [4], with latent dimensionality of 16.
- VAE-EnCodec$_{128}$ Data driven method with VAE implementation of EnCodec [4], with latent dimensionality of 128.

We train all four model on all 4 datasets, resulting in 16 trained models in total. The data is split into 90% for training and 10% for testing, as is in [16]. Each model is trained for 200,000 training steps on a Nvidia RTX A5500 GPU.

5.1.1. VAE-Noise Filtering Synthesizers

For the noise synthesis based models we use the Adam optimizer, a batch size of 16 and a learning rate of 0.001. For the first 10% of the total epochs only the reconstruction component of the objective function is optimized. This regularization strength is then linearly increased to its target value, $\beta = 0.3$, over a specified number of steps, in our case 50.

The VAE uses a hidden size of 128 for GRU and MLP layers and the latent space has 16 dimensions.

For the initial calculations of the MFCCs a STFT Frame Size of 512 samples is used with a STFT Hop Size of 128 samples. The number of mel bands used is 128, and the number of MFCCs used are 30.

This results in an atom size of 512 samples for DDSP Noise Filter Synthesizer, with a hop size of 128 in the overlap-add method, and an atom size of 128 for the Noise Band Synthesizer.

5.1.2. VAE-Encodec

We train all VAE-EnCodec models with the Adam optimizer, a batch size of 16 and a learning rate of $3e^{-4}$. We use the objective function introduced previously with $\lambda_t = 0.1, \lambda_f = 1, \lambda_g = 3$, $\lambda_{ft} = 3$, as in [4].

We train two models with different latent space dimensions, one with 16 dimensions, which offers a comparable rate of compression and latent dimensionality to DSP-informed methods, and one model with 128 latent space dimensions as was use in the original Autoencoder presented in [4].

An increased regularization weight of $\beta = 30.0$ is used, as experimentation showed this was required to achieve adequate regularization in training. We use the same incremental scheme as previously to introduce the regularization component.

5.2. Evaluation

We wish to investigate our models capabilities to provide high quality audio synthesis, and a generative latent space that offers smooth and fine grained temporal control.

To do so, we identify two evaluation strategies. The first is reconstruction evaluation. This should measure the model's ability to reconstruct signals that have been encoded to latent space.

The second is latent space navigation. This should investigate the general structure and traversal of the latent space.

5.2.1. Reconstruction

To evaluate our models reconstruction and synthesis capabilities we use two objective metrics. Firstly, we use the multi-scale STFT [2], to determine spectral similarity between real audio textures and synthesized audio textures.

Secondly we use the Fréchet Audio Distance (FAD), [19]. 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 a VGGish to directly compare the reconstructed audio produced by our models with the original target audio, [2].

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

	Sea Waves		*Rain*		*Thunder*		*Average*	
	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**
VAE-DDSP-NF	0.108±0.025	10.378	0.073±0.021	13.274	0.051±0.017	13.180	0.077	12.277
VAE-NoiseBand	0.083±0.034	3.842	0.064±0.020	2.880	0.035±0.012	2.879	0.061	3.200
VAE-EnCodec$_{16}$	0.052±0.013	2.045	0.058±0.024	1.299	0.019±0.017	3.996	0.043	2.447
VAE-Encodec$_{128}$	**0.025±0.005**	**0.430**	**0.029±0.011**	**0.687**	**0.004±0.003**	**1.796**	**0.019**	**0.971**

Table 1: $mSTFT$ loss (mean ± sd) and FAD results for reconstruction evaluation for models trained on individual sound categories. Comparing VAE implementation of DDSP Noise filtering, noisebandnet and EnCodec

	Sea Waves		*Rain*		*Thunder*		*Average*	
	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**	***mSTFT*↓**	***FAD*↓**
VAE-DDSP-NF	0.115±0.049	10.193	0.075±0.024	11.466	0.049±0.012	10.742	0.080	10.800
VAE-NoiseBand	0.086±0.023	5.028	0.057±0.012	3.583	0.037±0.007	2.945	0.060	3.852
VAE-EnCodec$_{16}$	0.058±0.014	2.115	0.048±0.019	1.875	0.007±0.004	2.652	0.038	2.213
VAE-Encodec$_{128}$	**0.024±0.004**	**0.430**	**0.031±0.016**	**0.501**	**0.003±0.002**	**0.479**	**0.019**	**0.470**

