CS 753: PROJECT REPORT

Facial Emotion Synthesis from Speech

Vamsi Krishna Reddy	Yaswanth Kumar	Vighnesh Reddy
160050064	160050066	160050090

1 Introduction

We first define the goal of our project which involves the following inputs.

1. A person's face as an image - called *Input A*
2. A Speech that has an emotion (emotion is not known) - called *Input B*

The output of our model is to generate the person's face with the emotion detected from the speech. We achieve this by using a cascaded architecture by generating Action Units to encode emotion as shown in Figure 1.

Literature exists for detecting emotion from speech, and viewing it as a classification task is a well known area of research. We however try to regress the Action Units (AU) corresponding to the face, which gives a more fine-grained control over emotion. This is required since we want to generate a face of an independent person with the same expression as in the voice (audio). Generative models have been extensively used in literature for generating images (in this case - faces) conditioned under a latent vector (like emotion / AU here).

1.1 Action Units

Figure 1: Effect of action units on facial features. (Source [1])

Facial Action Coding System (FACS) is a system to encode human facial movements by their appearance on the face. Movements of individual facial muscles are encoded by

Emotion	Action Units
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+16

Table 1: Dependence of various emotions over action units (Source: Wikipdia)

FACS from slight different instant changes in facial appearance. Using FACS it is possible to code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) that produced the expression. It is a common standard to objectively describe facial expressions.

Table 1 shows the dependence of emotions over action units. We ideally expect our model to trigger these action units when given an audio of a particular emotion. We encode the facial expression using Action Units corresponding to facial features. We utilize 17 AUs in our implementation *i.e.* AU - 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 45. This enables to get a fine-grained control over emotion that we generate.

2 Architecture Details

2.1 Cascaded Model

We now describe the inference mechanism here with the trained Model A, B available to us. The training and inference mechanism of the submodules is described later the section. The inference on the model involves the following steps:

1. Forward the speech input (*Input B*) to Model B to get the AUs corresponding to that speech's emotion.
2. Then get the desired output of an image with the face corresponding to that emotion by sending the image (*Input A*) and the AUs from above as inputs to Model A.

