EE599 - Deep Learning (Spring 2020)

Project report on

SpectroGAN: Emotion transfer on images and spectrogram of speech

By,

[1]Vineeth Rajesh Ellore, Ashwin Telagimathada Ravi, Karkala Shashank Hegde

[1] All authors contributed equally to this work

TABLE OF CONTENTS

Abstract

This project aims to perform emotion transfer mainly on spectrograms of speech signals using a Cycle Generative Adversarial Network (Cycle GAN). We have implemented a Patch GAN for the discriminator model, and an encoder-ResNet transformer-decoder for the generator model in our Cycle GAN. We have worked on optimising our Cycle GAN after running multiple configurations of the number of ResNet blocks in our generator model and the size of Fast Fourier Transform (FFT) length of audio signals. We have experimented with emotion transfer on various pairs of emotions. The performance of the model has been tested with unseen audio from RAVDESS dataset, lexically different audio from different speakers and audio samples from Kannada and Hindi languages. Noticeable emotion transfer has been recorded for emotion pairs: calm-fearful and calm-anger.

1. Introduction

1.1 Summary

Traditionally, GAN's have been used for style transfer, and they have shown to work really well on images. In this project, we have attempted to change emotions in spectrograms of speech signals. To do this we have used a cycle Generative Adversarial Network (GAN). As a baseline we have also experimented with cycle GAN based techniques to transfer emotions on face data.

1.2 CycleGAN

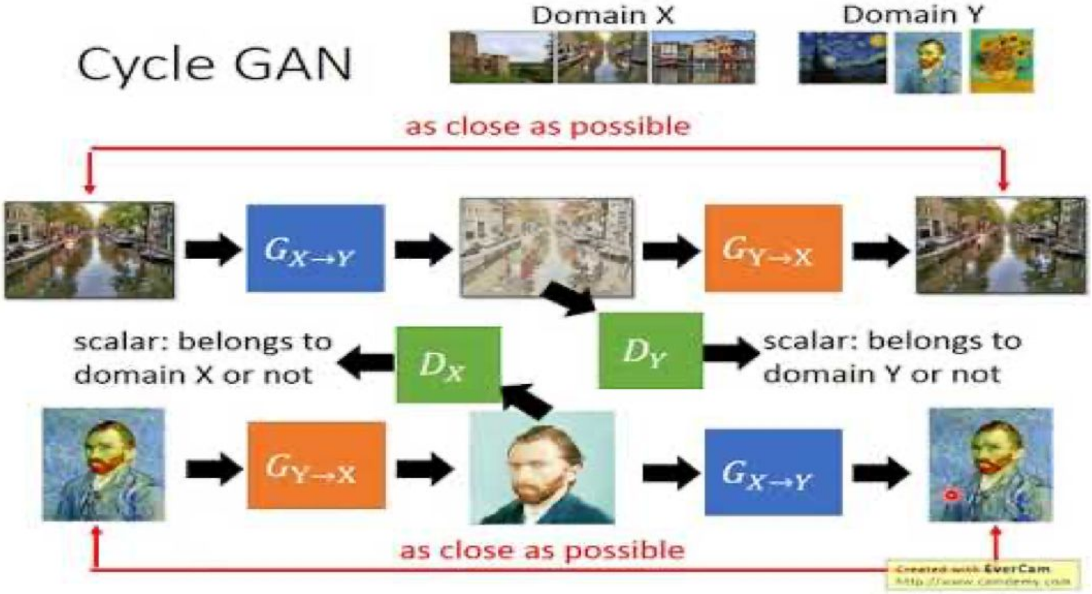


Fig 1. Flowchart of Cycle GAN [1]

A cycle GAN is an approach to learn a mapping between two domains of unpaired datasets by training a deep Convolutional Neural Network (CNN) [2]. A cycle GAN typically uses two generator-discriminator pairs, which are trained in an adversarial process, similar to normal GANs. In cycle GANs, apart from the generator and discriminator adversarial losses, these models are additionally regularized by forward and backward cycle-consistent losses. This is achieved by defining a composite model which links a generator with its discriminator, and also with the other generator. The model is trained by updating the weights of one generator while keeping the other generator non-trainable. The weights of the trained generator are shared later with its discriminator and the other generator. Both the generator-discriminator pairs are trained until an equilibrium is reached.