Table 2: *$mSTFT$ loss (mean ± sd) and FAD results for reconstruction evaluation for models trained on data from all sound categories. Comparing VAE implementation of DDSP Noise Filtering, NoiseBandNet and EnCodec*

For evaluating reconstruction capabilities we train one model on each dataset. We train two variants of VAE-EnCodec, with latent dimensions, z_{dim}, of 16 and 128. We do this as to have a VAE-EnCodec model with comparable latent space dimensionality to the noise filtering counterparts, ie $z_{dim} = 16$. We also train a model with $z_{dim} = 128$ as outlined in original EnCodec architecture to understand what affect compressing the latent space will have.

5.2.2. Latent Space Navigation

To investigate the structure and navigability of the learned latent representations, we perform interpolation between two encoded atoms in the latent space. We expects a well-structured latent space is to produce smooth transitions in the synthesized audio along the interpolation path.

To do this we propose taking two audio atoms from different audio texture recordings and encoding them to latent vectors, z_1 and z_2. We then generate a trajectory through the latent space by interpolating between the two encoded latent vectors, using linear interpolation, formulated by;

$$z_t = (1 - t)z_1 + tz_2 \tag{5}$$

where, z_1 and z_2 are the two latent vectors and t is the interpolation factor varying from 0 to 1.

We investigate the smoothness of interpolation by computing the STFT and the spectral centroid (brightness) over the generated samples. We propose that smooth interpolation between the two atoms can be observed in a smooth transition in both the STFT and spectral centroid.

The latent vectors, z_1 and z_2, are selected from two distinct audio recordings within the same environmental sound sub-category. Each pair is chosen to reflect a transition between contrasting spectral characteristics within that sub-category. We choose the following;

- Sea Waves: Crashing wave → calm shoreline
- Rain: Heavy Rain → light rain
- Thunder: Deep rumbling → quiet atmospheric texture

6. RESULTS

6.1. Reconstruction

$mSTFT$ and FAD results are reported in Table 1. This table shows results for models trained on the Sea Waves dataset, Rain dataset and Thunder Dataset individually, resulting in 12 models in total. Each model is evaluated on the audio texture category it was trained on. In regards to noise filtering synthesizers we see that VAE-NoiseBandNet out performs VAE-DDSP-NF (DDSP Noise Filtering). We observe that when it comes to reconstruction fidelity, VAE-EnCodec$_{128}$ shows considerably better results than all other models.

Table 2 reports $mSTFT$ and FAD metrics for models trained on the Combined dataset, which contains all three audio texture categories. This results in a total of four trained models for evaluation. Each model is evaluated on all three audio texture categories. Here we observe similar trends as to the previously discussed results. We also observe similar results to the models trained on each dataset individually.

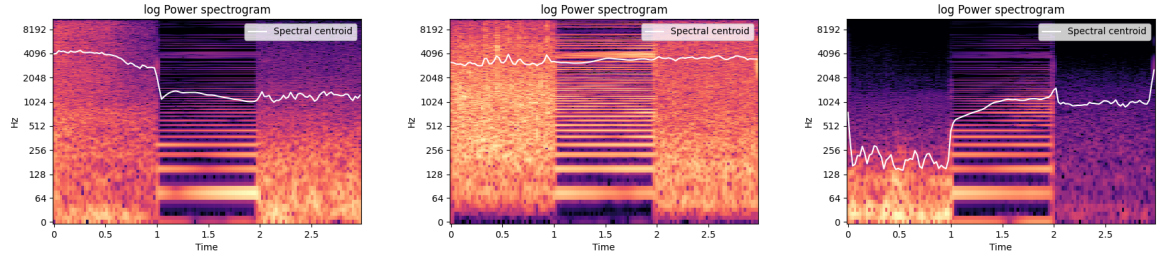
Samples can be found and listened to on our supplementary website [3].

