Speech Driven Video Generation from Facial Images

Kavita Jain[#1], Aloysius Alvares[*2], Oliver Andrades[*3], Ryan Dsouza[*4], Pratap Solanki[*5]

[#1]*Professor, Xavier Institute of Engineering, Mumbai*
[*]*Student, Xavier Institute of Engineering, Mumbai*

[1] kavita.j@xavierengg.com
[2] aloysiusreddragon@gmail.com
[3] oliver.andrades@live.com
[4] dsouzryan74@gmail.com
[5] solanki99p@gmail.com

***Abstract* - Generally, when a video is dubbed in other language the lip movement does not match the audio sequence. The problem of generating realistic talking heads is multifaceted, requiring high-quality faces, lip movements synchronized with the audio, and plausible facial expressions. We present a method for generating real time faces with synchronized lip movement as per the audio segment and static image provided by the user. Our focus is on capturing the facial coordinates in a frame and generate a new frame consisting of lip movement synced with desired audio. We implement separate frame and sequence discriminators to provide adversarial feedback to the generator based on correctness of face and synchronicity with the input audio. This multi-discriminator approach to GAN generates more realistic looking results.**

***Keywords*— Neural network, Machine Learning, CNN, RNN, Modal transformation**

I. INTRODUCTION

Facial animation in movies and animation is achieved manually or through the use of extremely resource intensive animation software. Manually animating the facial features requires manipulation at a frame-by-frame level and is tediously time consuming, on the other hand using animation software requires large numbers of GPUs usually on a render farm. Both approaches are extremely costly and inaccessible to common masses and small-scale businesses. Using machine learning to do this task can reduce the computational power and time required by more than an order of magnitude.

Currently, there exist multiple machine learning models that can do the task of lip animation (audio and image to video). The speech segment need not be spoken originally by the target person. Our method differs from previous approaches we learn the correspondences between raw audio and video data directly. By focusing on the speech portion of audio and tight facial regions of speakers in videos, our model is able to produce natural-looking videos of a talking face at test time even when using an image and audio outside of the training dataset.

The key idea of the approach is to learn a joint embedding of the target face and speech segment that can be used to generate a frame of that face saying the speech segment. Thus, the inputs are still images of the face and the target speech segment; and the generated output is the target face speaking the segment.

II. RELATED WORKS

Currently there are various works that proposed methods to synthesize videos of talking heads from either audio or text sources. **Lip reading in the wild [7]** focuses to recognize the words being spoken by a talking face, given only the video but not the audio. It makes two novel contributions: First, it develops a pipeline for fully automated large-scale data collection from TV broadcasts. Second, it develops CNN architectures that are able to effectively learn and recognize hundreds of words from this large-scale dataset.

Convolutional Neural Networks (CNN) are very similar to ordinary Neural Networks (NN): they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other. And they have a loss function (e.g. SVM/Softmax) on the last (fully-connected) layer. CNN allows us to process images of various sizes as well as reducing the amount of computation by reusing a small number of weights over the entire image.

Similarly, **Speech-Driven Facial Re-enactment Using Conditional GANs[3]** proposed using a recurrent neural network, they achieved movements of mouth landmarks based on audio features then exploit the power of conditional generative adversarial networks to produce highly-realistic faces conditioned on a set of landmarks.

Generative Adversarial Networks (GANs) are one of the most active areas in deep learning research and development due to their incredible ability to generate synthetic results. At its core, a GAN includes two agents with competing objectives that work through opposing goals. This relatively simple setup results in both of the agent's coming up with increasingly complex ways to deceive each other. This kind of situation can be modeled in Game Theory as a minimax game. Generative Adversarial Networks take advantage of Adversarial Processes to train two Neural Networks who compete with each other until a desirable equilibrium is reached.

The work in **Talking Face Generation by Conditional Recurrent Adversarial Network [4]** attempted to generate the talking face video with accurate lip synchronization while maintaining smooth transition of both lip and facial movement over the entire video clip. They deployed a multi-task adversarial training scheme in the context of video generation to improve both photo-realism and the accuracy for lip synchronization.

Recurrent neural networks (RNN) are a type of neural network where the outputs from previous time steps are fed as input to the current time step. This creates a network graph or circuit diagram with cycles, which can make it difficult to understand how information moves through the network. In order to handle sequential data successfully, we need to use recurrent (feedback) neural network. It is able to 'memorize' parts of the inputs and use them to make accurate predictions.

