Cycle-consistent Generative Adversarial Networks (CycleGANs) for the Non-Parallel Creation of Fake Voice Media

Daniel Fleury[1], Angelica Fleury[2]
[1]The John Marshall School
[2]University of Minensota Duluth
Corresponding author: daniel.fleury02@gmail.com

Abstract

The upsurge of Generative Adversarial Networks (GANs) in the previous five years has led to advancements in unsupervised data manipulation, sourced feature translation, and precise input-output synthesis through a competitive optimization of the discriminator and generator networks. More specifically, the recent rise of cycle-consistent GANs enables style transfers from a discrete source (input A) to target domain (input B) by preprocessing object features for a multi-discriminative adversarial network. Traditionally, cyclical adversarial networks have been exploited for unpaired image-to-image translation and domain adaptation by determining mapped relationships between an input A graphic and an input B graphic. However, this integral mechanism of domain adaptation can be applied to the complex acoustical features of human speech. Although well-established datasets, such as the 2018 Voice Conversion Challenge repository, paved way for female-male voice transformation, cycle-GANs have rarely been re-engineered for voices outside the datasets. More critically, cycle-GANs have massive potential to extract surface-level and hidden feature to distort an input A source into a texturally unrelated target voice. By preprocessing, compressing, and packaging unique acoustical voice properties, CycleGANs can learn to decompose speech signals and implement new translation models while preserving emotion, the intent of words, rhythm, and accents. Due to the potential of CycleGAN's autoencoder in realistic unsupervised voice-voice conversion/feature adaptation, the researchers raise the ethical implications of controlling source input A to manipulate target voice B, particularly in cases of defamation and sabotage of target B's words. This paper analyzes the potential of cycle-consistent GANs in deceptive voice-voice conversion by manipulating interview excerpts of political candidates.

Keywords: Generative Adversarial Networks (GANs) – Autoencoders –Deep Learning Ethics –Machine Learning

INTRODUCTION

The emergence of Generative Adversarial Networks (GANs) in 2014 marked the first implementation of a hypercompetitive minimax 'two-player' game between a discriminative and generative network to produce convergently similar outputs from an input distribution (Goodfellow et al., 2014). Analogically, the generator is akin to a 'scam artist/counterfeiter' attempting to increase the failure of the discriminator to detect fraudulent copies. Similarly, the discriminative network is analogous to a police/law enforcement, discerning fake copies. This dynamic one-on-one adversarial relationship between an extensive Generator network and the discriminative model allows the architecture to proportionately improve the cost functions of each player in the game, increasing the likelihood of similarity to the input distribution. Traditionally, this approach of adversarial loss has been applied to novel image synthesis, where a discriminative network consists of an image classification meta-architecture inversely competing against a generator network which adds/removes noise from the distribution.

Although Generative Adversarial meta-architectures are notorious for pixel-image synthesis, Cycle-Consistent GANs offer similar application in progressive voice transformation and conversion: Cycle-GANs marked a role in non-parametric modeling in image-to-image translation, by mapping relationships between a source input A and the target Input B (Zhu et al., 2018). Conventional GANs, however, simply exploit discriminative models and distribution samples to distinguish similarities between the dataset and the synthetic output. By mapping relationships between Input A and Input B domains, adaptation of styles, extracted features, and properties of the image can be understood cyclically between two Cyclical generators- *G* and *F* (Fang et al., 2018).

In a Cycle-consistent GAN, the *G* generator objectively maps the domain and features of input *A* to input *B*, while *F* maps input *B* to *A*. In convention with an adversarial model, the discriminators D_B and D_A forcibly reinforce the generator based on newly mapped output relationships. In other words, discriminator D_B forces *G* to transform inputs that come from domain input A (source dataset) into outputs that will be indistinguishable from input B (target data), while D_A respectively implements the inverse. This generative dynamic of the architecture in mapping A→ B can be modeled by B (G:A → B), while domain adaptation from B→ A can be modeled by *B* (G:A → B). Although domain adaptation has mainly been operationalized on style transfer experiments with image datasets (e.g. transferring and exchanging styles of different artists), the autoencoders and feature extractors of CycleGANs can be applied to simplecomplex acoustical features. More specifically, Voice Conversion (VC) can be attained by taking advantage of cyclical domain adaptation from inputs A$\boxed{\leftrightarrow}$B and 'style-transferring' acoustical features of the human voice while still maintaining inherent vocal integrity and linguistic information.

