Vocal Style Transfer using Pix2Pix GAN

Zachary Nguyen, Edward Kim

Drexel University, College of Computing and Informatics

Introduction

Voice synthesis has been a popular topic of interest in the machine learning community, and there has been monumental progress within the past few years. Most notably papers such as Tacotron and WaveNet have been particularly successful using neural networks such as CycleGANs to synthesize artificial voices. These implementations focus on synthesizing a voice completely from scratch. Comparatively, there hasn't been as much focus on vocal translation rather than synthesization. Vocal translation works off of paired audio sources to translate the audio from one voice to another. Recent developments in CycleGANs has resulted in the creation of a new network known as Pix2Pix, which uses paired images to train the network as opposed to unpaired images typically used in CycleGANs. Using paired images, Pix2Pix can be used to translate from one image source to another. While recent implementations of Pix2Pix have been mainly focused on image to image translation, this project seeks to approach this from an audio standpoint. Audio is converted into a spectrogram image from which features can be extracted and fed through the Pix2Pix network for training. Once trained, the network can translate audio from one speaker into another.

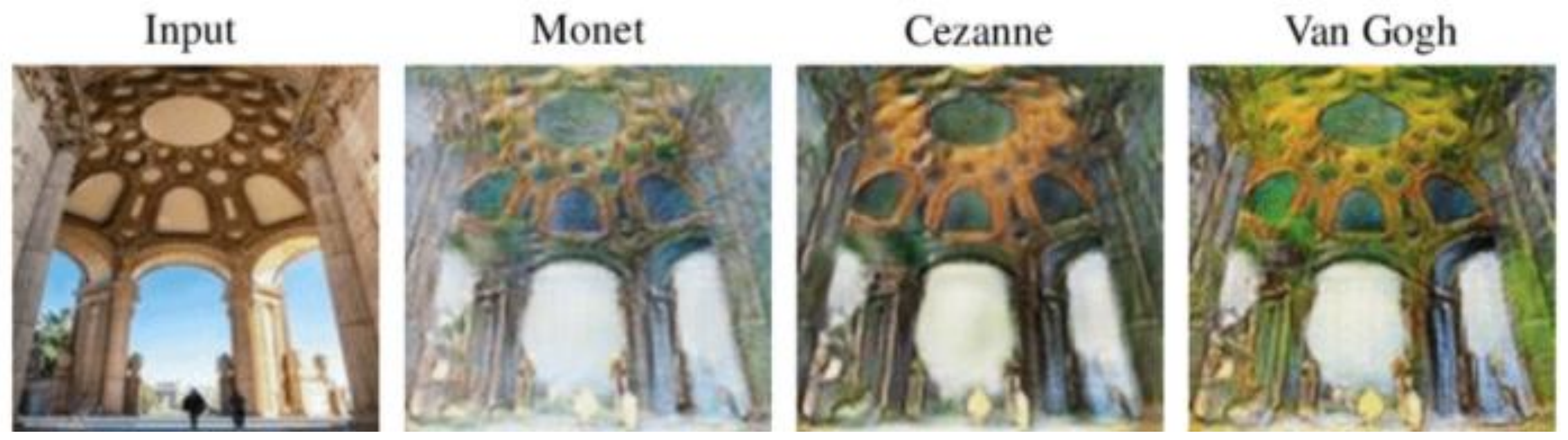

Figure 1. Example of image style transfer using Pix2Pix, 2019

Background

The neural network used for this particular project is known as a Pix2Pix generative adversarial network (GAN). Previously Pix2Pix has been used to translate images, but it has been adapted to handle audio sources in this project in the form of spectrograms. The audio signal is first passed through the Short-Time Fourier Transform function (STFT) in order to receive a matrix of complex numbers that represent the magnitude and phase of an audio signal. In order to reduce the matrix into one composed of only the real numbers, the element wise absolute value is calculated to output a matrix of magnitudes. While the STFT causes phase loss, this can be can be reconstructed later on using phase estimators. The magnitude can then be transformed into a magnitude spectrogram image which is a visual representation of the audio in the form of frequency over time.