1.3 Spectrogram

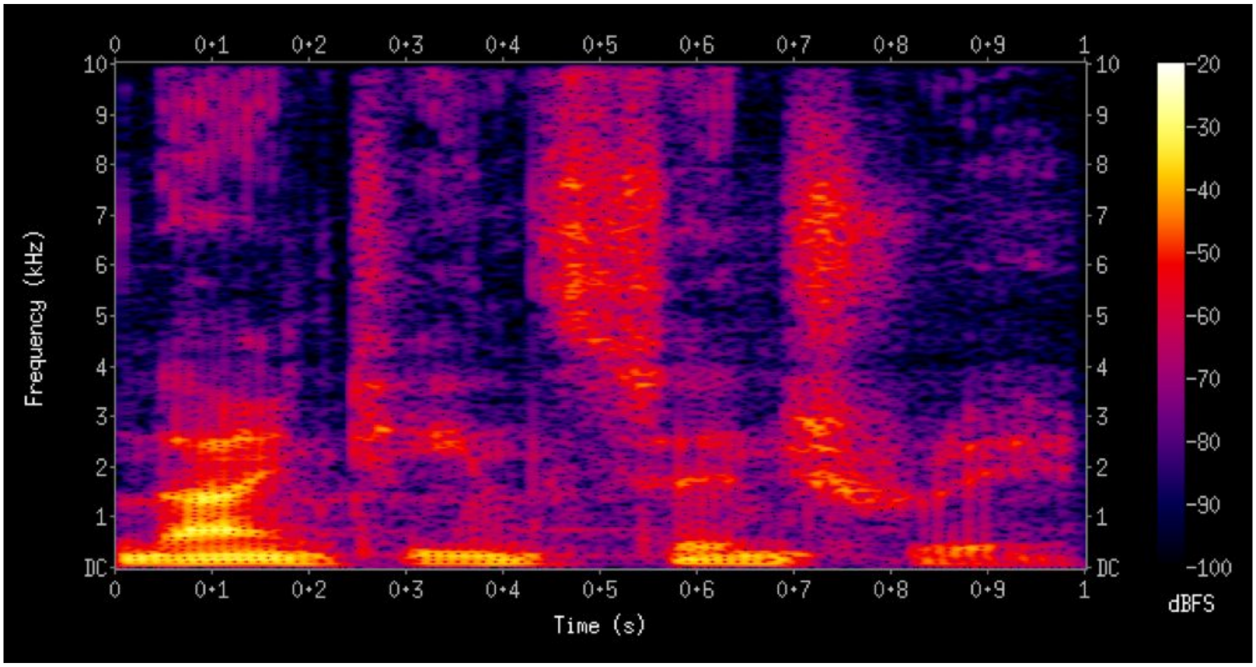


Fig 2. Sample spectrogram image [3]

Spectrogram is a method of representing the frequencies of an audio signal as it varies with time. It is constructed by calculating the Short Time Fourier Transform (STFT) on small overlapping time intervals of the audio signals. In this project, we use a magnitude only spectrogram, which does not contain any information about the phase of the signal.

1.4 Griffin-Lim algorithm

A widely used algorithm based on the redundancy of STFT for phase reconstruction of signals was proposed by Griffin and Lim [4]. It is an iterative algorithm, starting from an initial phase (often zero or random) and updating it every step after calculating the inverse STFT of corresponding length.

2. Literature Survey

Most of the research for image-to-image translation using GANs [5] has been supervised training with paired dataset from the two domains we want to translate between. The paper on Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [2] in 2017 alleviates this problem by introducing the idea of cycle GANs for unpaired image-to-image translation.

On the other hand, there has been a lot of experimentation on audio style transfer using NLP [6] and multi-layer perceptrons. These methods have leveraged hand-engineered features such as zero crossing rate, pitch, tempo, spectral features (centroid, contrast, roll-off), chroma features and Mel Frequency Cepstrum Coefficients (MFCC). Modern methods using automated feature extraction have also shown promising results for audio style translation. Emotion Recognition of speech spectrograms using Deep Convolutional Neural Networks [7][8][9] demonstrates this idea. [10] shows a GAN based approach for the reconstruction of an audio signal solely from a magnitude spectrogram.

In this project, we have combined these methods to demonstrate a novel approach to style transfer on speech signals using cycle GANs for image-to-image translation of spectrograms.

3. Dataset

For this project, we have used the Ryerson Audio-Visual Database for Emotional Speech and Song (RAVDESS) dataset [11]. This dataset consists of audio samples from twenty four professional actors, 12 male and female actors in total. These actors have vocalised two lexically-matched statements in a neutral North-American accent. The dataset contains audio for eight emotions namely anger, calm, disgust, sad, happy, fearful, neutral and surprised. There are approximately 200 audio samples for each emotion in which each audio sample is three seconds long, 16 bit and sampling rate of 48 kHz.

For the purpose of this project, we have downsampled the sampling rate to 16 kHz.

For the baseline experimentation on emotion transfer of facial images, we have used a dataset from kaggle [12].

4. Methodology

4.1 Model Architecture

The generator model consists of down-sampling convolutional blocks, six Resnet blocks and up-sampling convolutional blocks. The model outputs pixel values with the shape as the input and pixel values are in the range [-1, 1].

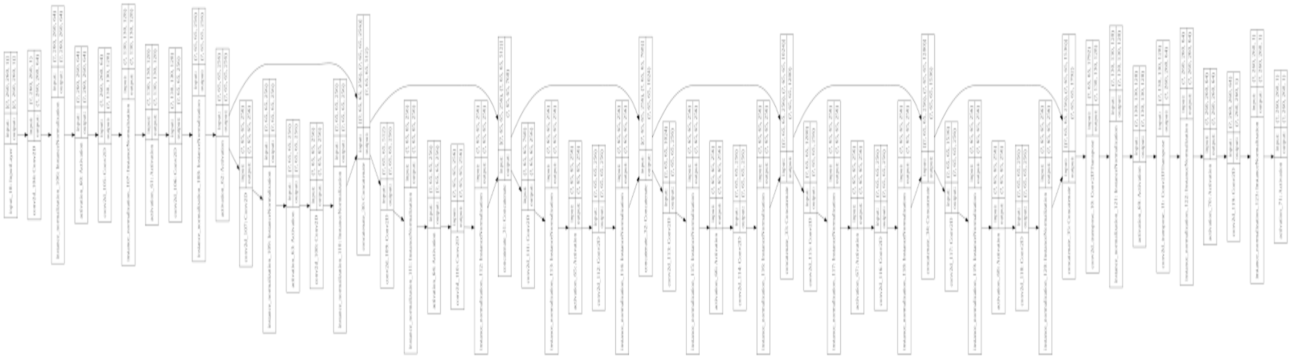


Fig 3. Generator model architecture. Consists of an encoder with 3 conv blocks, a transformer with 6 ResNet blocks, and a decoder with 3 conv blocks

The discriminator model implements a patchGAN architecture [2] which maps the entire input into individual patches of real or fake. For 260x260 image size, we get a 17x17 output patch of zeros or ones if the input image was fake or real respectively. For 520x520 images we get a output patch of 33x33. The loss function here is mse between the desired patch and the obtained patch.