Voice Conversion and Cycle-consistent GANs (CycleGANs

Voice Conversion (VC) is a generalized method of transforming speech signals of a source speaker (input *A*) into a target speaker (input *B*) by convergently exchanging layered acoustical features. Ultimately, the objective of VC is to preserve acoustical integrity and linguistic information while transforming nuanced accents and phrasing of target A's voice. Overarchingly, VC frameworks have an objective in balancing and mapping transferrable audio features between a source and target track. VC meta-architectures are versatile in dataset manipulation when provided with non-parallel and parallel systems of analyses. A parallel VC system maintains explicit linguistic information between adaptive domains of A and B, where sentences wording, and structure are identical between the source/target datasets. This coordination of a parallelized dataset creates inherently similar acoustical features between the source and target speakers of interest by having them record identical sentences of comparable duration, thus improving precision of mapped relationships.

However, non-parallel VC subsystems map and parameterize models of disjoint data, provided that a corpus of linguistic information has differentiable content, sentences, phrasing, and accents. In other words, a non-parallel architecture cyclically extracts acoustic features from the source and target speakers regardless if their recorded sentences are identical. Due to the non-mutual and unshared nature of non-parallel acoustic features, sensitive and accurate divergence-convergence of two voice types is difficult. Moreover, mapping discrete acoustical connections between the input and output samples becomes increasingly unstable for the model because of randomized discoordination of sentences, voice waveforms, etc. On the other hand, parallelized VCs are not scalable, as it is limited to source and target input voices with identical sentence types.

Non-parallelized training is notorious for weak acoustical feature extraction, subpar input-output mapping mechanisms, and inferior domain adaptation between source and target voice samples. Cycle-consistent Generative Adversarial Networks (CycleGANs), however, provide a solution to performance discrepancies and low levels of convergence during training. The premise of CycleGANs is that a forcible relationship between input and output distributions exist, so a cycle-consistency is established in that input information is invariant and input-output validations are indistinguishable. Due to this synchrony and forcible matching of source and target distributions (*A* and *B*), unpaired distribution mapping can be learned feasibly while still maintaining accuracy and believability of voice conversion.

Although Cycle-consistent Generators have been used adversarially for non-parallel voice transformation in already-established datasets, this paper considers the far-reaching ethical implications of exploiting Cycle-GANs for media and voice sabotage through an unsupervised mapping of relationships in an uncontrolled setting (i.e. the Cyclical generator manipulates unrelated voices of source A and target B). CycleGANs have mastered a theatre in the forced, unsupervised matching of voice distributions, however, can be negatively manipulated for audio recordings outside prominent datasets such as the Voice Conversion Challenge (VCC) 2012 and 2018 repositories. By reinforcing exploitive and manipulative voice recordings with the researcher's target track, the source speaker can be codominantly controlled in a non-parallel dynamic. In other words, execution of intergender training and validation can be accomplished with unrelated scraped recordings of the target voice (e.g. Donald Trump, Angela Merkel, or Barack Obama) and source voice (i.g. Researcher's voice). This project investigates the potential of CycleGAN non-parallel voice conversion in manipulative voice transformation to possibly defame individuals, distribute fake media, and deceive public listeners. The researchers employ both a survey of n=100 sample size and mean opinion scores (MOS) to quantify the misleads of voice transformation and tampered audio perception on a general public.

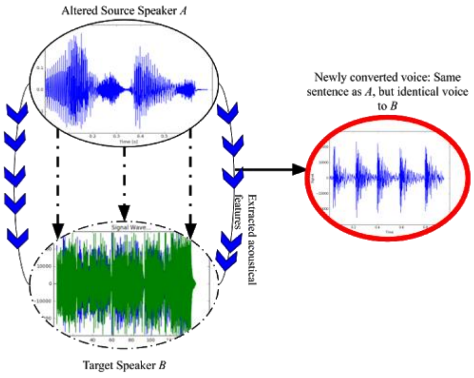

Altered Source Speaker *A*
Newly converted voice: Same sentence as *A*, but identical voice to *B*
Extracted acoustical features
Signal Wave
Target Speaker *B*

Figure 1.1: Holistic Cycle-GAN conversion with source and target domain adaptation accomplished by acoustic feature extraction.