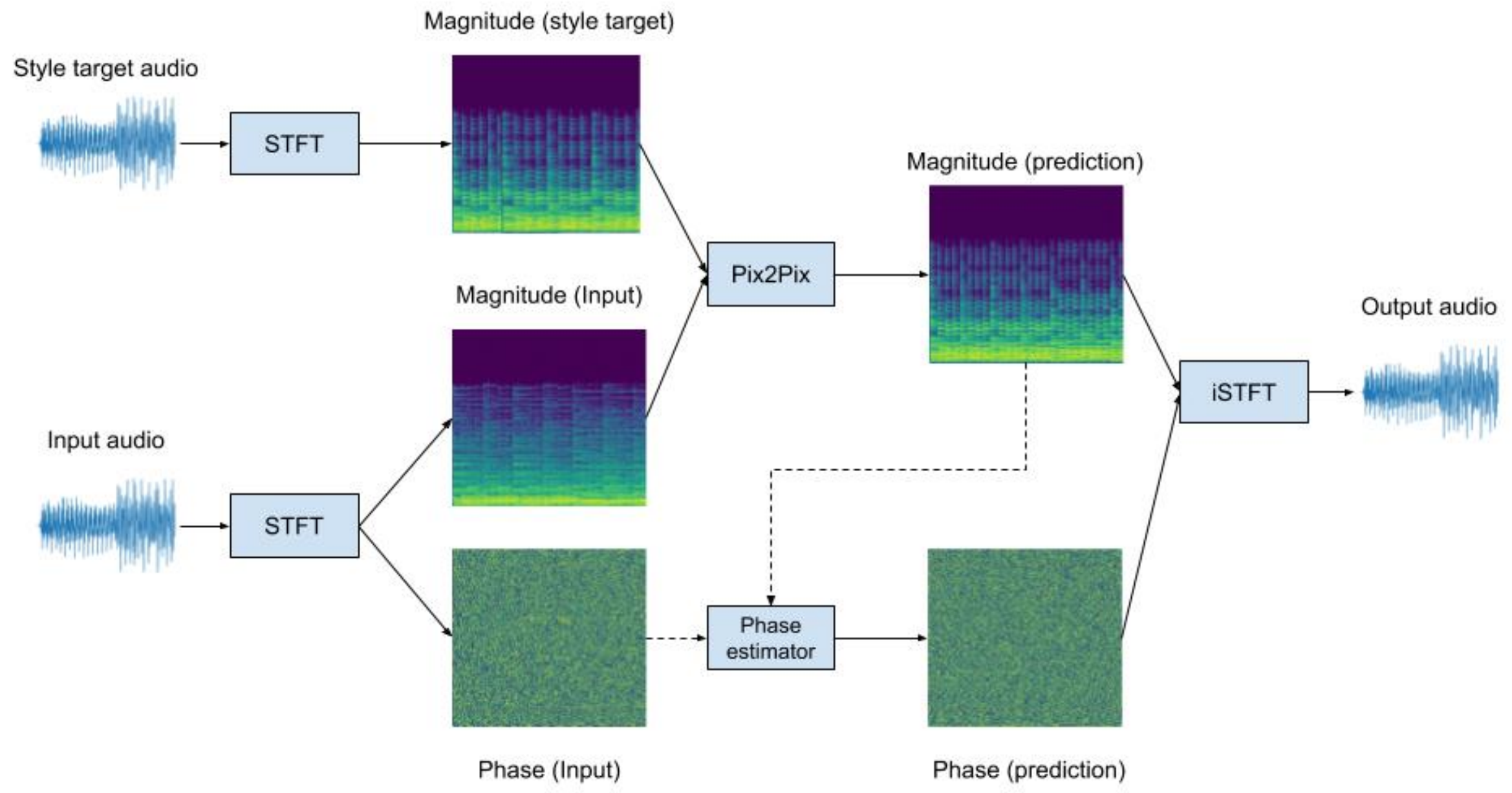

Figure 2. Example diagram of audio processing using STFT, 2019

The Pix2Pix network consists of two parts: a generator and a discriminator. The generator is the portion of the network that generates a new translated output given an input. The specific generator used in the Pix2Pix network is a U-Net as opposed to a standard encoder decoder network.

The discriminator is the portion of the network that determines if a given input is real or generated. In the Pix2Pix implementation, the discriminator itself is a PatchGAN which is a Markovian discriminator. An image describing the process can be seen below. The two parts of the Pix2Pix network together in tandem to generate new audio, with each subsequent iteration ideally improving the quality of the audio.