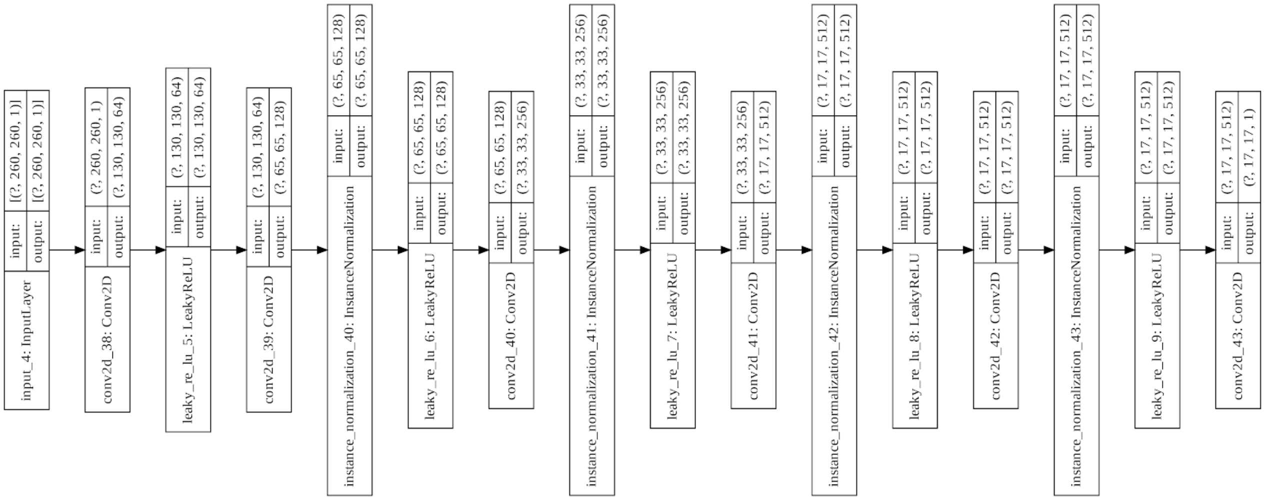


Fig 4. Discriminator model architecture

4.2 Style transfer on Facial expressions

We initially wanted to have a hyper parameter tuned model ready for style transfer on images. Thus, we implemented a patchGAN to have style transfer on emotions. We worked on the facial expressions data set [12]. The hyper parameters were fine tuned to get noticeable differences in the generated images.

4.3 Audio-Image Data Conversion

Once we had a model that was capable of performing style transfer, we focused on our audio dataset. The dataset was categorized and arranged domain wise. For Data conversion from audio to image and vice versa, the following configurations were used.

a) Audio to Image: To convert the audio to spectrograms we sampled the audio at 16000 Hz and performed stft with two experimental configurations.
    i) FFT length of 512 and used a hop length of 256. The source audio files were also trimmed to obtain a spectrogram of size 257 X 257. This image padded was with 0's to get a 260 X 260 array.
    ii) FFT length of 1024 and used a hop length of 128. The source audio files were also trimmed to obtain a spectrogram of size 513 X 513. This image padded was with 0's to get a 520 X 520 array.

b) Image to Audio: To convert the generated spectrograms to audio, we used the griffin-lim algorithm on the clipped image. We made sure that the FFT length and the hop length used in the ISTFT was the same as before.

4.4 Fine Tuning Generator size and FFT length

a) Generator size: The transformer portion in the generator is responsible for the style transfer. Therefore it is intuitive that different style transfer tasks can have different configurations for transformation. In our experiment we decided to have many Resnet blocks in our transformer. This was done since we wanted to avoid vanishing gradients. We experimented with 3 size configurations of ResNet blocks, namely, 3,6 and 9 blocks.

b) FFT length: Considering audio files can be converted to spectrograms using different FFT lengths, we wanted to test on two values, 512 and 1024.

5. Results

We have attached spectrograms for various emotion pairs in the following sections. The audio files for these spectrograms can be found on the project website.

5.1 Facial Emotion transfer

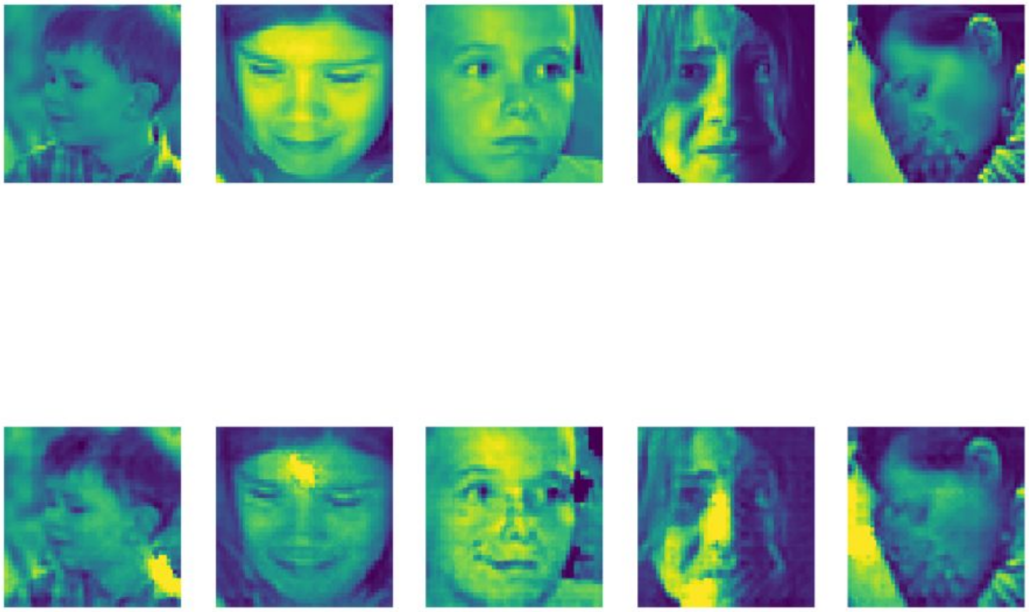


Fig 5.: First row consists of input images (sad) and second row consists of corresponding output images (happy).

The above figure represents our results from experimentation after running the model for 50 epochs. This Cycle GAN is not optimised for facial emotion transfer as our project focuses mainly on emotion transfer of speech signals.

5.2 Emotion transfer on spectrogram of speech

The following sections show the results obtained from various emotion transfers for different configurations of the model. The spectrograms show the change in emotions from the source domain to the target domain.