Voice Conversion and Cycle-consistent GANs (CycleGANs

The coinciding rise of polarity in the American media/news outlet landscape and the emergence of 'DeepFake' generative networks creates a high-risk plane of manipulation of video and images on a limited input. In other words, training and deploying a generative model may only require one source input of data to generate plausible results. In the theatre of pixel wise image-video transformation, Samsung, for example, used 'Few-Shot Adversarial Learning' to generate realistic and plausible talking head models (Zakharov et al., 2019). This strike of vulnerable face manipulation sparked relative concern in the public, news outlets, and general media, due to the vast implications of using one source image to potentially distort facial expressions and physical behavior. More importantly, Progressive Deep Convolutional GANs (PDCGANs) rapidly adopt facial photometries, anatomical features, and morphology to distort and create novel images of the human subject. In the realm of text generation, researchers, for example, from the University of Montreal, have exploited generative adversarial models for the determination and creation of natural language (i.e. language identical to natural human speech and writing) (Subramanian et al., 2017). Ultimately, long-established threats in videographic-image distortion have been made possible by mechanisms of Generative Adversarial Networks and the manipulation of dataset distributions.

Although the negative implications of GANs have been inspected primarily in image-video networks, little research into the distortion of human speech signals and waveforms have been considered on a scale of human defamation, defrauded media, and fake news generation. As presented, Cycle-consistent GANs have potential in fortifying non-parallel datasets, forcibly closening their distributions to eventually adapt domain A to domain B. In theory, this closening of distributions means that whoever's source voice can be freely manipulated to respectively alter a target voice. CycleGANs do not require lengthy audio samples, with a minimum of only 80 five to seven second snippets in both the source and target speakers' training datasets. Large repository Voice Conversion Challenge (VCC) datasets from both 2012 and 2018 are the only non-parallel use of voices in CycleGANs, however, the algorithm has never been reinforced for scraped voices of politicians in media and news.

By decisively manipulating the researcher's source voice, voices of Angela Merkel, Donald Trump, and Barack Obama, can be altered and controlled as target partitions. More importantly, once the domains have been successfully mapped and adapted between the source and target datasets, non-parallel conversion of new sentences can be created. This paper not only provides an overview to Cyclic generation and its potential in non-parallel voice manipulation but proposes an ethical discussion into its far-reaching impact into media and news, including legal testimonies, voice recorded evidence, newsrooms, and the changing of voices in large authority (e.g. a nation's president or prime minister).

Cycle-consistent GAN Architecture

As presented earlier, a Cycle-GAN typically consists of two disparate generators- *G* and *F* - with two discriminators associated with the source and target partitions, D_A and D_B. Moreover, generator G acts as a mapping function by connecting distribution A to distribution B, whereas generator F serves as an inverse mapping function from distribution *B*→*A*. Additionally, the A}=F(B) discriminators are able to distinguish between real and generated distributions between *A* and *B* cylically and consistently. The distinction for *A*→*B* can be described as $\hat{A} = F(B)$, whereas the distinction of the B (target) distribution can be described as $\hat{B} = F(A)$. Ultimately, the objective/goal of the model in mapping the distributions of A and B can be illustrated by and the mapping of B by $\{A_j\}_j{}^{N}{}_{=1}$ .This model pathway of mapping and connecting distributions is also demonstrated in the schematic below. Cycle-consistent networks manipulate circular autoencoding functions that intermatch the source and target distributions in accordance with the Discriminators (D_A and D_B).

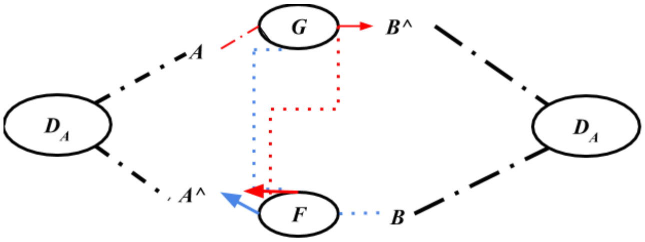


Figure 1.2: Discriminator-generator urgency and loss competition is demonstrated by generators G and F recursively outputting feedback into a loop with discriminators A and B.