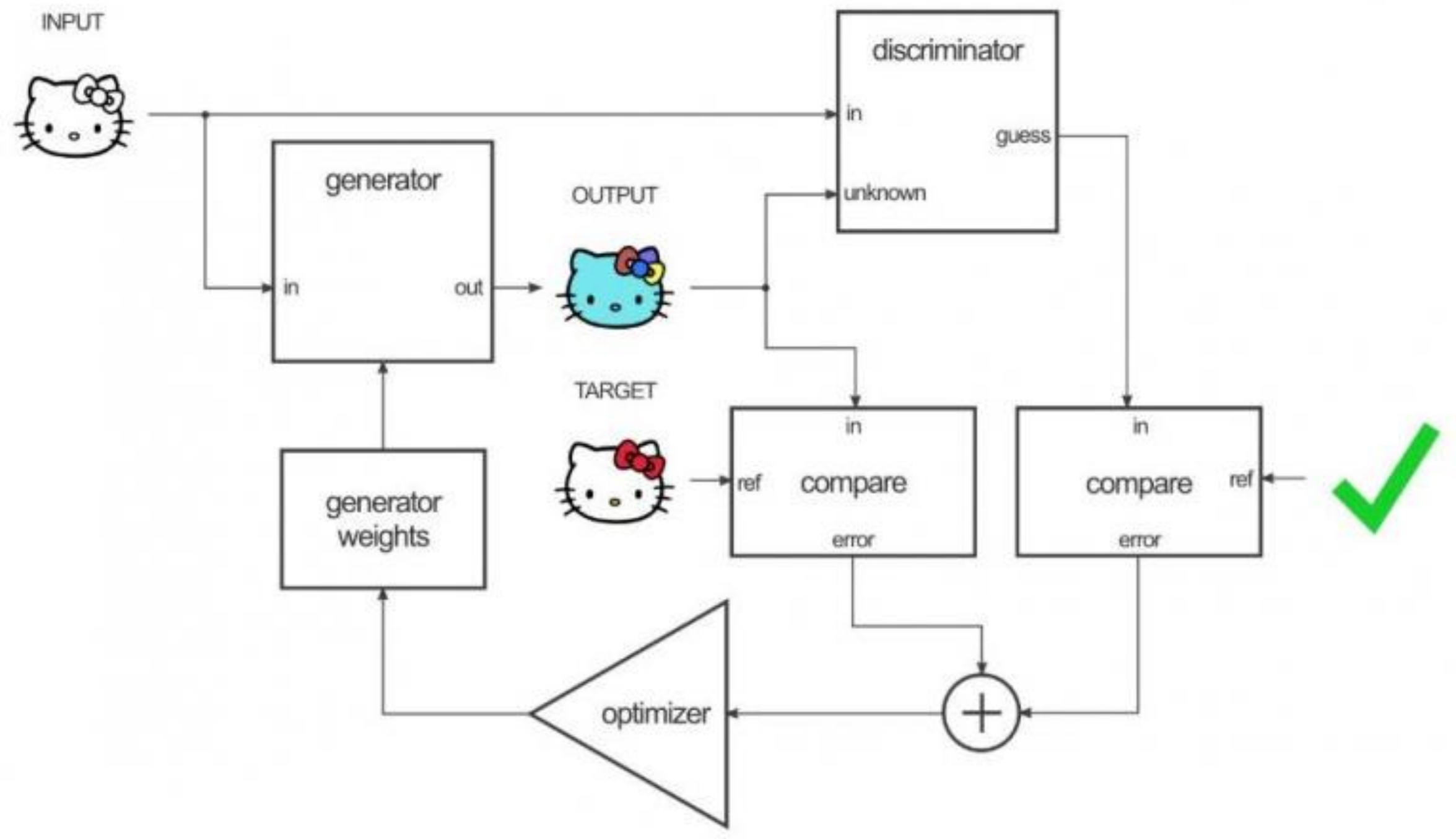

Figure 3. Example of training a Pix2Pix network, 2018

Once the training is complete, the output from the generator must also be processed to transformed back into audio. Thus we must reconstruct the phase from the magnitude matrix. This can be approximated using the Griffin-Lim algorithm, which alternates between both forward and inverse STFTs in order to approximate the phase of the original audio. Once the phase has been approximated, the audio can then be reconstructed from both the magnitude and phase.

Results

The dataset used to train the network was composed of 100 audio samples. To minimize variance in cadence, shorter audio samples were desired with a target of less than 5 seconds. The target was a naturally recorded female voice, which was padded if necessary to reach the target of 5 seconds. The starting input voice consistent of 100 female audio samples generated using Amazon Polly neural speech generator, once again padded to reach the standardized 5 seconds. Each audio sample was paired together and trained on the Pix2Pix network. The network used a batch size of 8, with a generator learning rate of 5e-3 and a discriminator learning rate of 5e-3. The program was run for approximately 4000 epochs on a K80. The an example spectrogram generated by the model is shown below.