5.2.1 Optimization of ResNet blocks

We have first tried to optimize the size of the generator model for our application of emotion transfer on spectrogram of speech. As mentioned in previous sections, our generator model consists of encoder- ResNet transformer- decoder configuration. Here, we have considered the number of ResNet blocks as a hyper-parameter. Three configurations with three, six and nine ResNet blocks have been tried and the results are shown in Table 1.

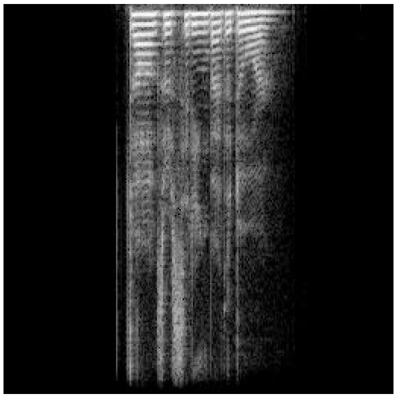
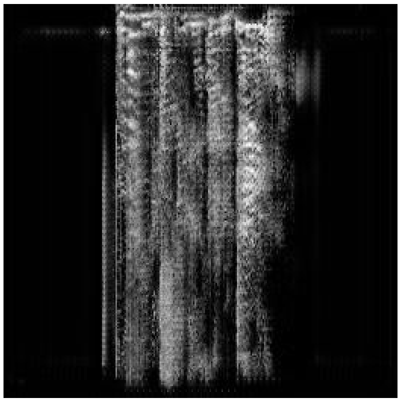
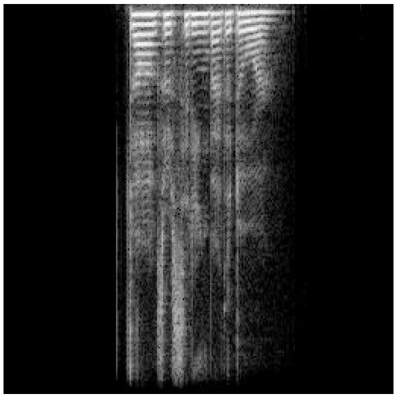
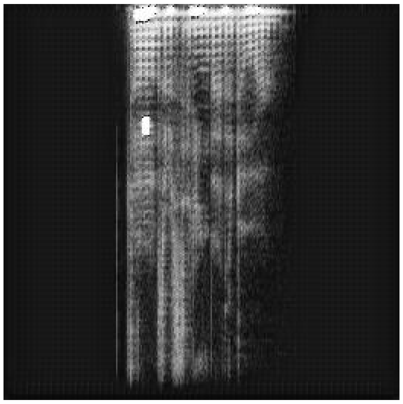
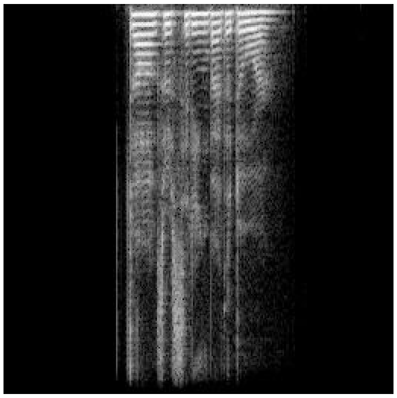
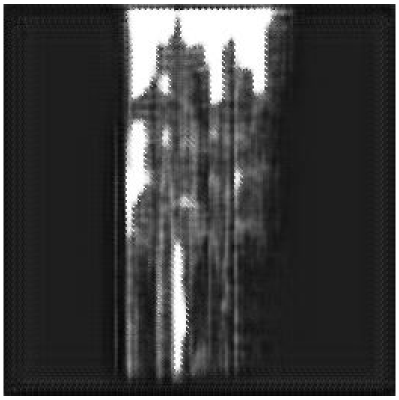
Number of ResNet Blocks	Original neutral Spectrogram	Generated Anger Spectrogram
3		
6		
9		

Table 1: Optimization of number of ResNet blocks in generator

We found that 3 ResNet blocks performed poorly without noticeable emotion transfer or perfect reconstruction of original audio. The model with 6 ResNet blocks performed better with satisfactory emotion transfer and reconstruction. Although 9 ResNet blocks gave very good results in terms of emotion transfer, the reconstructed audio suffered from noise. And this was computationally expensive too. Hence, we decided to proceed with 6 ResNet blocks which has a good compromise between style transfer, denoising and computational efficiency.

5.2.2 Testing optimal size of spectrogram

Table 2 shows a performance comparison between models that were trained to generate 260X260 spectrograms vs those that generate 520X520. These models were trained for a total of 100 epochs.

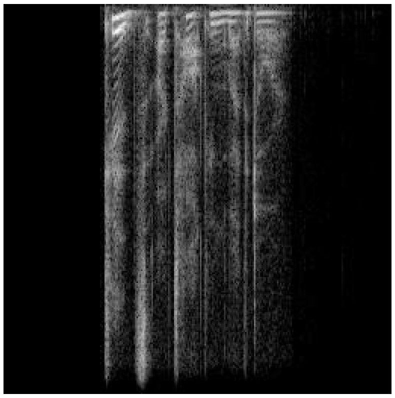
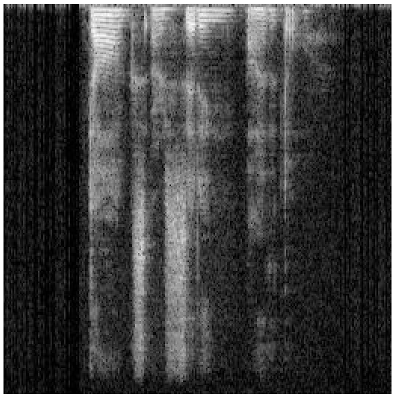
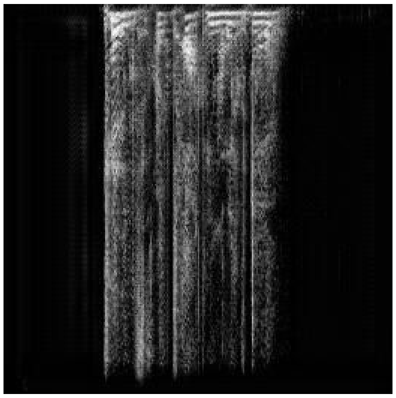
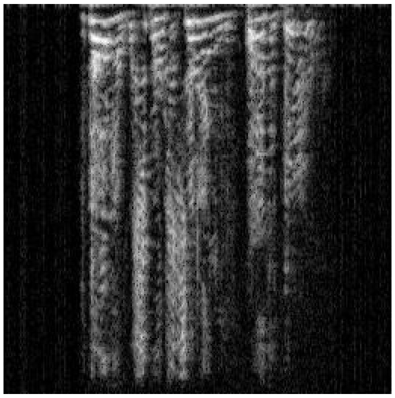
Emotion	“Dogs are sitting by the door” (260 X 260)”	“Dogs are sitting by the door” (520 X 520)”
Calm(original)		
Anger		

