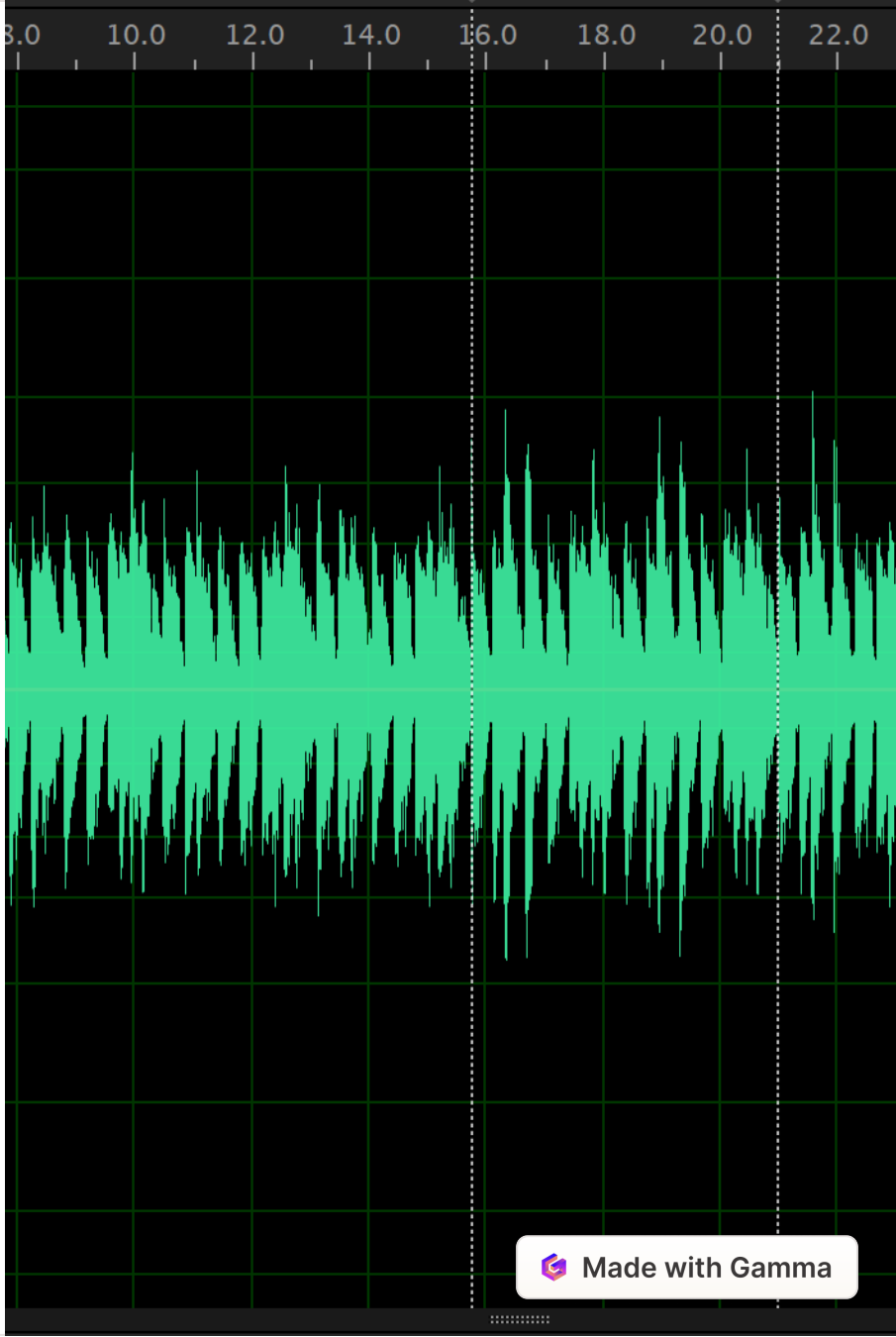
Adversarial Audio Synthesis

In this project, we explored the exciting field of audio synthesis using Generative Adversarial Networks, which can create realistic audio from latent vectors. We compared two approaches: PianoGAN and SpectoGAN, which operate on different representations of audio: notes and spectrograms.