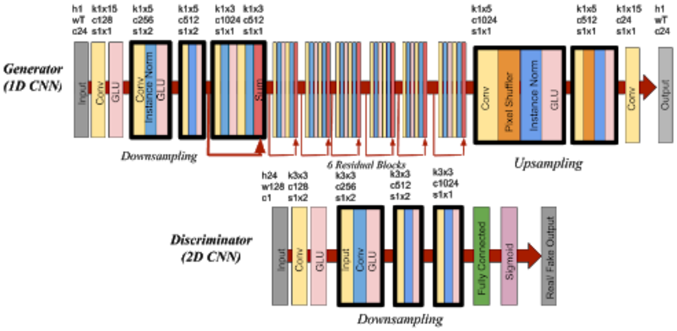


Figure 1.3: An extensive generator, defined as a 1-Dimensional CNN composed of downsamples, residual blocks, and upsamples, competes with a 2-Dimensional discriminator (comprised of downsamples).

In other words, the architecture progressively responds to discriminatory values from both D_A and D_B to reinforce the generators G and F. Additionally, the network exploits two objective loss functions- adversarial loss and cycle consistency loss. Adversarial and Cyclical-consistency loss defines a two-step process where adversarial loss makes A and A^ or B and B^ as close as possible. The cycle-consistency loss of the network ensures that input A_i or B_i grasps and retains its primary form after passing through the initial two generators. This combined mechanism of adversarial cycle-consistency solidifies non-parallel mapping techniques in unpaired training. Adversarial loss can be described by the following objective function, mapping generator G to its respective discriminator D_Y:

$$L_{cyc}(G,F) = E_{A\sim p\ data(A)}\left[\left|\left|F\big(G(A)\big) - a\right|\right| 1\right] \\ + E_{b\sim p\ data(b)}[\left|\left|G\big(F(b)\big) - b\right|\right| 1]$$

In addition to an adversarial loss mechanism which chronically synchronizes target transformations from Generators G and F (correspondence with input A and B), Cycle-consistency loss is used to retain previous unpaired voice information and extend it throughout the training process. Cycle-consistency loss is akin to the objective function of an autoencode, minimizing the ultimate difference between input and outputs A/B which eases the voice conversion process. Cycle-consistency loss is described as:

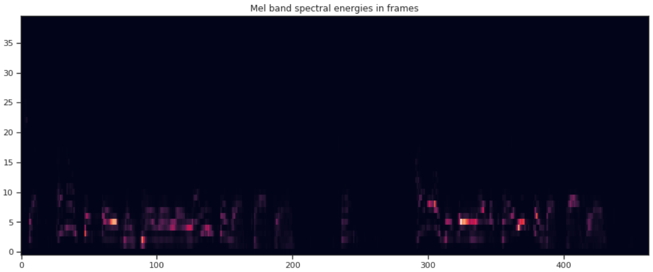
$$L_{GAN}(G, D_B, A, B) = E_{B\sim p\ data(b)}[\log D_b(b)] + E_{A\sim p\ data(a)}[\log(1 - D_B(G(x))], (1)$$

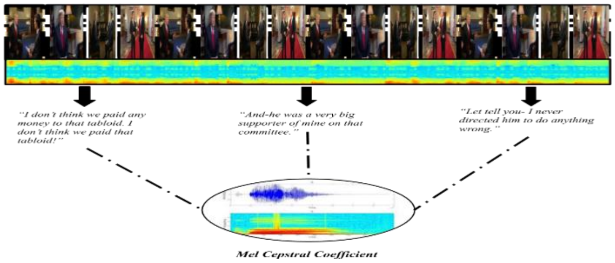
The combination of adversarial and Cycle-consistency loss enables rapid unpaired voice signal transformation and recursive conversion. In summary, adversarial loss allows the network to converge the two voices in preparation for intermatching, whereas Cycle-consistency loss retains original voice information and acoustical properties, allowing it to minimize the difference between the A and B distributions in the network. The combination of both the discriminator and generator network is illustrated in the diagram below. The architecture takes advantage of a 1D gated convolutional neural network (a gated CNN) for the generator, and a 2D Gated CNN for the discriminator. More importantly, the network uses Mel-cepstral coefficients as data inputs for voice conversion. The architecture inherently manipulates a large generator network with a smaller discriminator network using the same convolutions. The generator and discriminator networks are interconnected via functional mapping functions that operate on non-parallel data. Additionally, the CycleGAN network is a six-layered feed forward neural network, with around 128, 256, 256, or 128 neurons in each layer. Sigmoid activation was implemented in each block/hidden layer of the network. The most accessible implementation of CycleGAN resources Pytorch libraries, however, we found it more practical to exploit Tensorflow for visualization.