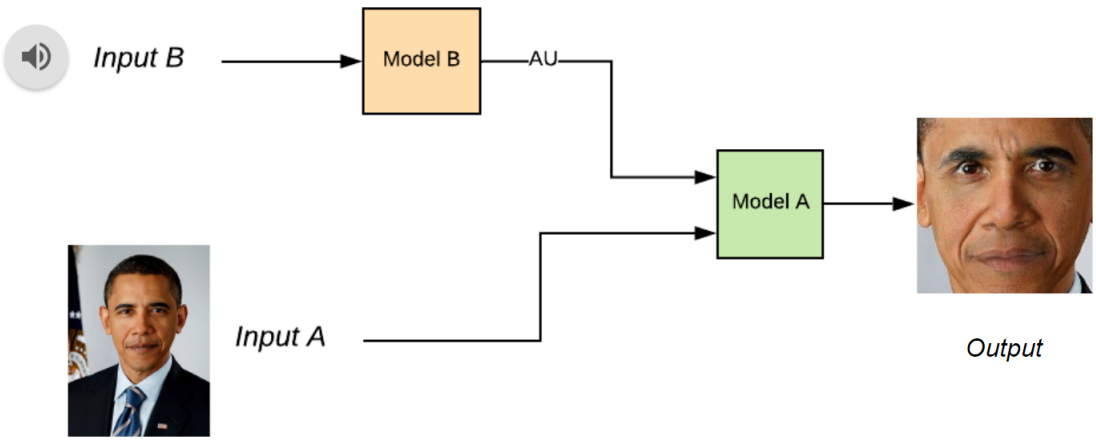

Figure 2: Inference on our cascaded model

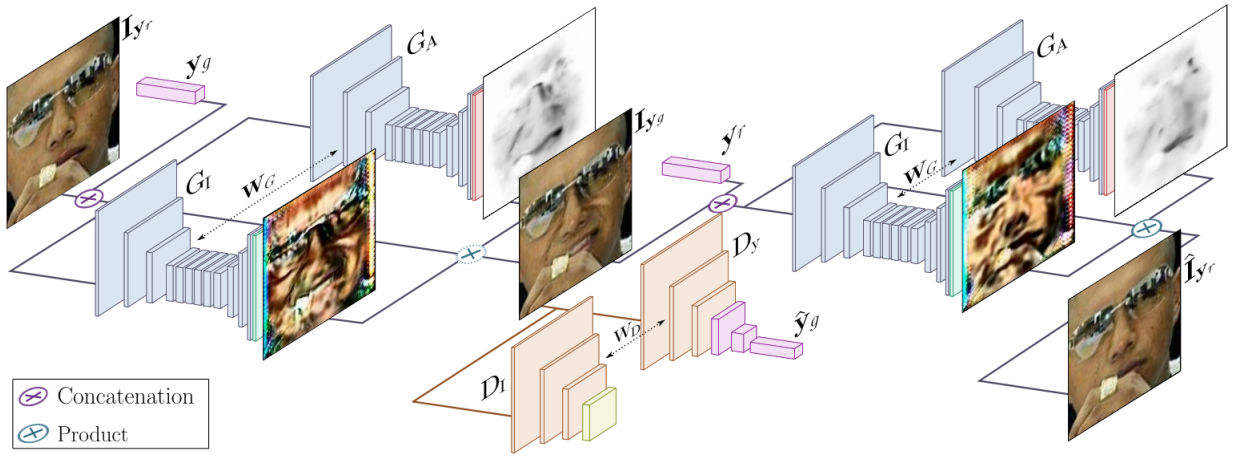

Figure 3: GANimation Architecture from Pumarola *et al.* [2]

2.2 Model A

This model is essentially a GAN. The goal of the generator network of this GAN is to generate facial images similar to input image conditioned on AUs from Model B. We employ the architecture presented in [2] for this task which they call GANimation as shown in Figure 3. We now describe the training and inference mechanism in the following subsections.

2.2.1 Training

The generator consists of a conv-deconv architecture backbone with deconv heads to generate two images. One denotes the features to be changed according to the AU vector given as input over the input image. Other is an attention mask which denotes where to change the pixel corresponding to that location for that emotion. This is needed to ensure that the generated images actually look similar to the original face image.

The discriminator then has a conv backbone layer and two heads, one for a PatchGAN and other to discriminate against the original AU. The PatchGAN gives us a 32×32 matrix, which discriminates whether each patch of image is fake or real.

Also, similar to CycleGAN [3] there is a cycle reconstruction loss for the generator. Another smoothing loss exists to ensure that the attention mask generated from the generator is smooth, *i.e.* doesn't contain any abrupt changes in pixel intensities (which denotes the importance of the pixel). All these losses together generate gradients which are back-propagated through the network, with the optimizers updating the parameters.

2.2.2 Inference

During inference, only the generator is used to generate the desired images and the discriminator is discarded. The image and AUs are given as input to the generator, which then generate the face with the new emotion.