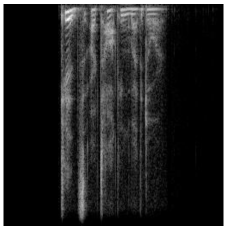
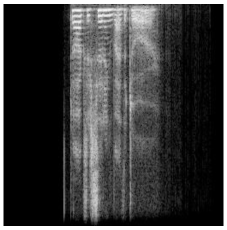
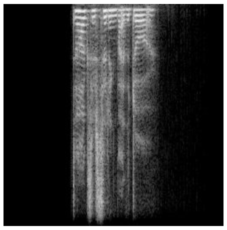
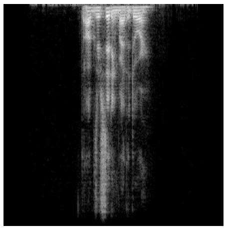
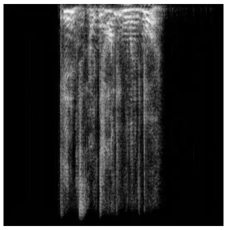
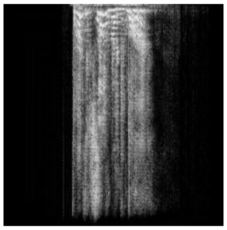
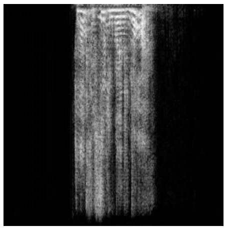
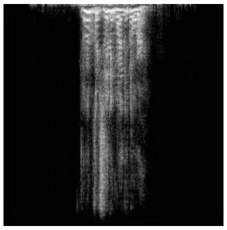
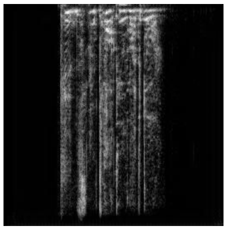
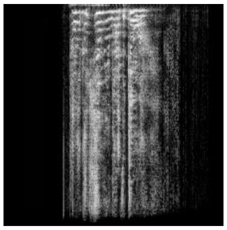
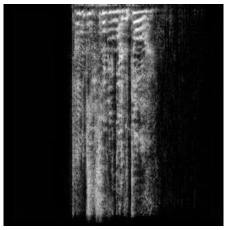
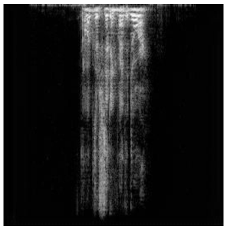
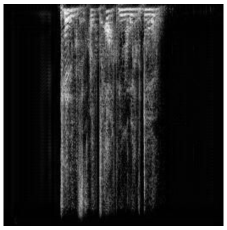
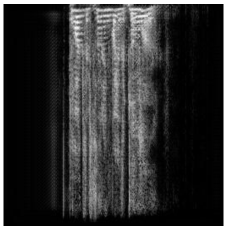
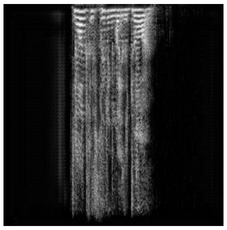
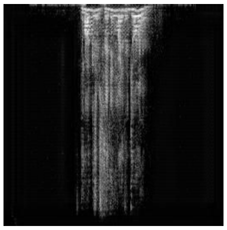
Table 2: Optimization of Spectrogram resolution

It is evident that the model which generated a 260X260 spectrogram had a better reconstruction compared to the other. This finding also helped us in reducing computation for further experiments.

5.2.3 Epoch Optimization

We trained models to perform emotion transfer from calm to other emotions given in the dataset, for 500 epochs on AWS p3.2xlarge instances. We noticed that after 150 epochs, the models' output started to mimic the input. This suggests a form of mode collapse [13]. Therefore, from checkpoints during runtime, for this dataset, we found that there was peak style transfer in the spectrograms at 150 epochs. Though it must be noted here that at the end of 500 epochs, the generated spectrograms and converted audio files had no noise.

Table 3 shows the conversion to the other emotions from calm, for different audio files.

Emotion	"Kids are talking by the door"	"Dogs are sitting by the door"	"Dogs are sitting by the door"	"Dogs are sitting by the door"
Calm (Original)				
Surprised (150 epochs)				
Fearful (150 epochs)				
Anger (150 epochs)				