6.2. Latent Space Navigation

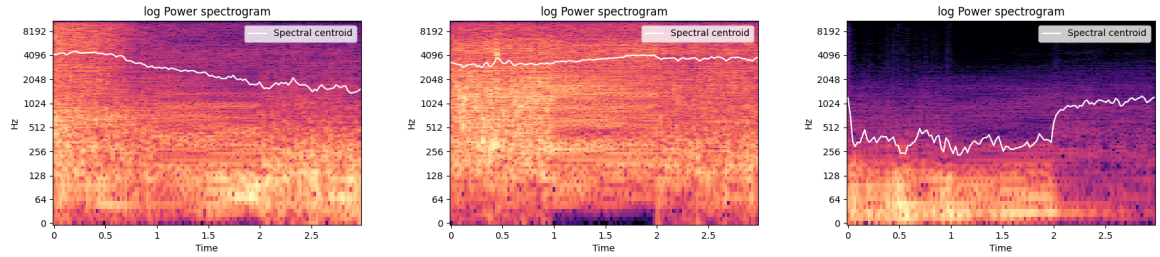
For latent space navigation we compare two models, one for each paradigm, we choose VAE-EnCodec$_{16}$ and VAE-NoiseBand due to their comparable latent space sizes. Figure 1 presents the log Power spectrograms and spectral centroids of three synthesized audio textures, generated by our models.

Each spectrogram captures a transition between two real audio segments, with the middle 1-second segment synthesized via linear interpolation in the latent space. The results reveal a distinct difference in interpolation behavior. VAE-EnCodec$_{16}$ introduces structured frequency bands during interpolation, producing an unintended tonal quality. In contrast, VAE-NoiseBand maintains a noisier, more natural transition, truer to the training data.

[3] https://aaron-dees.github.io/generativeLatentSpaces/



(a) STFT and Spectral Centroid for synthesized audio generated by VAE-EnCodec$_{16}$



(b) STFT and Spectral Centroid for synthesized audio generated by VAE-NoiseBand

Figure 1: *Comparison of the STFTs and Spectral Centroids of synthesized audio generated from interpolated latent trajectories of Sea Waves, Rain and Thunder.*

Examining the spectral centroids, both models exhibit a smooth directional trend through the interpolation segment. However, VAE-EnCodec$_{16}$ shows minimal variation, while VAE-NoiseBand more accurately reflects the fluctuations in spectral centroid between consecutive atoms of the original textures. We also note that VAE-EnCodec$_{16}$ exhibits relatively large jumps in Spectral Centroid as the interpolation segment begins.

Subjective listening confirms these findings, VAE-NoiseBand produces a smoother and more perceptually natural transition between textures compared to VAE-EnCodec$_{16}$. Samples can be found and listened to on our supplementary website.

7. DISCUSSION

In this work, we proposed three VAE based frameworks for two common paradigms in neural audio synthesis. Our results demonstrate that, for DSP-informed synthesis, integrating the Noise Band synthesizer within a VAE framework, inspired by [3], produces superior reconstruction performance compared to the the original DDSP Noise Filter synthesizer, according to the metrics used, this aligns with similar results seen in [3]. The data-driven approach, VAE-EnCodec shows superior reconstruction performance for both its models, compared to the DSP-inspired approaches. We observe that reconstruction quality in these models depends on compression rate. When the latent dimensionality is 128, VAE-EnCodec achieves high-quality reconstructions. However, as the compression rate increases and the latent dimensionality is reduced to 16, comparable to that of VAE-NoiseBand, the reconstruction quality degrades and scores marginally higher than VAE-NoiseBand.

Beyond reconstruction, we also investigated the generative properties of the learned latent spaces. Our results reveal a distinct difference between the two approaches: VAE-NoiseBand exhibits smoother and more perceptually accurate interpolations between known textures. We argue that this advantage stems from two key factors, firstly using MFCCs as input features encourages a more structured latent representation in terms of spectral distribution, and secondly, leveraging a DSP-informed noise synthesizer implicitly provides a more accurate spectral modeling of the target audio textures.

7.1. Limitations and future work

While our proposed VAE-based frameworks demonstrate promising results in both DSP-informed and data-driven neural audio synthesis, several limitations remain.