It is important to notice that we can unroll an RNN network as many times as elements in the input sequence. Furthermore, the parameters of each "realization" of the RNN cell are the same, making the number of parameters of the model independent of the length of the sequence. Essential to the success of a model like this is the operations we performed inside the RNN unit. This allows us to use variable length sequences.

III. PROPOSED MODEL

The proposed architecture is given in Figure 4. We describe and explain below, how the modules (audio encoder, identity encoder, noise generator, image decoder, frame discriminator and sequence discriminator) are designed and trained together.

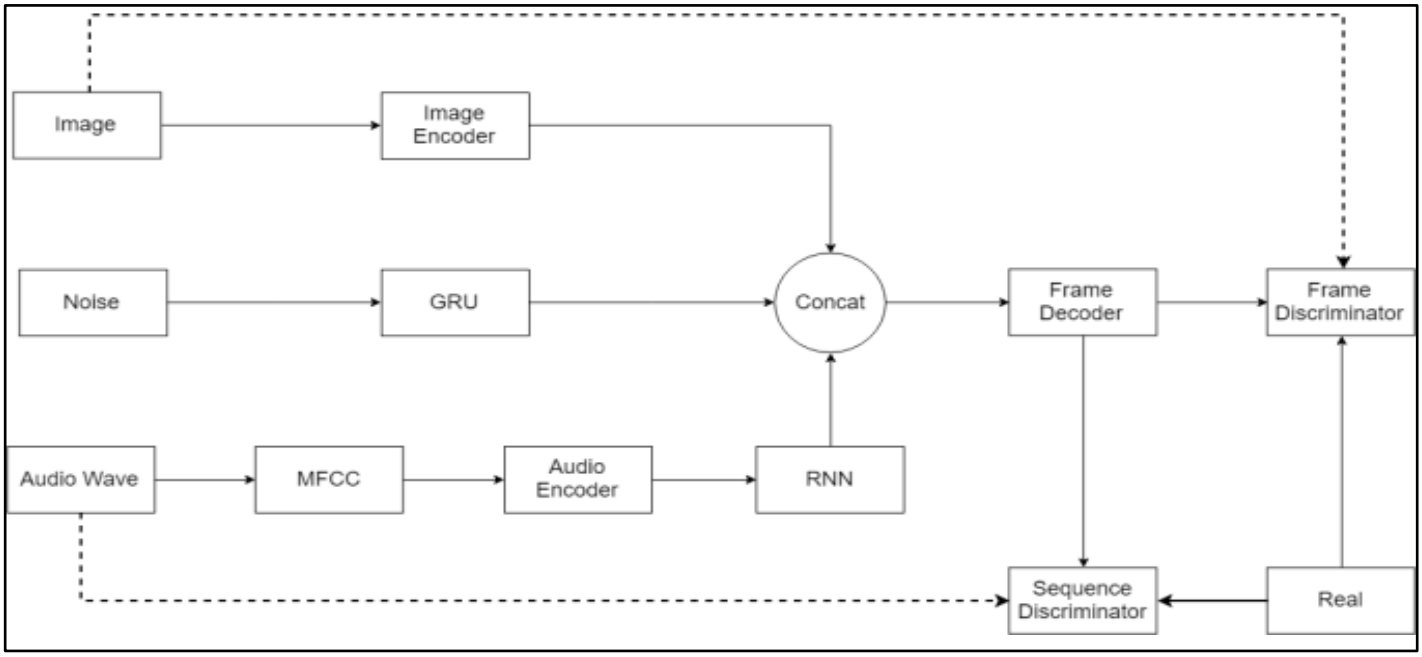

Fig. 5 – Proposed Model

1. Audio encoder: MFCC is used to encode key features of Human speech. It is typically used in speech recognition pipelines. We use a 7-layer neural network that applies 5 layers of convolution, and 2 Fully-Connected layers to produce an instantaneous encoding of the MFCC input. We used a 2 layer Gated Recurrent Unit (GRU) to produce an encoding such that it encodes time varying information obtained from the previous 7-layer network. This is to be used as one of the inputs to the decoder. The CNN is based on AlexNet [6] and VGG-M [8], and is similar to the audio encoder mentioned in You Said That? [1].

