Variational Autoencoders for Musical Spectrogram Modeling: MUSE - VAE

Tony Farrand
Duke University
STA 571: Advanced Probabilistic Machine Learning

Abstract

This paper investigates the use of Variational Autoencoders (VAEs) to learn probabilistic latent representations of musical audio using spectrogram segments extracted from the GTZAN dataset. Each 30-second audio clip is divided into six 5-second windows and transformed into mel spectrograms. The probabilistic model is defined through the joint distribution $p_\theta(x, z) = p_\theta(x \mid z)p(z)$ with a Gaussian prior and encoder-defined approximate posterior. Particular emphasis is placed on latent-space structure to analyze encoder behavior. Results show meaningful separation of genres in latent space and recognizable but regularized spectrogram reconstructions, consistent with known VAE trade-offs between fidelity and structure.

1 Introduction

Variational Autoencoders (VAEs), introduced by Kingma and Welling [1], are latent-variable generative models that combine neural networks with variational Bayesian inference. VAEs learn a continuous, smooth latent space by optimizing a tractable lower bound on the data likelihood. This makes them useful for tasks such as interpolation, clustering, and generative audio modeling.

In this project, I apply a convolutional VAE to musical audio represented as mel spectrograms derived from the GTZAN dataset [3]. The main focus is on reconstruction quality and on how genre information and musical structure appear in the learned latent space.

2 Dataset and Preprocessing

I use the GTZAN genre dataset, a benchmark collection of ten genres, each containing 100 audio clips. Each 30-second audio file is segmented into six fixed-length windows and transformed into mel spectrograms using a 128-bin mel filterbank.

Because nearly all genres in GTZAN use similar instrumentation such as voice, guitars, bass, and drums, the resulting mel spectrograms share similar global structures. This is important context when interpreting the latent space and the degree of genre separation.

3 Previous Work

Deep generative models for music have advanced steadily over the past decade. MusicVAE [2] showed that hierarchical latent structures can support smooth interpolation and long-range musical coherence. While MusicVAE operates mainly on symbolic sequences, my work focuses on spectral audio representations. This introduces separate challenges such as phase reconstruction and preservation of fine temporal detail, which must be addressed differently in the audio domain.

4 Methods

4.1 Probabilistic VAE Model

A VAE models data x using a latent variable z sampled from a prior $p(z)$, usually a standard Gaussian. The joint model is

$$p_\theta(x, z) = p_\theta(x \mid z)p(z).$$

Because the true posterior $p_\theta(z \mid x)$ is intractable, I use a neural encoder to learn an approximate posterior

$$q_\phi(z \mid x) = \mathcal{N}(z \mid \mu_\phi(x), \sigma^2_\phi(x)I).$$

Training optimizes the Evidence Lower Bound (ELBO)

$$\mathcal{L}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}(q_\phi(z \mid x) \,\|\, p(z)).$$

The first term encourages accurate spectrogram reconstruction, and the KL term regularizes the latent space toward a smooth Gaussian manifold. I use KL annealing to prevent early posterior collapse.

4.2 Model Architecture

I use a convolutional encoder and decoder tailored to 128×216 mel spectrograms. A moderately high latent dimensionality helps avoid oversmoothing and supports variation across genres. A Softplus output activation ensures valid, positive mel coefficients.

4.3 Latent Space Analysis Tools

To interpret the learned representation, I apply PCA projections to the latent encodings. These projections give a human-interpretable view of how genres cluster and how the model arranges different musical styles in the latent space.

4.4 Mode Collapse Mitigation

Audio contains high-frequency and transient detail that is difficult to model. To reduce collapse to a narrow set of spectra, I use:

- KL annealing across 50 epochs,
- a latent dimension in the range of 512 to 2048,
- a log-magnitude reconstruction loss,
- architectural choices that limit excessive smoothing.