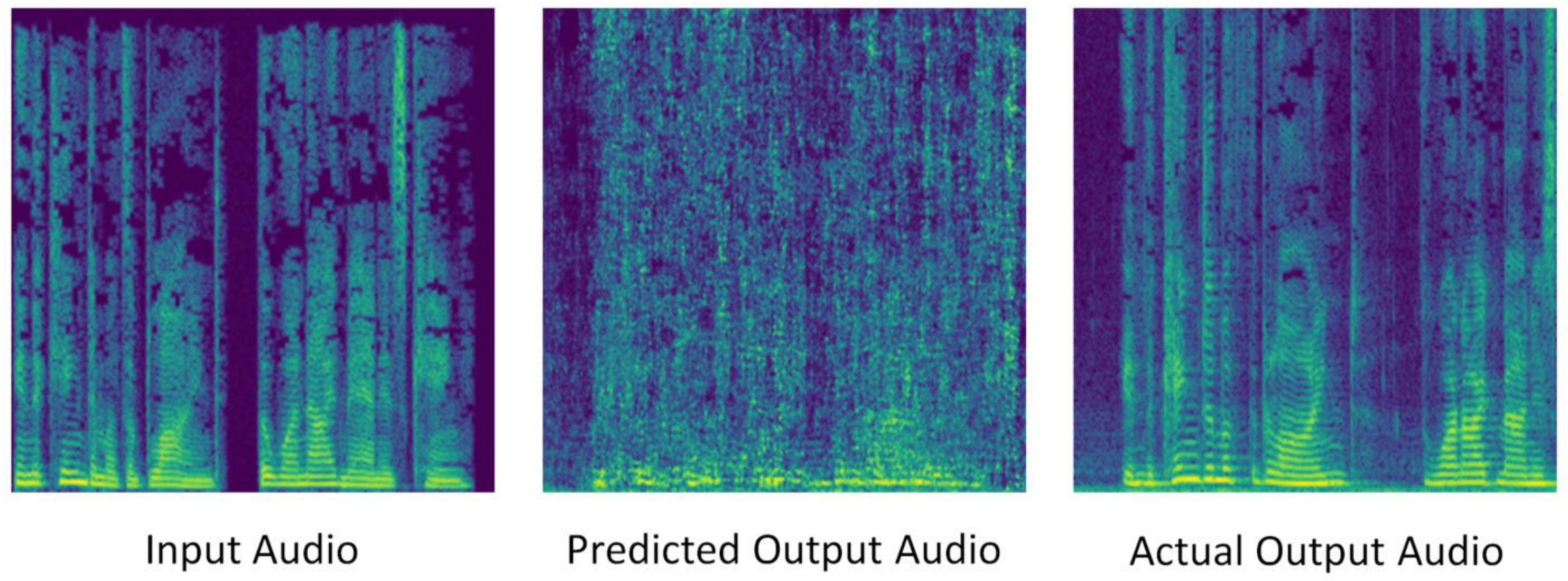

Figure 5. Graph of discriminator and generator error over training

As seen from the spectrogram, similar features can be seen developing between the predicted output from the model and the actual output. However, the bulk of the image still displays a large amount of noise. This can be confirmed not only quantitatively through spectrograms but qualitatively through the audio as well. While some structure can be discerned, it is heavily distorted and noisy.

The error over time for the generator and discriminator is displayed below. As you can see generator and discriminator error sharply decrease until roughly 500 epochs, where it reaches a joint GAN error of 11.89. The minimum joint GAN error is 5.35 with 4.19 for the generator error and 1.16 for the discriminator error at around 3900 epochs. At this point both the discriminator and generator error begins to increase