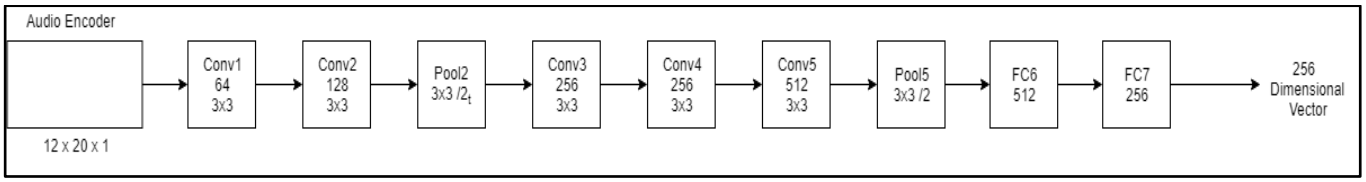

Fig. 1 – Audio Encoder

2. Identity encoder: The identity encoder is a 6-layer CNN which takes a 3x128x96 dimensional input image and outputs a 50-dimensional encoding vector. It is part of a U-net [5] architecture, such that the output of the first 4 layers is passed to the decoder via Skip-connections. This is used to preserve identity information and to improve segmentation accuracy. The identity encoder is the same as used in [2].

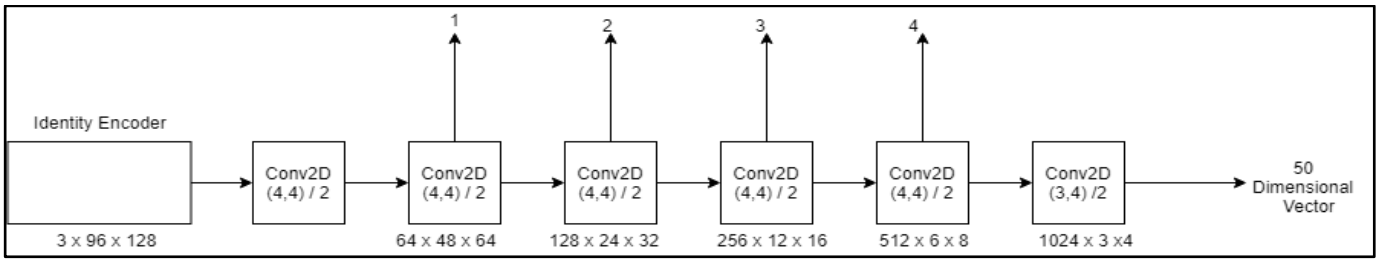

Fig. 2 – Identity Encoder

3. Noise Generator: We use a noise generator to produce sequentially coherent noise. The purpose of this is to provide smoothness to the facial transitions. The noise generator is a Recurrent Network that utilizes a 2-layer GRU that produces a temporarily coherent 10-dimensional noise vector. It takes as input a 10-dimensional vector of mean 0, and standard deviation of 0.77.

4. Image Decoder: Image decoder is the latter half of the U-net based generator, which takes in the skip connection passed by the Identity encoder. It is used to generate the output image by concatenating the output of Audio encoder, Identity encoder and Noise generator. The decoder consists of 6 Transposed Convolutional layers and 4 Convolution layers arranged in an alternating fashion. The skip connections passed from the identity encoder are concatenated to the output of the previous transpose layers before every convolution layer, as suggested in [5].

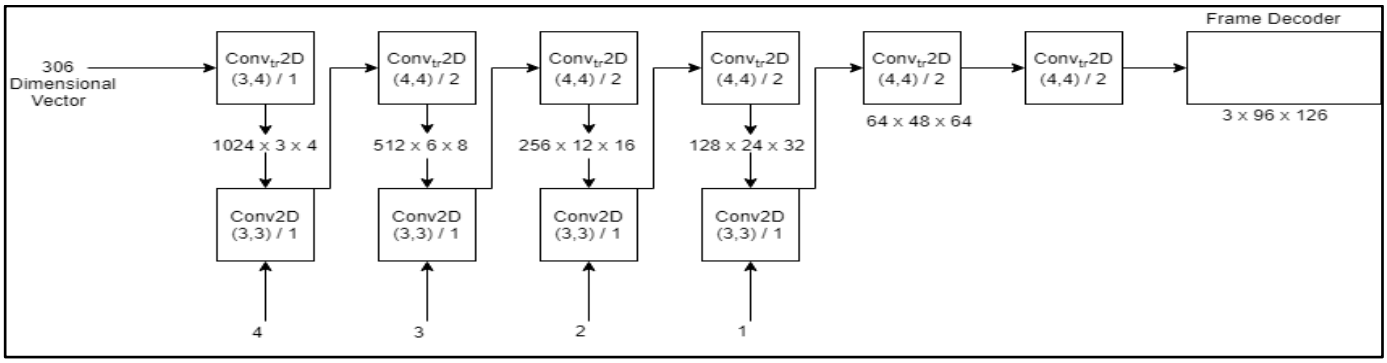

Fig. 3 – Frame Decoder

5. Frame Discriminator: It is a 6-layer CNN that takes the generated or real image as input and the supplied static image as conditional input. It is based on the frame discriminator used in "Attribute Augmented Convolutional Neural Network for Face Hallucination" [6]. It is used to compare the identity of the generated image against the input image. We concatenate the generated image with the reference image to pass it through 4 Convolutional and Pooling layers and 2 Fully Connected layers to get a single binary output representing the measure of similarity between the conditional and generated faces.