Methodology and Setup

I. Website Data Scraping: Accessing Source and Target Audio Tracks

Scraping, extraction, and a finalized compression of audio snippets required comprehensible audio tracks ranging at most of ~10 seconds each. The integral Cycle-consistency of the network required high level acoustical feature to operationalize on non-parallel data distributions, so it was ensured that audio snippets included at least one full length sentence with no abrupt linguistic cutoffs. In other words, the audio snippets were representative of natural human sentences easily expected in a conversion. The target candidate of interest in the experiment was Donald J. Trump, the current 45th U.S. president. Due to the massive wealth of audio recordings and interviews from public media news outlets (e.g. CNN, Fox News, MSNBC, etc), localizing and extracting at least 80 audio tracks from one or two interviews was feasible. The following diagram illustrates example phrases and sentences extracted from the interview(s) into acoustical Mel-Cepstral coefficients.




"I don't think we paid any money to that tabloid. I don't think we paid that tabloid!"
"And-he was a very big supporter of mine on that committee."
"Let tell you- I never directed him to do anything wrong."
Mel Cepstral Coefficient

Figures 1.4

Mel-Cepstral acoustical feature extractions require consistent and acoustically lucid data with clarity in enunciation, pronunciation, and intonation. The voice signals and combined waveforms of the Cepstral files required not only full-length sentences but sentences variety of vocal capability. Videotaped interview candidates, such as Donald J. Trump, provided essential emotional and jagged phrasing for Mel-Cepstral coefficients to operate on. By extracting discernible snippets of speech extracts from both the source and target speakers, Cycle-consistency and adversarial loss is expected to decrease. The following diagram demonstrates the removal and snipping of speech examples from Donald Trump's interview.

II. Acoustical Feature Preprocessing

As noted earlier in the background, Mel-general Cepstral Coefficients are convenient in compressing and serializing acoustical feature-sets into understandable units for the Cycle-GAN. Mel-cepstral coefficients are primarily designated by frame separations, indicating disjoints, breaks, new phrases, or significant changes in the audio spectrogram. These resulting spectrogram changes are an essential acoustical feature mapped between the source and target distributions (A & B). The following figures demonstrate MCEPs frames on both a standard 5.4 second data snippet of audio recordings, figures 1 and 2 representing Donald Trump, while 3 and 4 correspond to the researcher's source voice. Additionally, the energy frames, normalized frames, and MCC frames are revealed in figures 1.6-1.15 (1.6, 1.7, 1.10-1.12=Donald Trump while 1.8, 1.9, 1.13-1.15 represents the researcher's source voice).