5 Results

5.1 Latent Space Structure

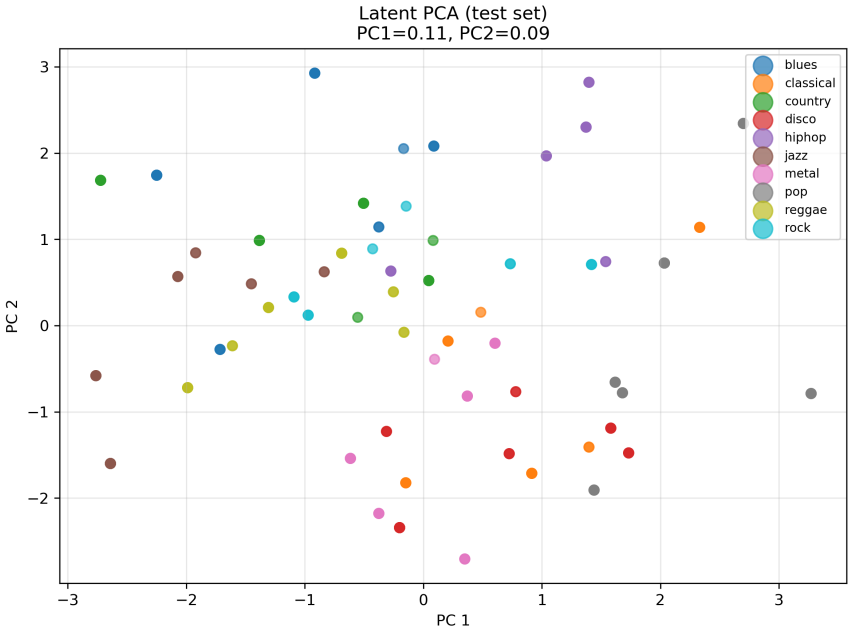


Figure 1: PCA projection of the VAE latent space showing emergent genre clusters.

Several musically meaningful clusters emerge from the latent representation:

- Rock and Country lie adjacent to Blues, reflecting shared musical origins and similar instrumentation.
- Blues and Jazz appear near one another, consistent with their overlapping harmony and rhythm.
- Hip-Hop and Pop occupy distinct regions, both dominated by strong and unique rhythmic structure.

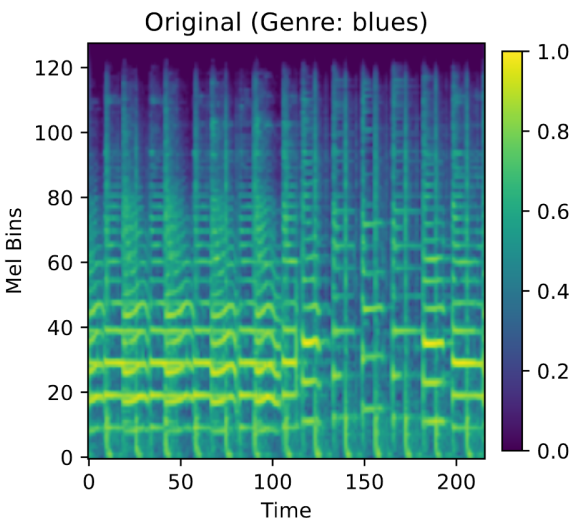
An especially interesting observation is that Classical and Metal lie close to each other in the latent projection. Many musicians informally note that metal often resembles classical music played with distortion, and both styles frequently use fast scalar passages, harmonic minor modes, dramatic dynamic contrast, and dense layering. The fact that the VAE recovers this relationship from short mel spectrogram segments suggests that it captures genuine similarities in time–frequency structure rather than just surface timbre.

A broader theme is that GTZAN groups genres that share core instruments such as guitars, human voice, and a drum kit. This naturally limits separability compared to datasets with larger timbral diversity, such as solo instruments or ensemble-specific collections.

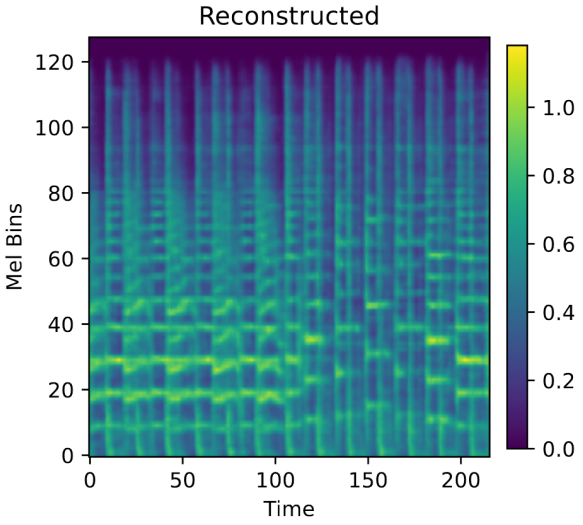
5.2 Reconstruction Quality

The VAE reconstructs mel spectrograms with reasonable fidelity. As expected, reconstructions exhibit some blurring due to the Gaussian prior and the KL term. KL annealing and a higher latent dimension were important in avoiding posterior collapse and preserving recognizability in both time and frequency.