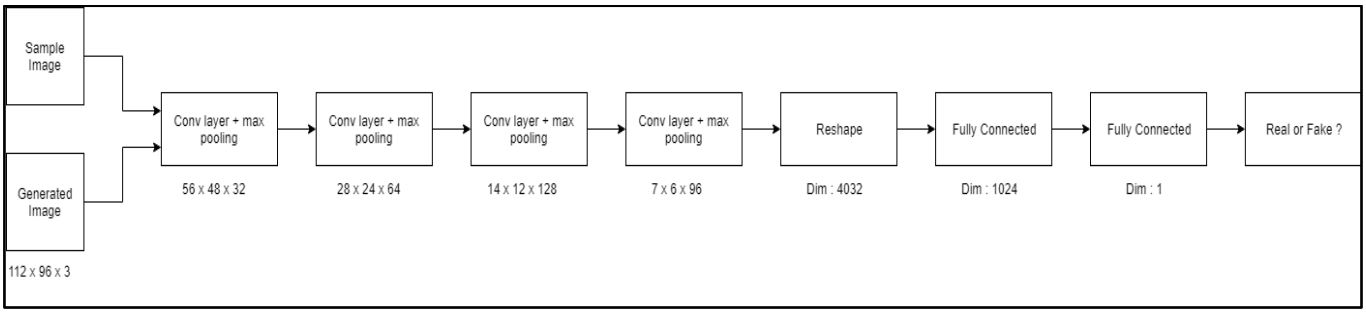

Fig. 4 - Frame Discriminator

6. Sequence Discriminator: The sequence discriminator in [2] uses a recurrent network over the entire video sequence and passes that encoding through a 2-layer fully connected network to get a binary truth value. Since fully connected layers require a fixed size input, their network only works on fixed length sequences. We solve this problem by using a customized version of Spatial Pyramid Pooling [9]. Our customized version works on 1-Dimensional vector. This allows us to obtain a fixed sized output. This module takes the generated or real image as input and the given audio as a condition. It verifies that the output image matches the audio input.

IV. TRAINING

To train the neural network, we used the vxceleb2 dataset of videos, which are cropped to the speakers face. The dataset consists of videos of about 6000 people of various ethnicities and a balance gender split. The Frame discriminator (Dimg) is trained on frames that are sampled uniformly from a video x using a sampling function S(x). The Sequence discriminator (Dseq), classifies based on the entire sequence x and audio a. The loss of our GAN is an aggregate of the losses associated with each discriminator. The modules in generator uses a learning rate of 2e-3 and the modules in discriminator uses a learning rate of 2e-5.

V. CONCLUSION

We have demonstrated that our model is able to generate videos of any identity speaking from any source of input audio. This work shows that there is promise in generating video data straight from an audio source. We have also shown that re-dubbing videos from a different audio source (independent of the original speaker) is possible. One clear extension is to add a quantitative performance measure of our models. This is not a straightforward task as there is no definitive performance measure of generative models for a specific domain. One possible option is to have a lip-specific inception score using networks trained on a lip-specific task [3]. Moving forward, this model can be applied to computer facial animation relying only on audio.

REFERENCES

[1] You said that? Chung, Jamaludin and Zisserman.
[2] End-To-End: Speech-Driven Facial Animation with Temporal GANs: Pantic, Petridis, Vougioukas.
[3] Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks: Aghajan, Hasani, Jalalifar.
[4] Talking Face Generation by Conditional Recurrent Adversarial network: Qi, Song, Wangand Zhu.
[5] U-net: Convolutional Networks for Biomedical Image Segmentation: Brox, Fischer and Ronneberger.
[6] Attribute Augmented Convolutional Neural Network for Face Hallucination: Cheng-Han Lee, Kaipeng Zhang, Hu-Cheng Lee, Chia-Wen Cheng, Winston Hsu.
[7] Lip Reading in the Wild: Joon Son Chung and Andrew Zisserman.
[8] Return of the devil in the details: Delving deep into convolutional nets: Chatfield K., Simonyan K., Vedaldi A., Zisserman A.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Spatial Pyramid Pooling in Deep ConvolutionalNetworks for Visual Recognitio