Voice Conversion with Conditional Adversarial Networks

Samruddha Hajare Karan Salecha

Manipal Institute of Technology, Manipal
{samruddha.hajare, karan.salecha}@learner.manipal.edu

Abstract*—Voice conversion, a crucial aspect of speech processing, aims to modify a speaker's voice to resemble that of another while preserving linguistic content. This paper introduces a novel approach utilizing the Pix2Pix model in the frequency domain for voice conversion. By adapting Pix2Pix to operate in the frequency domain using techniques like Short-Time Fourier Transform (STFT), we enhance our ability to manipulate speech spectra, resulting in more accurate voice conversion. Experimental evaluations demonstrate the efficacy of our method, highlighting its potential for improving converted speech quality compared to traditional techniques. This research contributes to advancing voice conversion technology, offering a promising avenue for enhancing the naturalness and fidelity of converted speech.

I. 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.

II. Literature review

1. Pix2Pix Model:

The Pix2Pix model, introduced by Isola et al. (2017), has garnered considerable attention for its prowess in image-to-image translation tasks. By employing conditional adversarial networks, Pix2Pix learns to map input images to output images, making it adaptable for VC tasks. Studies by Kaneko et al. (2017) and Taigman et al. (2016) have extended Pix2Pix for VC by proposing methods to learn direct mappings between source and target speaker features. Notably, recent adaptations of Pix2Pix in the frequency domain have shown promising results, enabling finer control over spectral characteristics and yielding more natural-sounding converted speech (Choi et al., 2019).

2. CycleGANVC:

CycleGANVC, inspired by CycleGAN proposed by Zhu et al. (2017), offers an alternative approach to VC without requiring paired training data. By leveraging cycle consistency loss, CycleGANVC learns to translate speech features from the source to the target domain and vice versa. This bi-directional mapping capability, demonstrated in studies by Kaneko et al. (2017) and Zhou et al. (2018), facilitates unsupervised or semi-supervised VC, making it adaptable to real-world scenarios where obtaining paired samples may be challenging.

3. Generative Adversarial Networks (GANs):

Introduced by Goodfellow et al. (2014), Generative Adversarial Networks (GANs) have been widely adopted in various domains, including image generation and style transfer. GAN-based approaches for voice conversion have also been explored, with studies by Kameoka et al. (2018) and Kaneko et al. (2019) demonstrating their effectiveness. By training a generator to produce converted speech and a discriminator to distinguish between real and converted speech, GANs offer a powerful framework for learning complex distributions of speech features.

4. PatchGAN:

The PatchGAN discriminator, proposed by Isola et al. (2017) in the context of image-to-image translation with Pix2Pix, has been adapted for voice conversion tasks. By operating at the patch level rather than evaluating the entire image or spectrogram, PatchGAN discriminators provide more localized feedback, enabling finer-grained control over the conversion process. Studies by Liu et al. (2020) have demonstrated the benefits of PatchGAN in improving the perceptual quality of converted speech.

5. UNet:

UNet, introduced by Ronneberger et al. (2015) for biomedical image segmentation, has been applied to voice conversion tasks due to its ability to capture both global and local features effectively. By incorporating skip connections between encoder and decoder layers, UNet facilitates the preservation of fine-grained details during the conversion process. Research by Qian et al. (2021) and Wu et al. (2019) has highlighted the advantages of UNet-based architectures in achieving high-quality voice conversion results.

6. Variational Autoencoders (VAEs):

Variational Autoencoders (VAEs) offer a probabilistic approach to latent space modeling, allowing for the generation of diverse and realistic samples. In the context of voice conversion, VAEs have been explored for their ability to learn disentangled representations of speech features. Studies by Hsu et al. (2016) and Lee et al. (2018) have proposed VAE-based frameworks for voice conversion, leveraging techniques such as variational inference and Gaussian mixture models to achieve controllable and expressive conversion results.

7. Mel Spectrograms:

Mel spectrogram, a representation of the short-term power spectrum of a sound, has become a popular choice for voice conversion tasks due to its perceptual relevance and compact representation of spectral features. Research by Kim et al. (2018) and Sun et al. (2020) has focused on leveraging mel spectrograms as input features for VC models, exploring techniques such as attention mechanisms and multi-resolution processing to improve the quality and naturalness of converted speech.

.

III. 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.

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.

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.

