VIP Week 2

Charbel El Haddad

September 2025

1 Introduction

During week 2 of VIP, I experimented with generative models for music, focusing specifically on *oud* music. The goal was to explore probabilistic models and understand the complexity of musical data.

2 Gaussian Model for Music Generation

I first experimented with a Gaussian model of the form:

$$x = \hat{\mu} + EA^{1/2}z, \tag{1}$$

where z is a latent variable, and the covariance matrix is defined as:

$$\Sigma = EAE^T. \tag{2}$$

This corresponds to sampling from a Gaussian distribution:

$$x \sim \mathcal{N}(\hat{\mu}, \Sigma). \tag{3}$$

The results showed that music is more complex than what a simple one-layer Gaussian distribution can capture, with $\hat{\mu}$ representing the average over the data.

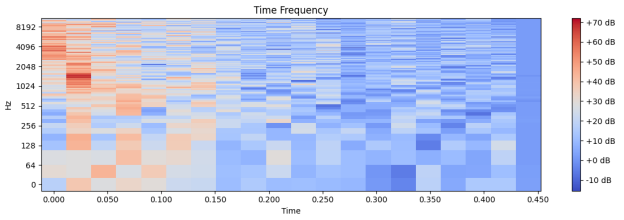

Figure 1: Results of the Gaussian music generation model.

3 Diffusion Models for Music

Due to the limitations of the Gaussian model, I began studying diffusion models, mainly the *DDPM* paper [1], to better understand how these models function.

A key observation was that most image diffusion models generate 256×256 images and are trained on that resolution. Music, however, is more complex and requires awareness of the entire piece.

4 Proposed Model Architecture

I proposed a context-aware architecture for music generation:

- Input: raw audio $(128, T)$
- Autoencoder compresses to latent space
- Diffusion model operates on the compressed mel-spectrogram latent space
- Decoder reconstructs the audio
- Extra neuron in input and output layers represents the current window index
- Sliding window design: each new window preserves $\alpha\%$ of the previous window to maintain continuity

This design ensures the model is both context-aware and capable of generating music token by token.

5 References

References

[1] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models (DDPM)*. Advances in Neural Information Processing Systems, 2020.