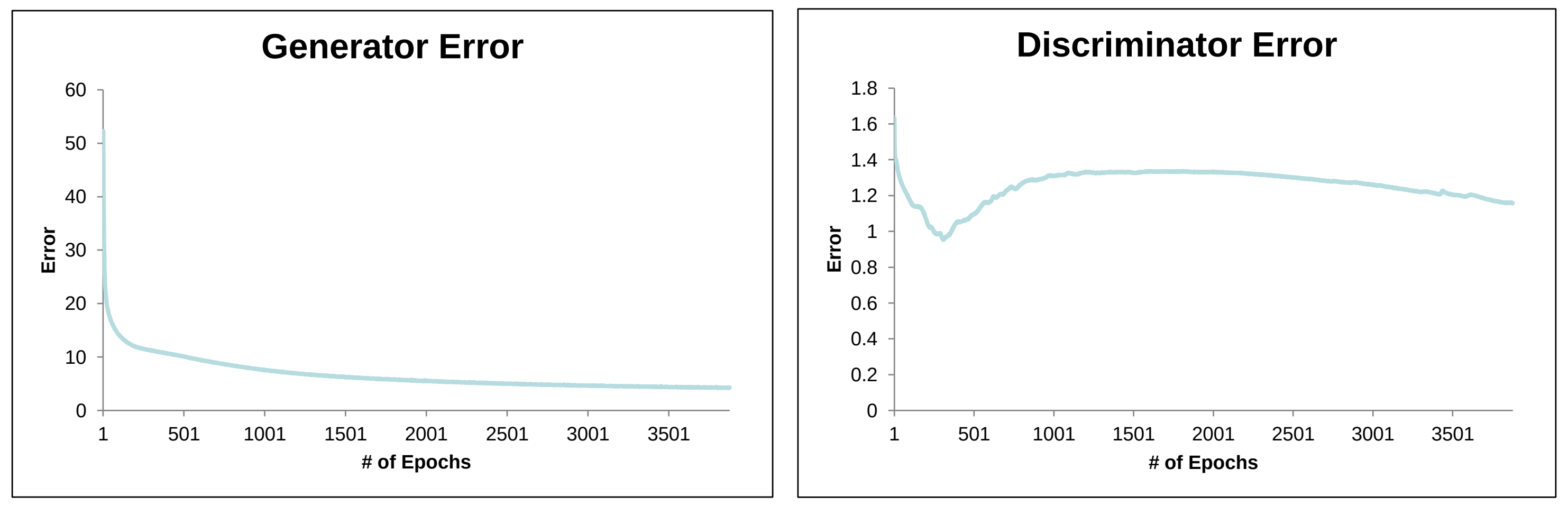

Figure 5. Graph of discriminator and generator error over training

Conclusion

While the Pix2Pix network has worked successfully on networks using synthetically produced musical audio, it struggled to produce clear audio when tasked with translating between different voices. However the results are promising, as seen in the development of similar features in the predicted and target spectrograms. With more fine tuning, the Pix2Pix network could serve as an effective alternative to pure voice synthesis in the form of vocal style translations.

Future Work

In the future several adjustments may yield more promising results. As opposed to using natural voices, translating from purely synthetic voices may yield better results. Synthetic audio would standardize cadence and sample length negating the requirement for padding or upsampling.

In addition, using larger datasets may be helpful. The dataset used in this study was roughly 8 minutes of audio data which may not be prevent generalizing. A spatial filter such as Gaussian blur may also be considered in order to reduce noise at the cost of fidelity. Lastly, a deep Griffin Lim approach may be also improve phase reconstruction

References

Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., & Grosse, R. B. (2019, May 2). *TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer*. arXiv.org. https://arxiv.org/abs/1811.09620.
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018, November 26). *Image-to-Image Translation with Conditional Adversarial Networks*. https://arxiv.org/abs/1611.07004.
Martel, H. (2019, September 14). *Pix2Pix Timbre Transfer*. https://github.com/hmartelb/Pix2Pix-Timbre-Transfer.
Masuyama, Y., Yatabe, K., Koizumi, Y., Oikawa, Y., & Harada, N. (2019, March 10). *Deep Griffin-Lim Iteration*. arXiv.org. https://arxiv.org/abs/1903.03971.
Pix2Pix - Image-to-Image Translation Neural Network. (2018, November 30). https://neurohive.io/en/popular-networks/pix2pix-image-to-image-translation/.
D. Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, April 1984, doi: 10.1109/TASSP.1984.1164317.
Yu, Xiaoming & Cai, Xing & Ying, Zhenqiang & Li, Thomas & Li, Ge. (2019). SingleGAN: Image-to-Image Translation by a Single-Generator Network Using Multiple Generative Adversarial Learning. 10.1007/978-3-030-20873-8_22.

Acknowledgements

I would like to thank Dr. Edward Kim (Drexel University) for acting as a both a mentor and consultant throughout this project. I would also like to thank the Drexel STAR Scholar program for funding this research.