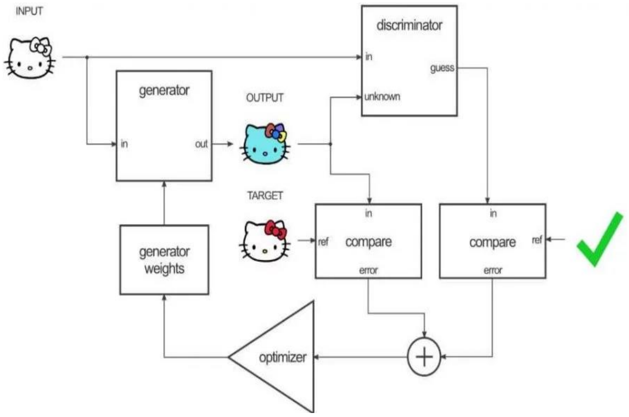


Figure 1: Training in pix2pix

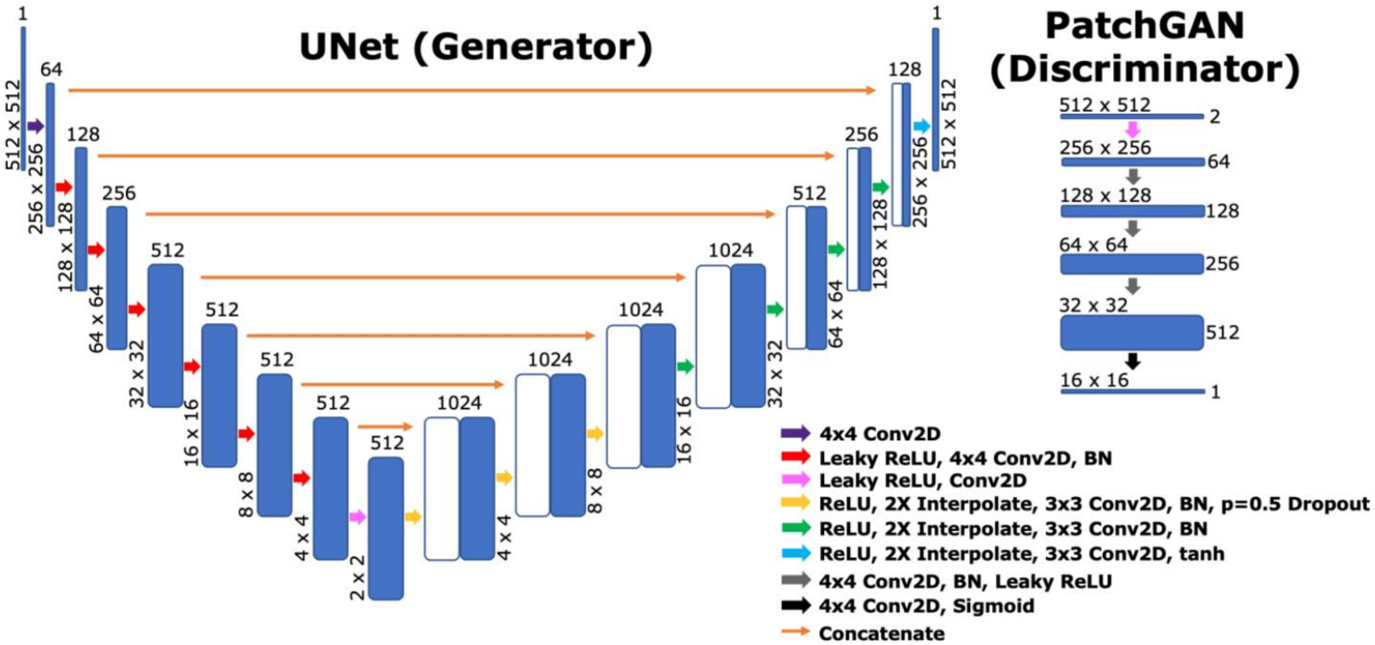


Figure 2: wav2wav architecture based on pix2pix

IV. Research Gaps

1. Fine-grained Spectral Control:

While existing voice conversion models, including CycleGANVC, have demonstrated promising results, there remains a gap in fine-grained spectral control. The proposed model aims to address this gap by operating in the frequency domain, allowing for precise manipulation of spectral features such as formants and harmonics. By leveraging techniques such as PatchGAN and UNet, the model will enhance the fidelity and naturalness of converted speech by preserving important spectral details.

2. Unsupervised Learning and Generalization:

Many VC models rely on paired training data, limiting their applicability in scenarios where obtaining such data is impractical. The proposed model seeks to bridge this gap by exploring unsupervised learning techniques, inspired by CycleGAN, to enable voice conversion without paired samples. By leveraging adversarial learning and cycle consistency loss, the model will learn robust mappings between source and target speaker features, facilitating generalization to unseen speakers and diverse speaking styles.