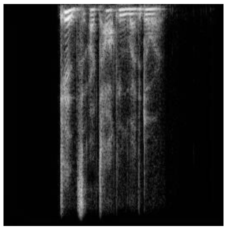
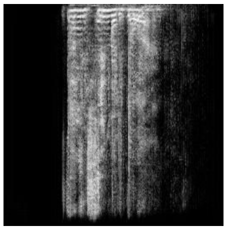
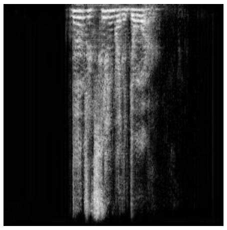
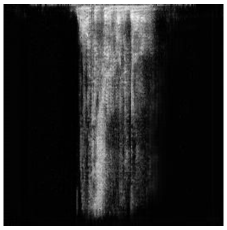
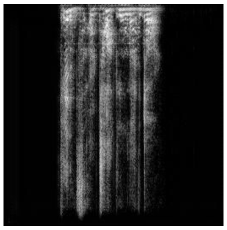
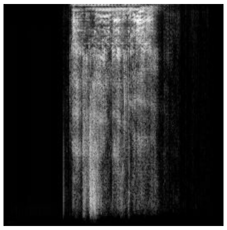
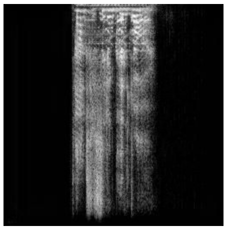
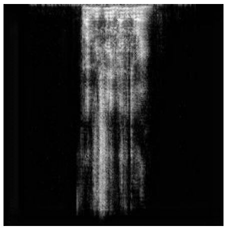
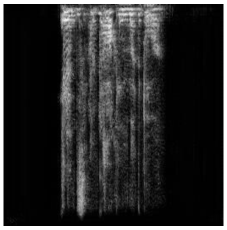
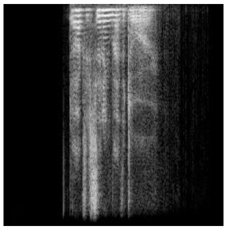
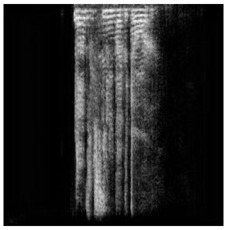
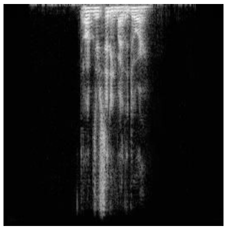
Disgust (150 epochs)				
Happy (150 epochs)				
Sad (150 epochs)				

Table 3: Performance of model trained to transfer emotion from calm to other emotions

From the above table we see two conversions, calm to fearful and calm to surprised, gives the best emotion transfer. Also there were noticeable characteristic changes in the harmonic structure of the input speech.

5.2.4 Unseen audio from same dataset

A few audio files from the dataset were held back for testing and optimizing our model. The spectrograms shown in table 4 are generated for these unseen input audio files.

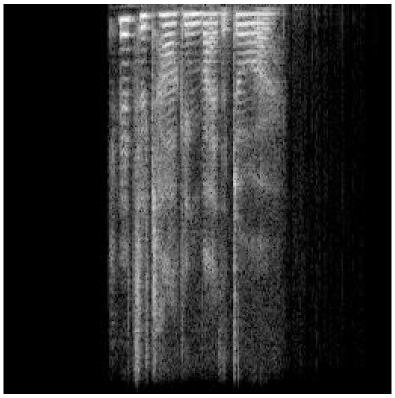
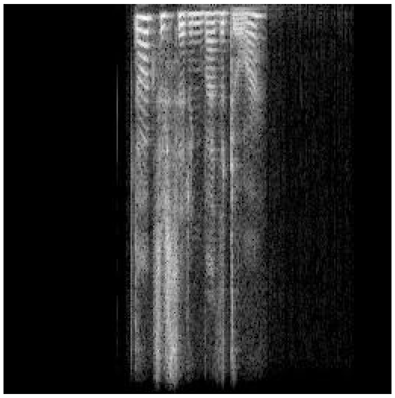
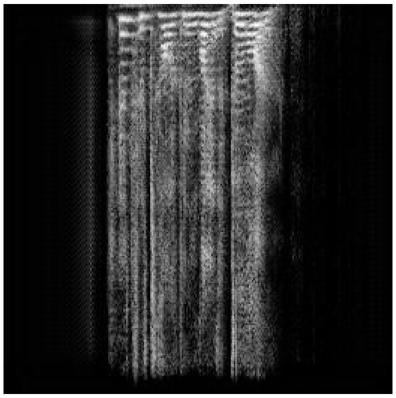
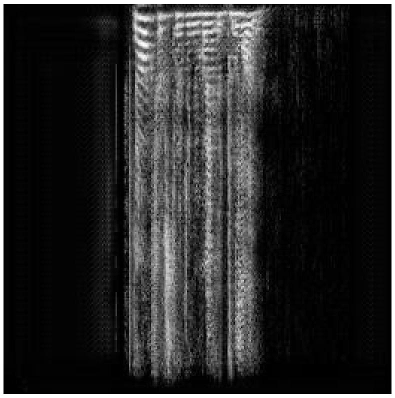
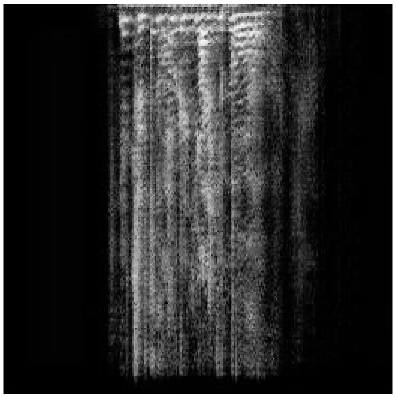
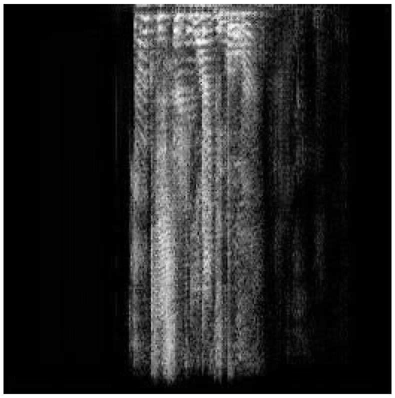
Emotion	“Kids are talking by the door”	“Dogs are sitting by the door”
Calm		
Anger		
Fearful		

Table 4: Testing models on unseen data

5.2.5 Same script by unseen actor

Table 5 below shows the performance of the model on an audio file of the same script, but by an actor not from the dataset. The model shows good performance even on unseen data.