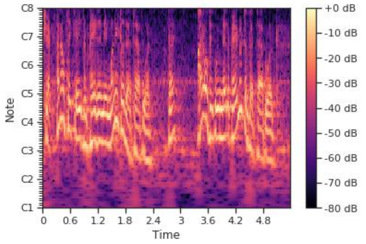


Figure 1.6

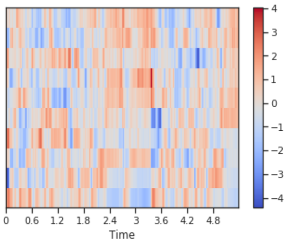


Figure 1.7

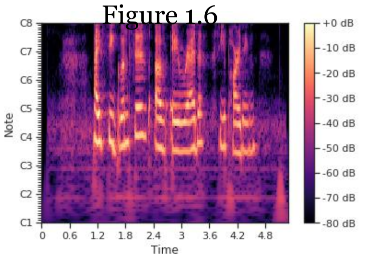


Figure 1.8

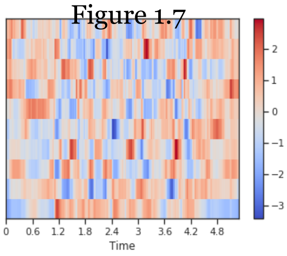
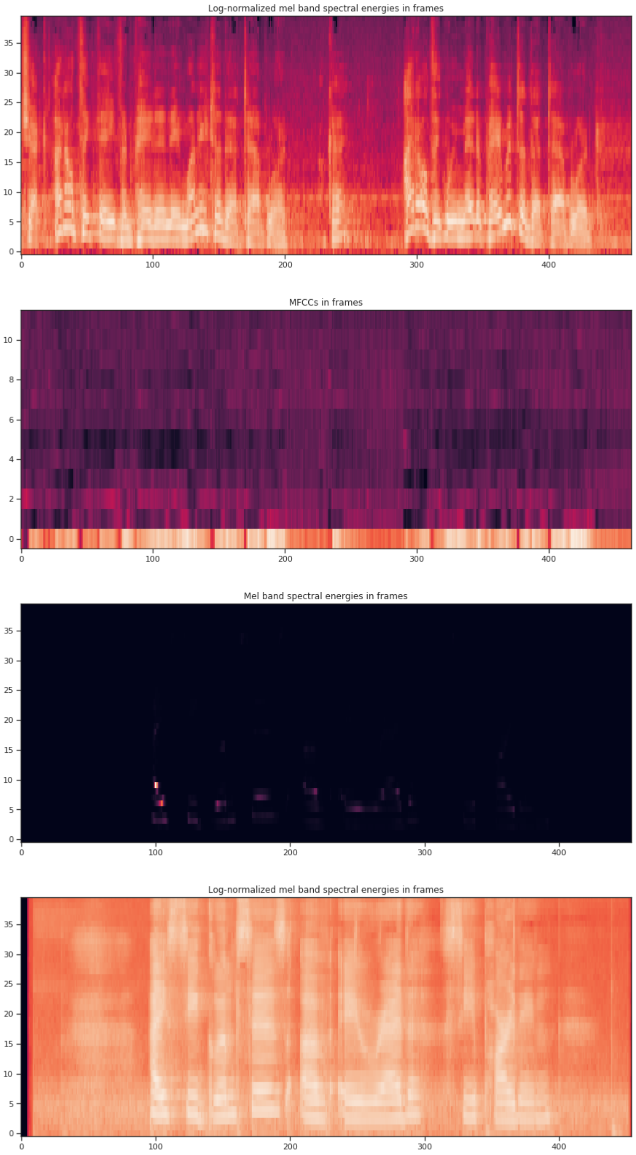
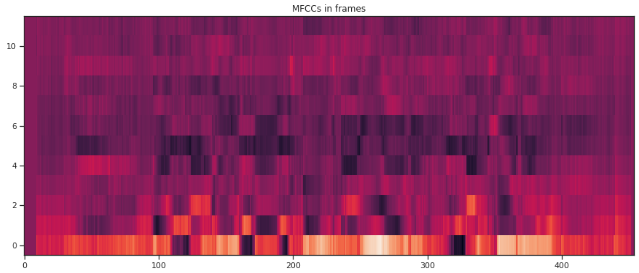


Figure 1.9

Figures 1.10-1.15





Implementation of CycleGAN via Tensorflow

The execution of a Cycle-consistent Generative Adversarial Network involved various levels of prepackaging the architecture, installing dependencies, and configuring hyperparameters/variables. Traditionally, CycleGANs are exploited with high-order Pytorch libraries due to adjustability of training, however, this project exploits Tensorflow for ease of visualization and monitoring via a Tensorboard interface. Integral visualization techniques allowed elucidation of voice spectrograms and progressive training sensitivity over time. More importantly, monitoring average cost functions and loss functions with detectable divergent-convergent patterns indicated progress of source-target signal transformation. The Cycle-GAN architecture was cross-integrated and containerized with a Tensorflow-Keras backend which enabled easy feature extraction of Mel-Cepstral coefficients due to inherited Python libraries.

The default hyper parameterization for the cyclical-GAN backbone involved a rebalancing of critical learning rate and source-target batch size uptakes. Simply put, the network started with a mini batch size of 1, with two distinct learning rates valued for the discriminator and generator variables. The generator learning rate = 0.0002, while the discriminator = 0.0001. More importantly, based on training examples and features extracted, the discriminator decayedh at a rate of the previously stated learning rate/200,000, respectively the same for the generator learning decay variable. In order to amplify robust acoustic feature extraction in a relatively noise-polluted target dataset, the number of Mel-Cepstral coefficients was increased to around 36 to collect variated voice characteristics and solidify unpaired intergender translation between a completely distinct pair of source-target voices.