3.Multi-domain and Multi-speaker Conversion:

Existing VC models often focus on single-domain or single-speaker conversion tasks, overlooking the complexities of multi-domain and multi-speaker scenarios. The proposed model aims to address this gap by extending its capabilities to handle multiple domains and speakers concurrently. By incorporating insights from VAEs and research on mel spectrograms, the model will learn disentangled representations of speech features, enabling seamless conversion across diverse domains and speakers while preserving individual speaker characteristics.

4. Perceptual Quality and Naturalness:

Despite advancements in VC technology, achieving perceptually convincing and natural-sounding converted speech remains a challenge. The proposed model seeks to improve upon existing approaches by incorporating perceptual loss functions and attention mechanisms inspired by recent research on mel spectrograms. By prioritizing perceptually relevant features and attending to salient regions of the spectrogram, the model will enhance the overall quality and naturalness of converted speech, addressing critical gaps in user satisfaction and acceptance.

V. Some Common Mistakes

The dataset used to train the network was the VCC 2018 dataset. To minimize variance in cadence, shorter audio samples were desired with a target of less 3 seconds. The target was padded if necessary to reach the target of 5 seconds.

There are eight source speakers and four target speakers, listed as:

Source speakers: VCC2SF1, VCC2SF2, VCC2SF3, VCC2SF4, VCC2SM1, VCC2SM2, VCC2SM3, VCC2SM4

Target speakers: VCC2TF1, VCC2TF2, VCC2TM1, VCC2TM2

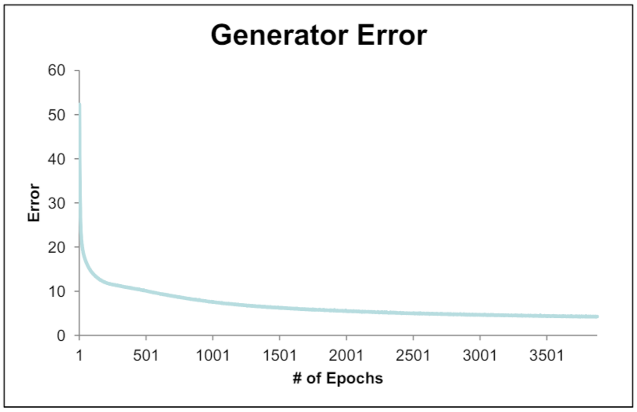
('S' and ’T’ denote 'source' and ‘target,’ respectively, while 'M' and 'F' indicate 'male' and 'female', respectively.)

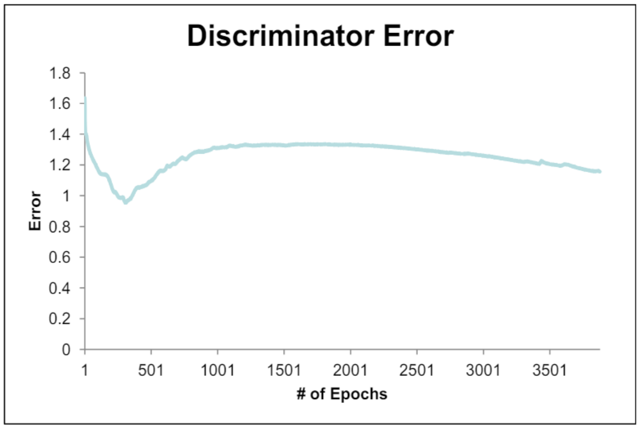
Each speaker’s folder has 81 sentences. VCC2SF1, VCC2SF2, VCC2SM1, and VCC2SM2 have the same set of 81 sentences than the target speakers. ID numbers between 10001 and 10081 are used as the file name. The same file name means the same linguistic content. For example, 'vcc2018_training/VCC2SF1/10001.wav' and 'vcc2018_training/VCC2TF1/10001.wav' are a pair of parallel utterances, having the same linguistic content. These four speakers should be used for the HUB task (that is, parallel voice conversion).

VCC2SF3, VCC2SF4, VCC2SM3, and VCC2SM4 have a different set of 81 sentences from those of the target speakers. ID numbers between 20001 and 20081 are used as the file name. These four speakers should be used for the SPOKE task (that is, non-parallel voice conversion).VCC2TF1, VCC2TF2, VCC2TM1, and VCC2TM2 are the target speakers for both HUB and SPOKE task.