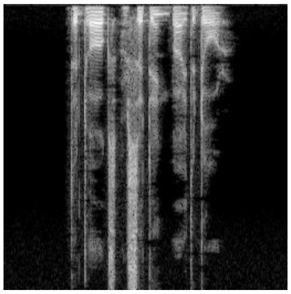
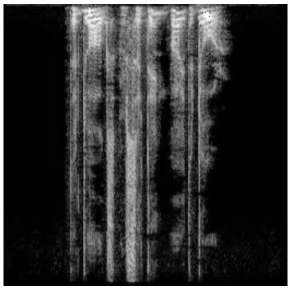
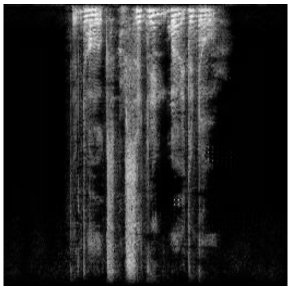
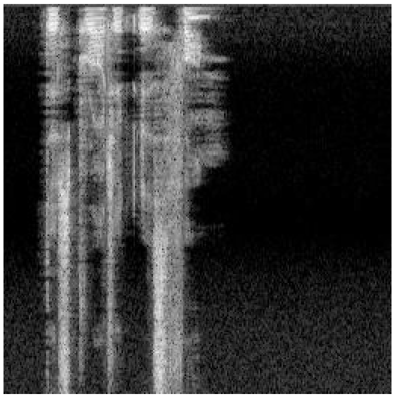
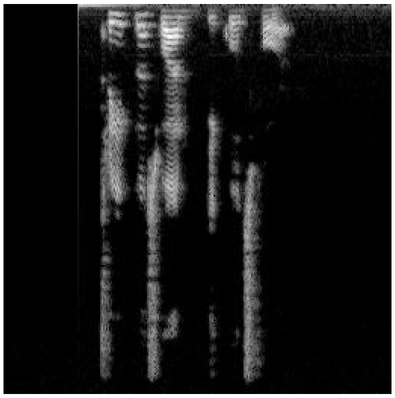
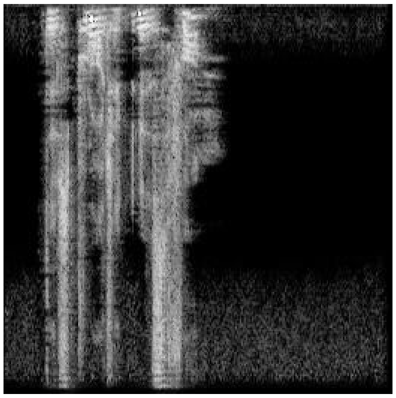
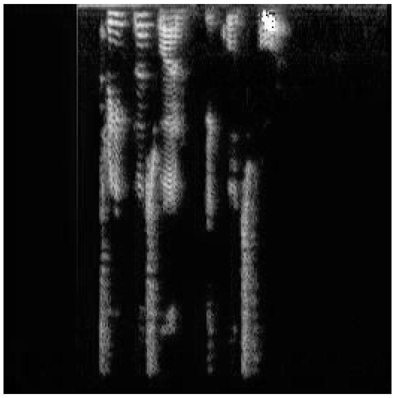
Script	Calm	Anger	Fearful
“The dogs are sitting by the door”			

Table 5: Testing model on unseen speaker

5.2.6 Lexically similar script by unseen actor

The following results in table 6 demonstrate the ability of our model to transfer emotions on audio clips of unseen actors speaking lexically similar sentences.

Emotion	“This project is fun”	“Three plus one equals four”
Calm		
Anger		

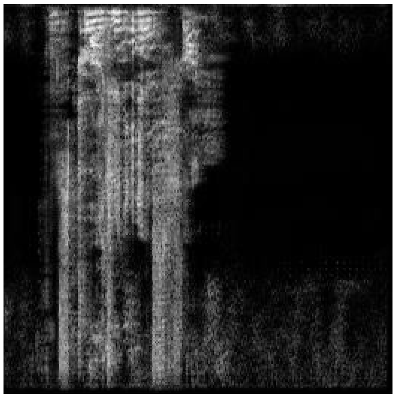
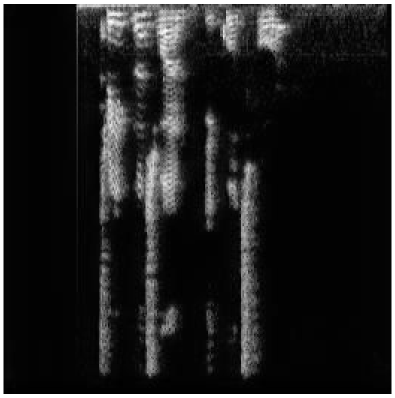
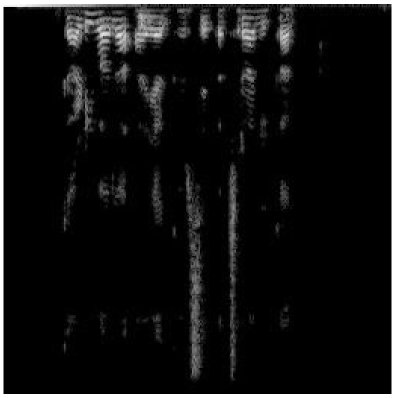
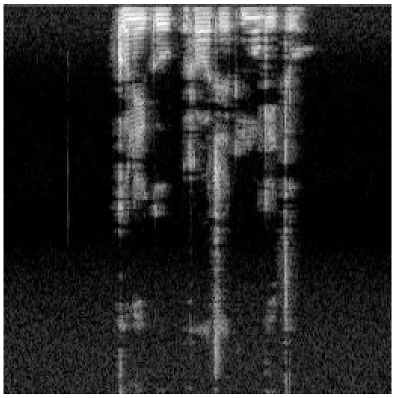
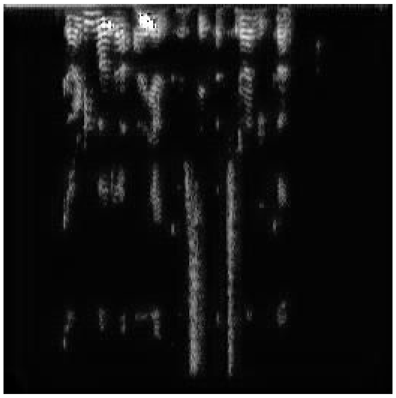
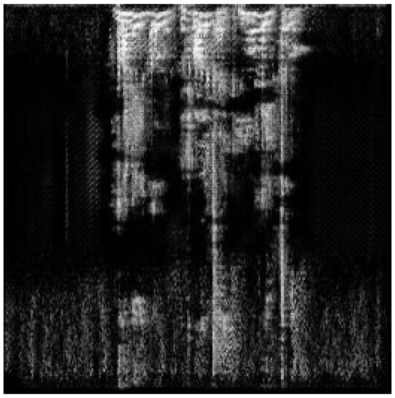
Fearful		