Iterative hyper parameterization required observance of negative generative-adversarial model outcomes, including modal collapse, mediocre domain adaptations, misbalanced domination of the generator and discriminators, and inadequate divergence-convergence trends. Although disruptions during the early-middle-late phases of training were possible, the Cycle-consistent GAN architecture was not as susceptible to training/computational inconsistencies. In other words, data distributions did not lean towards collapsed biases, variances, or generator-discriminator dominations compared to traditional vision models (e.g. DCGANs). The critical scalability, jumpstarting nature, and low-interference behavior of the CycleGAN proves it to be optimal in domain adaptation, especially across low-dimensional data distributions such as voice vectors, signals, spectrograms, etc. The ease of launching CycleGANs was accompanied by autoencoded domain adaptation, low incidences of parameter model collapse, and the malleability of the algorithm on non-paralleled and unstructured data.

Generator and Discriminator Domain Adaptation Losses

Discriminative and generative adversarial loss, as previously characterized in the objective functions, was attributed by the difficulty of deception and feature synthesis in relation to the discriminator. In other words, a decrease in discriminative loss equates to a lower cost value in the minimax competition, indicating a success of the generator in deceiving the discriminator. Although discriminative loss decreases in believable magnitudes of domain adaptation and style transfer, immediate rapid loss can be indicative of a mediocre discriminator failing to capture authentic acoustical features. The following figures (1.4-1.7) illustrate divergence of discriminator and generator network losses between domains A and B.

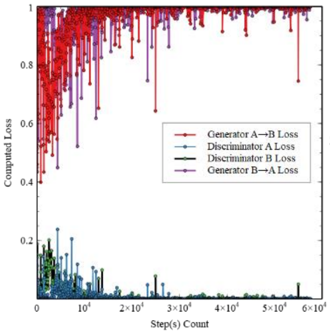


Figure 1.16

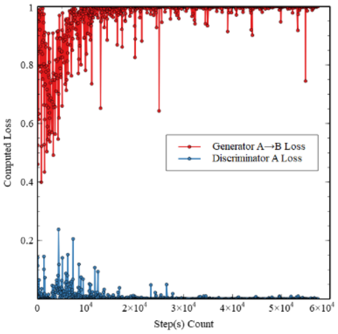


Figure 1.17

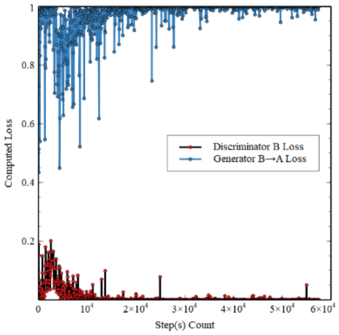


Figure 1.18

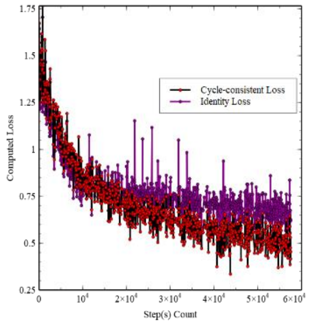


Figure 1.19

Ultimately, figures 1.16-1.19 delineated above reveal long-term convergence-divergence trends of the Cycle-consistent-Generative Adversarial Network with the range of 60,000 steps. As seen broadly in figure 1, generator vs. discriminator loss regressions diverged consistently until they both converged at minimal values of ~0.01 or ~0.99. Long term divergence patterns, at least in the range of 10,000 steps, are indicative of well-competing generator and discriminator models in the general GAN framework. In other words, we can deduce that the observed divergence patterns indicates a successful domain adaptation from A→ B by manipulating pre-packaged Mel Cepstral Coefficients and Non-parallel Cycle-consistent mapping. In addition to divergences between the generator and discriminator, there is a coordinated convergence between discriminators A and B. Discriminators A and B plateau near loss values of 0.01 in synchrony, highlighting similar rates of domain learning and mapping between two Non-parallel distributions.

Moreover, convergence in generative mapping between A→ B and B→ A can be acknowledged in the figures above. Generators A and B converge at high loss values, sustaining competition with the discriminator agent network and responding effectively between domain adaptations in A and B. Ultimately, these coordinated convergence patterns between A and B in both theaters of discrimination and responsive generation show low levels of collapse in the model, high consistency, and successful mapping of the Non-Parallel distributions in the dataset between Donald Trump's voice and the researchers source dataset. The following images show the Mel-generalized Cepstral Coefficient frames and demonstrates their connection to domain adaptation.