Firstly, the VAE-NoiseBand model showed smooth interpolation capabilities between points in the latent space, with this said, due to the fine grained temporal structure of our framework the interpolated trajectories synthesized audio produces a smooth and consistent spectral transitions between atoms, but lacks the temporal variations and coherence that are observed in natural audio textures. Future work could include a component to model these temporal variations, to provide a framework for modeling temporal variation and long term coherence as well as spectral changes in the synthesized audio as we navigate the latent space.

The reconstruction capabilities of the data driven approach, VAE-EnCodec, are hard to ignore, as they out-performed their DSP-inspired counterparts. Future work could explore alternative model architectures for more efficient latent space encoding and better regularization, to improve performance in low-dimensional settings..

Given the nature of our data-driven approach, it is difficult to comment on the behavior of the VAE-EncCodec interpolation experiments and structure of the latent space. Further consideration and investigation would be required to do so.

While our evaluation of latent space interpolation provides insight into the generative properties of both approaches, a more systematic perceptual study would further validate the observed differences. Future work could incorporate listening tests to better assess the subjective quality of generated sounds.

Our proposed method exposes only a generative latent space for synthesizing audio. Although we showed that VAE-NoiseBand offered a structured latent representation, we could expose additional controls by disentangling extracted audio features from the latent space and conditioning the decoder with them, as is done in [13].

As well as traditional controls, such as extracted audio features, we also note the useful nature of higher level, user defined controls, such as rain intensity of wave frequency. We aim to explore the possibilities of such a framework in future work through introducing hierarchical architectures that enable high level control with coherent long term temporal evolution. Ultimately we aim to introduce a neural audio synthesis framework that will enable both high level and low level control over audio texture synthesis, whilst providing high-fidelity synthesis capabilities with realistic temporal coherence.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] Keunwoo Choi, Sangshin Oh, Minsung Kang, and Brian McFee, "A proposal for foley sound synthesis challenge," 2022.

[2] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," *ArXiv*, vol. abs/2001.04643, 2020.

[3] Adrián Barahona-Ríos and Tom Collins, "Noisebandnet: Controllable time-varying neural synthesis of sound effects using filterbanks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 1573–1585, Feb. 2024.

[4] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," 2023, Featured Certification, Reproducibility Certification.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[6] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, José Sotelo, Aaron Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," 12 2016.

[7] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2410–2419, PMLR.

[8] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "Audiogen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*, 2023.

[9] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts, "Gansynth: Adversarial neural audio synthesis," in *International Conference on Learning Representations*, 2019.

[10] Javier Nistal, Stefan Lattner, and Gaël Richard, "Comparing representations for audio synthesis using generative adversarial networks," *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 161–165, 2020.

[11] Antoine Caillon and Philippe Esling, "Rave: A variational autoencoder for fast and high-quality neural audio synthesis," 2021.

[12] Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton, "Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics," in *Proc. Digital Audio Effects (DAFx-18)*, 2018.

[13] Ninon Devis, Nils Demerlé, Sarah Nabi, David Genova, and Philippe Esling, "Continuous descriptor-based control for deep audio synthesis," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] Chitralekha Gupta, Purnima Kamath, and Lonce Wyse, "Signal representations for synthesizing audio textures with generative adversarial networks," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2020, pp. 158–163.

[15] M. Huzaifah and L. Wyse, "Mtcrnn: A multi-scale rnn for directed audio texture synthesis," 2020.

[16] Yunyi Liu, Craig Jin, and David Gunawan, "Ddsp-sfx: Acoustically-guided sound effects generation with differentiable digital signal processing," in *Proc. Digital Audio Effects (DAFx-24)*, 2024, pp. 216–221.

[17] Chitralekha Gupta, Purnima Kamath, Yize Wei, Zhuoyao Li, Suranga Nanayakkara, and Lonce Wyse, "Towards controllable audio texture morphing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] Purnima Kamath, Chitralekha Gupta, Lonce Wyse, and Suranga Nanayakkara, "Example-based framework for perceptually guided audio texture generation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 2555–2565, Apr. 2024.

[19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," 09 2019, pp. 2350–2354.

[20] Seán O'Leary and Axel Röbel, "A montage approach to sound texture synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 6, pp. 1094–1105, 2016.

[21] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," 2019.