- Aditya Bhangale (210070004)
- Vinay Sutar (21d070078)
- Vikas Kumar (210070093)



How GANs Work

1

The Generator

Produces fake data resembling the real data distribution

2

The Discriminator

Distinguishes between real and fake data and provides feedback to the generator

3

The Minimax Game

The generator and discriminator compete to optimize their performance

4

Loss Function for GANs:

$$V(D,G) = \mathbb{E}_{x\sim P_X}[\log D(x)] + \mathbb{E}_{z\sim P_Z}[\log(1 - D(G(z)))].$$

Loss Functions for GANs

1 Jensen-Shannon Divergence

Measures the similarity between two probability distributions

2 Wasserstein Distance (Used)

Measures the distance between two probability distributions, is smoother in nature compared to Jensen-Shannon Divergence

$$W(P_X, P_G) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P_X}[f(x)] - \mathbb{E}_{x \sim P_G}[f(x)],$$

3 Other Loss Functions

Can be used for different objectives, such as image-to-image translation or style transfer

2D and 1D convolutions

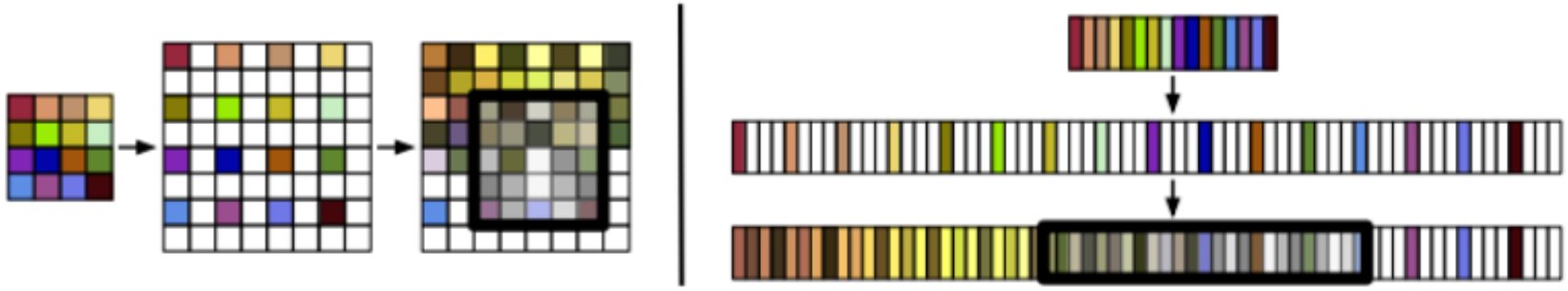


Fig. 2. Depiction of the transposed convolution operation for the first layers of the SpectoGAN (left) and PianoGAN (right) generators. DCGAN uses small (5x5), two-dimensional filters while PianoGAN uses, one-dimensional filters and a larger upsampling factor. Both strategies have the same number of parameters and numerical operations.