The post-processed results of eventual convergence-divergence in the GAN architecture were notable due to successful avoidance of domain collapse, modal collapse, and inconsistent mapping activations. The trends of generator vs. discriminator divergence was indicative of a source A and source B data distribution that functioned well in non-parallel intergender voice transformation. Although generative loss values underwent expected gradual divergence from the discriminator and convergence approaching 1.00 loss values, the discriminator rapidly converged to a near-zero value threshold, causing a long term case of potential distribution overfitting and low-quality validation output. The unprecedented convergence of the discriminator may signal a need for a larger and more varied dataset. Discriminator loss values typically converge rapidly due to quick deceptivity by the generator, nearly immediately tricking the discriminator into a low loss and cost threshold. For future improvement, the number of audio snippets/examples will be increased to ~1000 for both the source and target distributions, enabling more diverse MCEPs fed into the network.

In relation to holistic loss values, figure four illustrates the comparable magnitude fluctuations of identity and cyclical loss in a downward trend. Identity los and cyclical loss both mutually coincide in the process of domain adaptation and mapping parameters of source A to source B. The regressive negative slope of both the identity and cycle-consistent loss highlights success in adaptively mapping distributions A and B on a closed training set. Identity loss is loosely defined as the gradual mapping of the source-target domains while maintaining the previous identity source voice throughout the process (i.e. not distorting the identity signals and spectrograms). Cycle-consistent loss is broadly defined as the cost in mapping distribution A→

B with regards to sacrifices made by the generator and the discriminator.

Conclusion and Ethical Discussion

The non-parallel transfer of acoustical features, parameters, and styles are critical processes of discriminator-generator competition in a Cycle-consistent generative network. By implementing CycleGANs across an unstructured dataset of 80 batched voice snippets in both the source and target domains, unsupervised mapping and intergender connections between the distributions was possible. More importantly, Cycle-GANs prove to be malleable and scalable as a network architecture due low modal collapse, quick divergence across the generator, and output of believable output validation voices. Additionally, the CycleGAN was capable of exploiting Mel-general Cepstral coefficients as packaged acoustical feature data, preprocess it, and use it in mapping distributions. The use of Cycle-consistent adversarial loss allowed both competitors (the generator and the discriminator) to progressively improve in both distinguishing validation output and producing more deceptive results.

Although Cycle-consistent GANs are advantageous in a series of fields such as hearing assistance, unsupervised vocal feature transfer, and rapid intergender domain adaptation, the architecture presents outweighing ethical implications and costs. CycleGANs use adversarial and generative loss, along with autoencoders, to compress increasingly complex layers of vocal features into convertible inputs. The capacity of a CycleGAN to rapidly adopt domain-level features and transfer them in an intergender nature creates possibilities of media fraud, sabotage of recordings, legal proceedings, and general media speech tracks. More importantly, the source dataset can easily be manipulated (as demonstrated) and created with false sentences by simply recording what the individual would like the target track to vocalize/speak. Cycle-consistent GANs prove to be effective in recognizing and transferring unstructured/non-parallel data across distributions, regardless of wide variations in the MCEPs and gendered acoustical features.

Bibliography

Feng, F. (2018). High-quality nonparallel voice conversion based on cycle-consistent adversarial network. *Arxiv,1*, 1-5. Retrieved April/May, 2019, from https://arxiv.org/pdf/1804.00425.pdf.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. NPIS, 2014, pp. 2672–2680.

L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice conversion using deep neural networks with layer-wise generative training,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 12, pp. 1859–1872, 2014.

Subramanian, S. (11). Adversarial Generation of Natural Language. Arxiv, 1, 1-18. Retrieved 2019, from https://www.aclweb.org/anthology/W17-2629.

T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion based on speaker-dependent restricted Boltzmann machines,” IEICE Trans. Inf. Syst., vol. 97, no. 6, pp. 1403–1410, 2014.

Zakharov, E. (2019). Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *Arxiv,* 1-19. Retrieved from https://arxiv.org/pdf/1905.08233.pdf

Zhu, J. (18). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Arxiv,1*, 1-18. Retrieved 2019, from https://arxiv.org/pdf/1703.10593.pdf.