Table 6: Testing on lexically similar script by unseen actors

5.2.7 Different Language by unseen actor

We also experimented on audio clips of unseen actors speaking in a different language (Hindi and Kannada). The model did not produce results with sufficient style transfer. However, the model was still able to reconstruct the audio clip of an unseen language without much noise.

Emotion	“Gaadi waala aya ghar se kachra nikal”	“Konegu project mugithu”
Calm		
Anger		

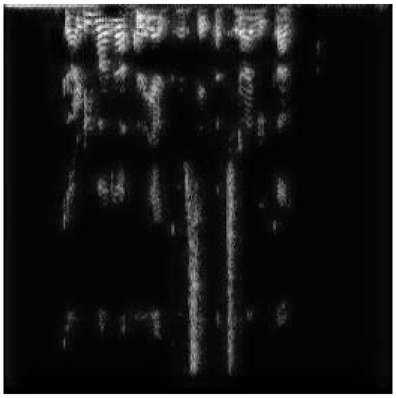
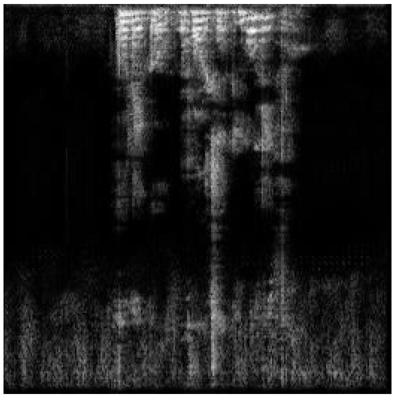
Fearful		

Table 7: Testing on unseen language

5.3 Computation resources

For these experiments we recorded the following train times for different model configurations on Google CoLab:

- 260X260 image resolution with 3 resnet transformer blocks (~20.8M) - about 8 Hours for 250 epochs (50,000 steps)
- 260X260 image resolution with 6 resnet transformer blocks (~43.8M) - about 12 Hours for 250 epochs (50,000 steps)
- 260X260 image resolution with 9 resnet transformer blocks (~77.4M) - about 12 Hours for 120 epochs (24,000 steps)

For these experiments we recorded the following train times for different model configurations on AWS p3.2xlarge:

- 520X520 image resolution with 6 resnet transformer blocks (~43.8M) - about 12 Hours for 120 epochs (24,000 steps)
- 260X260 image resolution with 6 resnet transformer blocks (~43.8M) - about 14 Hours for 500 epochs (100,000 steps)

6. Future Work

A step towards improving the performance of the model to make it generalise better would be to use a dataset which isn’t lexically similar, like the IEMOCAP dataset [14]. Some of the reconstructed audio files suffer from noise. Denoising methods can be implemented to reduce this effect. This GAN model suffers from mode collapse after 150-200 epochs. This can be reduced by using a better regularization or loss function.

7. Conclusion

We have implemented a Cycle GAN to transfer various emotions in speech using spectrograms. After experimentation on various model configurations, we found out that the optimal model consists of six ResNet blocks in the generator model for an input of image size 260x260. Satisfactory emotion transfer and audio reconstruction on unseen inputs shows that this method can be used to train a versatile Cycle GAN on lexically limited data content.

References

1. https://www.youtube.com/watch?v=9N_uOIPghuo
2. Jun-Yan et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (2017) https://arxiv.org/abs/1703.10593
3. https://en.wikipedia.org/wiki/Spectrogram
4. D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. ASSP, vol.32, no.2, pp.236–243, Apr. 1984. https://ieeexplore.ieee.org/document/1164317
5. Goodfellow, Ian J. et al. “Generative Adversarial Networks.” *ArXiv* abs/1406.2661 (2014): n. pag. https://arxiv.org/abs/1406.2661
6. Felix Burkhardt and Nick Campbell, "Emotional Speech Synthesis https://people.ict.usc.edu/~gratch/CSCI534/Readings/ACII-Handbook-SpeechSyn.pdf
7. Gao, Jian et al. “Nonparallel Emotional Speech Conversion.” Interspeech 2019 (2019): n. pag. Crossref. Web. https://arxiv.org/abs/1811.01174
8. [Satt et al. (2017) "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms https://pdfs.semanticscholar.org/de47/fc09bc8dcd032c8b3450a0b2a816c376e07e.pdf
9. Wyse et al. "Audio Spectrogram Representations for Processing with Convolutional Neural Networks" (2017) https://arxiv.org/abs/1706.09559
10. Oyamada et al. "Generative adversarial network-based approach to signal reconstruction from magnitude spectrograms" (2018) https://arxiv.org/abs/1804.02181
11. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391
12. Challenges in Representation Learning: Facial Expression Recognition Challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
13. Arora, Sanjeev and Yi Zhang. “Do GANs actually learn the distribution? An empirical study.” *ArXiv* abs/1706.08224 (2017): n. Pag. https://arxiv.org/abs/1706.08224
14. C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008. https://sail.usc.edu/iemocap/index.html