PianoGAN for Note Generation

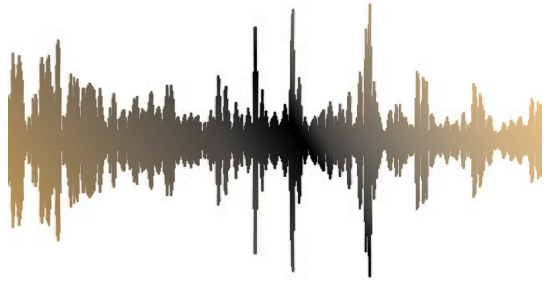
Architecture-Based on DCGAN

MAESTRO Dataset

Contains MIDI files of piano performances

Pitch, Step, and Duration

Variables extracted from each note

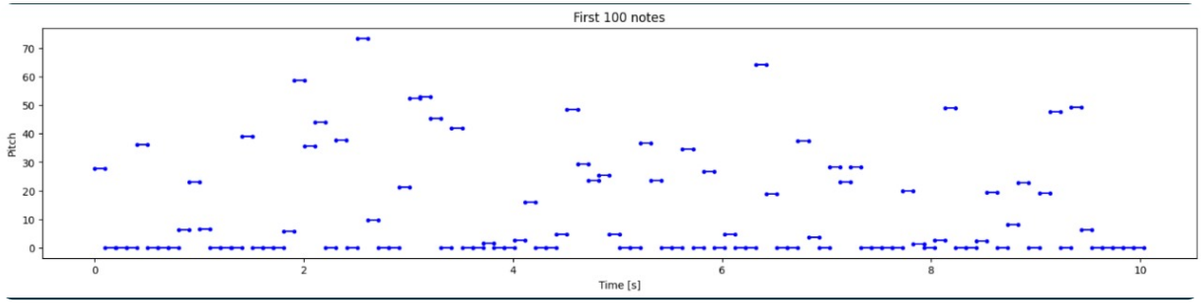


1D Convolutions

Process raw audio waveforms to generate notes

PianoGAN Output

Output:



256 Notes

Generated for each latent vector

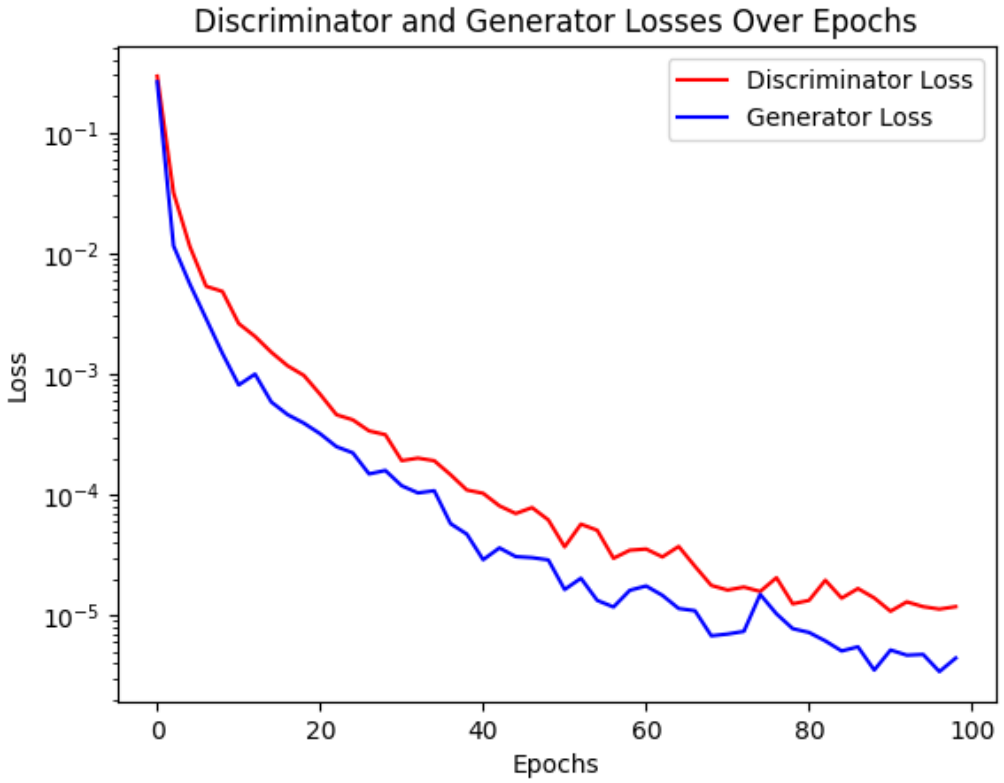
Pitch, Step, and Duration

Obtaining a variable pitch keeping step and duration constant.