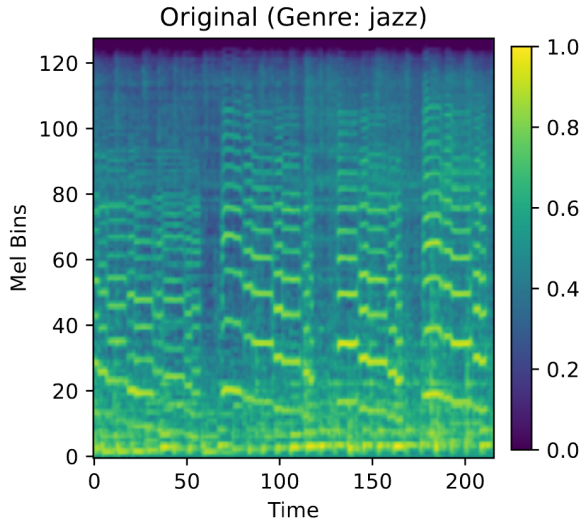
The model reproduces broad spectral envelopes, energy contours, and band structures that align well with the originals, although fine texture is softened.



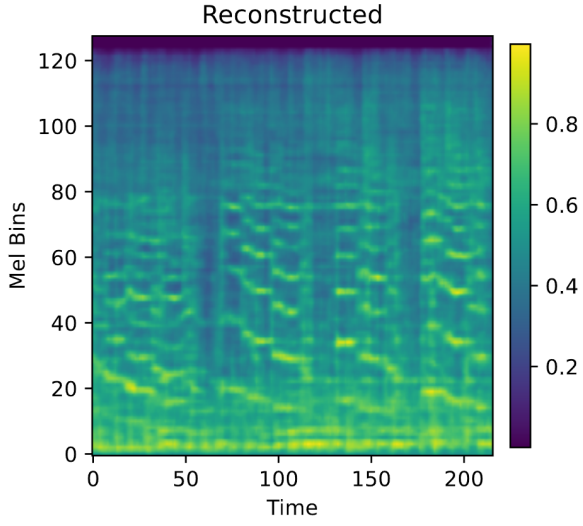
(a) Actual Spectrogram 1



(b) Reconstruction 1



(c) Actual Spectrogram 2



(d) Reconstruction 2

Figure 2: Original mel spectrograms (left) and VAE reconstructions (right).

6 Future Work

A major limitation of this work is the reliance on the Griffin–Lim vocoder to convert mel spectrograms back to audio. Griffin–Lim does not model phase and therefore introduces audible distortion. I confirmed this with a small test in which original audio clips were converted to mel spectrograms and then inverted, yielding noticeably degraded audio even without the VAE.

Modern audio systems typically use neural vocoders such as HiFi-GAN, WaveGlow, WaveRNN, or diffusion-based decoders. Integrating one of these models would likely improve perceptual quality, but training and validating such a system was beyond the scope of this project.

7 Conclusions

This project explores how a variational autoencoder behaves when trained on short mel spectrogram segments drawn from the GTZAN music genre dataset. Even though the model only sees two-second windows, the learned latent space reflects patterns that align with basic music theory and genre structure, such as the close relationship between Blues, Rock, and Country and the structural similarity between Classical and Metal.

It is notable that a relatively simple convolutional VAE, combined with standard regularization and KL annealing, can capture these relationships while still reconstructing spectrograms with enough detail to recover recognizable audio when paired with a vocoder. Overall, the project provides a practical demonstration of how VAEs trade off sharpness, diversity, and latent organization when modeling real audio data.

8 References

[1] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[2] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Doug Eck. Hierarchical variational autoencoders for music. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[3] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, volume 10, pages 293–302, 2002.