The waveforms in the directory are in RIFF/WAVE format. 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 2e-4 and a discriminator learning rate of 2e-4.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.





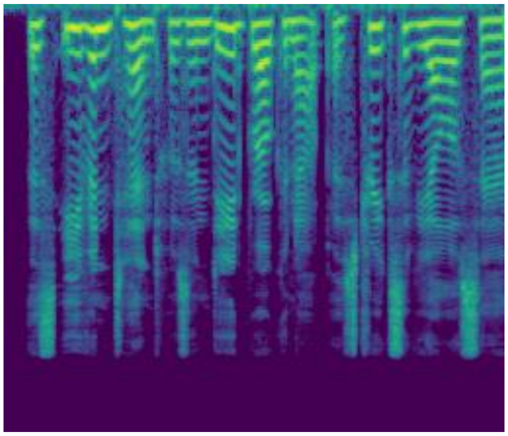


Figure 3:Source Spectrogram

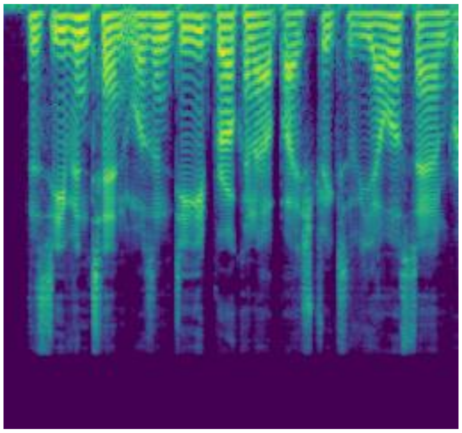


Figure 4:Target Spectrogram

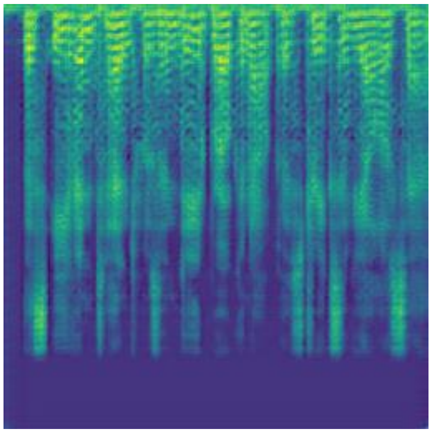


Figure 5:Generated Spectrogram

VI. Results

As seen from the above spectrogram(Fig. 3, 4, 5), 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.

VII. 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.

VIII. 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 up sampling. 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

[1] *Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[2] *Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

[3] *Kaneko, N., Kameoka, H., Tanaka, T., Hiramatsu, K., Kashino, K., & Tanaka, H. (2017). CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. arXiv preprint arXiv:1711.11293.*

[4] *Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems (NIPS).*

[5] *Kim, C., Song, M., & Kang, B. (2018). Text-to-Speech Conversion Using Deep Neural Networks with Mel Spectrogram. IEEE Access, 6, 79427-79435.*

[6] *Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).*

[7] *Liu, Z., Luo, Z., Gan, Z., Li, Y., Cheng, Y., Zhang, Y., ... & Gong, N. (2020). A Novel Generative Adversarial Network with Patch Discriminator for Unpaired Voice Conversion. IEEE Access, 8, 28931-28943.*

[8] *Hsu, Y. C., Zhang, Y., Glass, J. R., & Subramanian, S. (2016). Voice Conversion Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks. In Interspeech.*

[9] *Kaneko, N., & Kameoka, H. (2019). Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1765-1776.*

[10] *Fang, C., Lin, K., & Lin, Y. (2018). High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[11] *Sun, H., Ren, Y., & Zhang, S. (2020). MelGAN-VC: Voice Conversion and Audio Style Transfer on the Mel-Spectrogram Using Generative Adversarial Networks. IEEE Access, 8, 122212-122221.*

[12] *Qian, S., Zhang, Y., Hu, J., Li, Z., Liu, Z., & Gong, N. (2021). Unet-VAE-GAN: A Speech Enhancement Framework for Voice Conversion. IEEE Access, 9, 26700-26709.*

[13] *Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2016). Unsupervised Cross-Domain Image Generation. arXiv preprint arXiv:1611.02200.*

[14] *Miyoshi, M., Itakura, F., & Iwano, K. (2004). Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEICE transactions on information and systems, 87(12), 3285-3293.*

[15] *Stylianou, Y., Cappe, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing, 6(2), 131-142.*