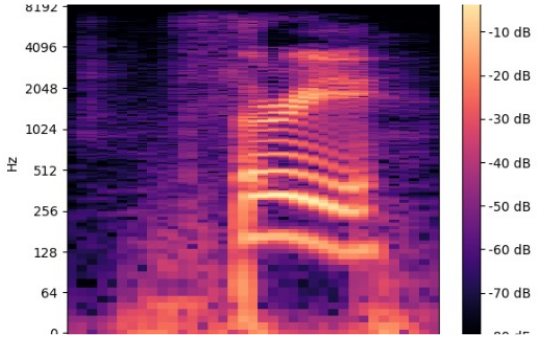
Realistic Piano Compositions

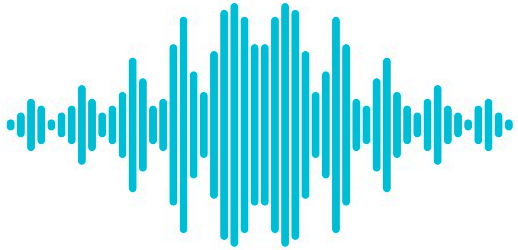
Can be converted to audio using pretty-midi library

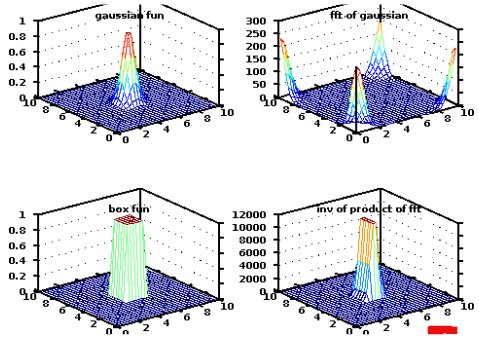
Training of PianoGAN:



