Project Summary:

My Project was exploration, reproducing research papers and training models related to Text to Speech model using few minutes of the speaker's voice, Voice conversion model using few minutes of the speaker's voice and Video Generation using GANS, Audio Generation using GANS.
The Primary focus was to train a TTS model or a VC Model using few minutes of the speaker particularly for the Indian Accent.

Week (1-2) :
Worked on Video generation using GANS (Predict missing frames during a call):
Read about GANs and their applications
Critically read and compiled a list of research papers :

FutureGAN
Stochastic Adversarial Video Prediction (savp)
Generating Videos with Scene Dynamics
Improving Video generation for multi-functional applications:
MoCoGAN: Decomposing Motion and Content for Video Generation
Probabilistic Video Generation using Holistic Attribute Control

Github Repositories for all the papers are available

From here to there: Video in between using direct 3d convolutions.
Predicting Future Frames using Retrospective Cycle GAN
Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks
Deep Video Generation, Prediction and Completion of Human Action Sequences

Generating the Future with Adversarial Transformers

Github Repositories are not available for these papers but does address our problem

Week 3:

Worked on Audio generation using Gans :

Critically read and compiled a list of research paper and researched about the feasibility of these models in real time :

1) Audio inpainting with generative adversarial network
2) Adversarial generation of time-frequency features with application in audio synthesis
3) Adversarial audio synthesis
4) WaveRNN (Efficient Neural Audio Synthesis)
5) SEGAN: speech enhancement generative adversarial network
6) A context encoder for audio inpainting
7) Audio super-resolution using neural nets

Github Repositories for all the papers are available

Speech Loss Compensation by Generative Adversarial Networks
ConcealNet: An End-to-end Neural Network for Packet Loss Concealment in Deep Speech Emotion Recognition

Github Repositories are not available for these papers but does address our problem

Week 4:

Worked on TTS Models using few minutes of data for unseen speakers

Critically read and compiled a list of research papers :

1) Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis
2) Neural Voice Cloning with a few samples
3) Voice Loop
4) Flowtron
5) Tacotron 2 + Waveglow
6) Deep Voice 3
7) Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention

Github Repositories for all the papers are available

Week 5:

Worked on reproducing the results of TTS Models

Reproduced the Results of the following papers:

- Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis
  Result : Could Clone the Voice of Former President Obama with 5 mins of data, works well for american accent
- Speaker Adaptation using Deep Voice 3
  Result: Trained the pretrained model for 2000 Steps, the results were satisfactory
- Tacotron 2 using Waveglow
  Result: Satisfactory results for a single speaker on whom the model was trained on

Week 6:

- Worked on Voice Conversion Models using few minutes of data for unseen speakers

Critically read and compiled a list of research papers:

1) One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization
2) Zero-Shot Voice Style Transfer with Only Autoencoder Loss
3) Star-Gan Voice Conversion
4) GAN-based text-to-speech synthesis and voice conversion (VC)
5) Voice Conversion Based on Cross-Domain Features Using Variational Autoencoders

Github Repositories for all the papers are available

Reproduced the Results of the Following Repositories:

- One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization
  Result: Moderate Results for Voice Conversion to Unseen Speakers
- Voice Conversion using Star-GAN
  Result: Good Results for Voice Conversion to Seen Speakers
- Zero-Shot Voice Style Transfer with Only Autoencoder Loss
  Result: Moderate Results for Voice Conversion to Unseen Speakers

Week 7:

- Preprocessing Indian Languages Dataset

Researched and compiled a list of Indian Datasets:
http://www.openslr.org/resources.php
https://www.iitm.ac.in/donlab/tts
Drawbacks: (Absence of speakers who speak English in Indian Accent)

- Preprocessing script to load the Indian datasets and resample their sampling rate and generate their mel-spectrograms

```
import os
from shutil import copy
import librosa

directory_1 = "/Users/jay/Documents/Real-Time-Voice-Cloning/data/ta_in_male/"

for test in os.listdir(directory_1):
    sr = 16000
    if ".DS_Store" not in test and "LICENSE" not in test and "line_index.tsv" not in test:
        path = directory_1 + test
        print(path)
        y, s = librosa.load(path, sr=16000)
        librosa.output.write_wav(path, y, sr)
my_set = set()


for files in os.listdir(directory_1):
        d = files[4:9]
        my_set.add(d)
print(my_set)


for item in my_set:
    if item != "" :
        os.mkdir(item)

for file in os.listdir(directory_1):
    d = file[4:9]
    path_1 = directory_1 + file
    if len(d) == 5:
        copy(path_1, d)



directory = "SV2TTS_Indian/encoder"
directory_1 = "SV2TTS_Indian/encoder/"

for file in os.listdir(directory):
    if "DS_Store" not in file:
        count = 0
        path = directory_1 + file
        for files in os.listdir(path):
            if "_sources.txt" not in files:
                # print(files)
                count = count-1
        if count<-5:
            print(file)
```

Week 8:

The following papers can perform voice cloning for unseen speakers:

Text to Speech Model:

1. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis
2. Neural Voice Cloning with a few samples

Voice Conversion Model:

1. One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization

Training of the following model:

(Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis)

Training the model for Indian Datasets

Trained the Pretrained Model Speaker Encoder Model that was trained from 1.56M Steps to 1.575M Steps using the following datasets (100+ Speakers in Gujurati, Kannada, Marathi, Tamil and Malayalam:

http://www.openslr.org/63/ http://www.openslr.org/64/ http://www.openslr.org/65/ http://www.openslr.org/66/ http://www.openslr.org/78/ http://www.openslr.org/79/

The Model for the Mandarin dataset : https://github.com/KuangDD/zhrtvc

The Model for Swedish dataset:

https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/400

Results: The Model works well for voices that have American Voices. For Indian Voices, the results were better than the initial results, the accent was different than the initially robotic voice and the american accent was reduced, but the accent was still not Indian because of the unavailability of Indian accent speakers. The voice sounded more natural and smooth.

Further Work:
By training the Speaker Encoder, Synthesizer and Vocoder on a dataset composed of multiple speakers audio and transcripts in Indian accent would bring in results that can clone unseen Indian voices very well.