SpectoGAN for Spectrogram Generation







SC09 Dataset

Contains spoken audio clips of digits from 0 to 9

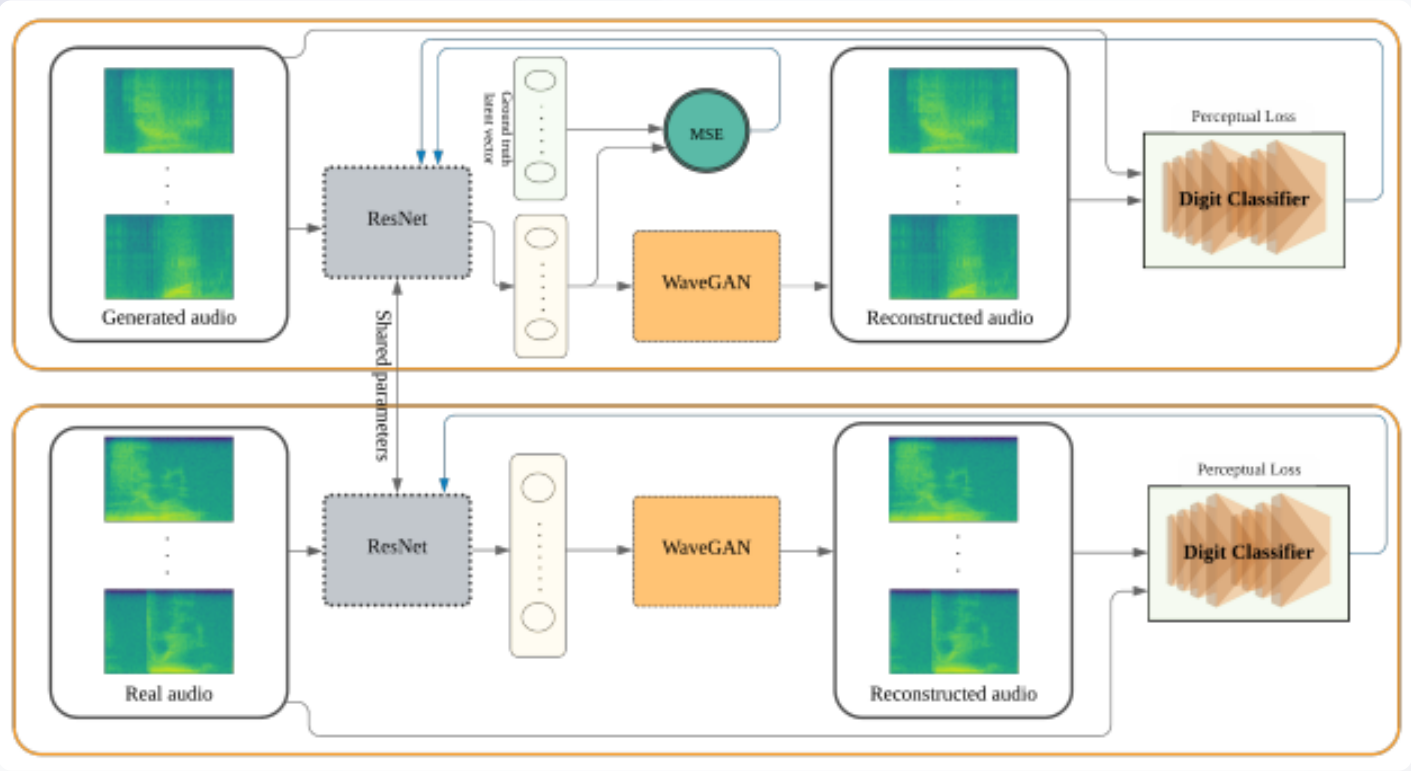
Melspectrograms

Image-like representations of audio frequency and amplitude

2D Convolutions

Process spectrograms to generate unique ones

Architecture:



SpectoGAN Output

Unique Spectrograms

Generated for each latent vector

Varying Frequency, Amplitude, and Duration

Represent spoken digits from the training data

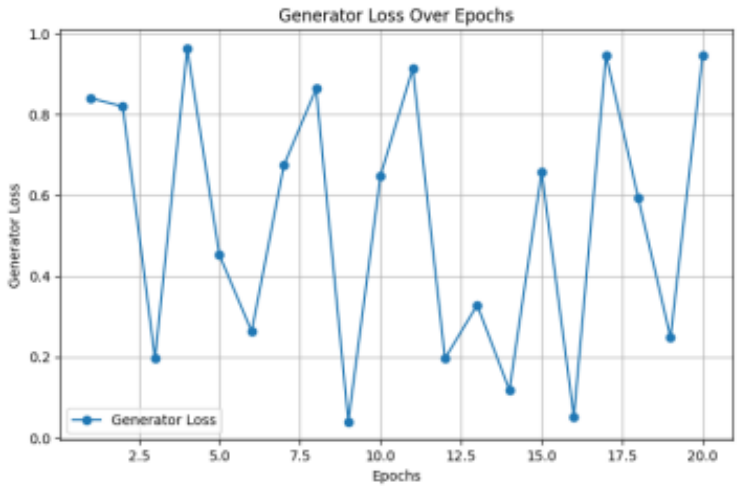
Challenges in Recovering Audio

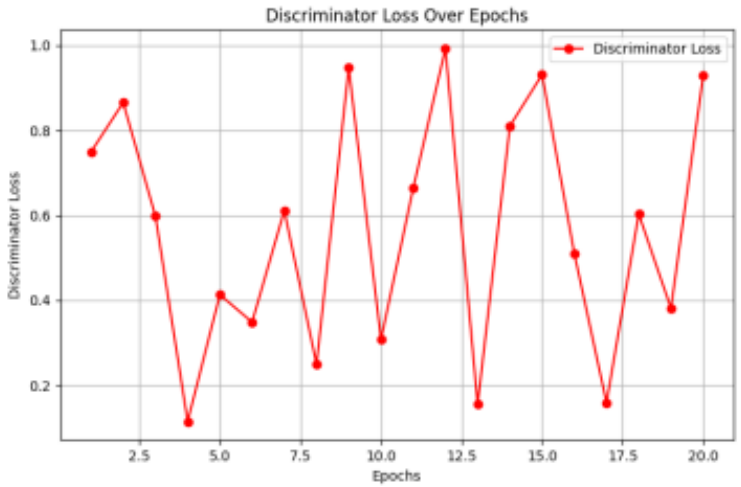
Phase information is lost in the transformation

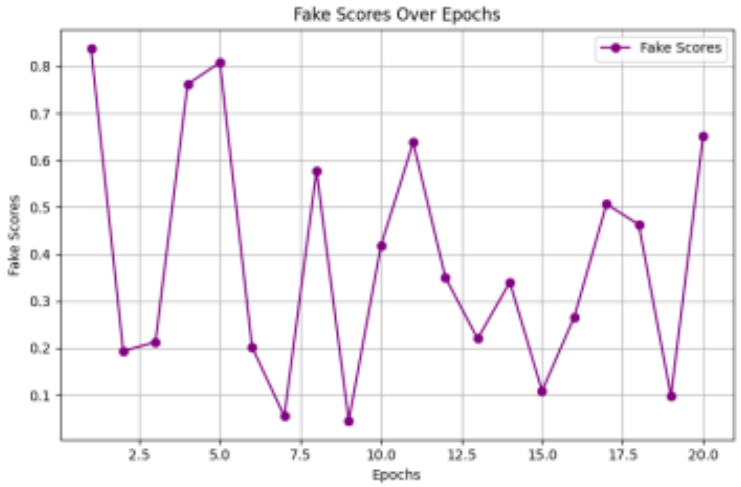
Challenges Faced

1. Using Various Libraries together in the same code for handling of Audio data
2. Unpredicatible behaviour of Generator, for which we trained multiple times
3. Matching the dimensions of the data throughout the code

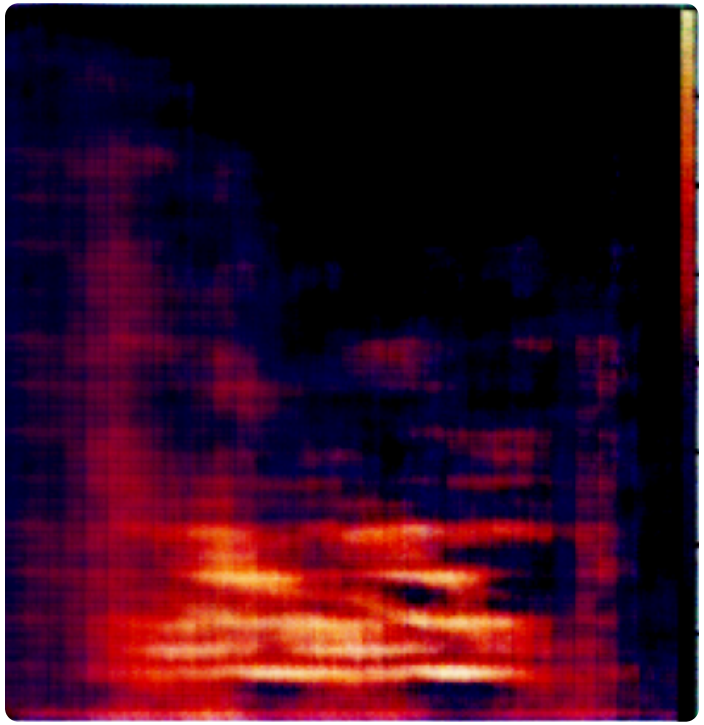
Training:







Generated spectrogram:



Future Directions

1 **Experimenting with Different GAN Architectures**

Such as WaveGAN, GANSynth, or DiffWave

2 **Conditional GANs**

Can generate audio based on labels or attributes

3 **Transfer Learning and Pre-trained Models**

Can improve the performance and efficiency of GANs