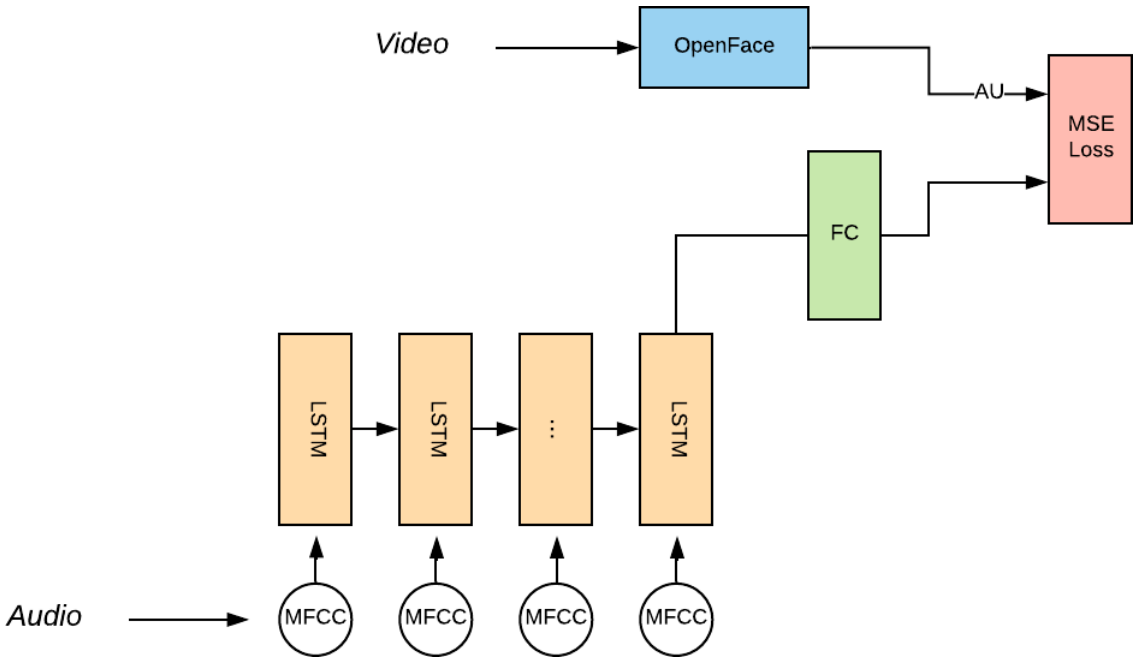

Figure 4: Architecture of Model B

2.3 Model B

The goal of Model B is to regress the AUs given an MFCC features extracted from audio as input. The architecture of this model is shown in Figure 4. It is an LSTM [4] with a fully connected layer at the end as shown in Figure 3. The LSTM encodes the MFCC feature vectors of video frames into single vector which is fed to fully connected layer at the end. The ground truth labels are the mean of AUs of video frames extracted using the Openface toolkit [1]. To train this model, we simply consider a regression loss like MSE error as our objective function and minimize it.

3 Datasets

We trained Model A on 75,000 images from Emotion-Net dataset [5]. The resized 128×128 input image face (obtained using `face_recognition` package) along with face AUs obtained from Openface is fed to Model A as input. The original dataset is much larger, but owing to the computational resources available, we restrict overselves to this limited dataset trained for about 24 hrs. The results don't seem to have affected due to this simplification.

We trained Model B on 1440 videos from RAVDESS dataset [6] which contains video clips of actors uttering phrases with a particular speech. The input to Model B are the 39-sized MFCC vectors. Hence, the MFCC vectors corresponding to audio compoment of videos is extracted using `torchaudio` package from PyTorch. It is trained under MSE Loss over the averaged AUs obtained for each frame using Openface.

Angry - low	Angry - mid	Angry - high
Happy - low	Happy - mid	Happy - high
Surprise - low	Surprise - mid	Surprise - high

Figure 5: Outputs of the model for various strengths of emotions. Left image is input A. Every audio *i.e.* input B is hyperlinked to the text. The corresponding output images are shown here. The text denotes the audio emotion and strength based on human interpretation.

4 Results

We use the RAVDESS dataset [6] to train and evaluate our model. Since the task is generative, a precise evaluation metric for the whole task is difficult to define over. Hence, the evaluation of the model is manual (as is the case for many tasks handled by generative models). Figure 5 shows the variation in our model's output with changing emotions and strengths. As a sanity check, we also note that varying the various AUs components to 1 and visualizing the images gives interpretable results. Thus, indicating that our Model A is sufficiently trained and working well as shown in Figure 6.

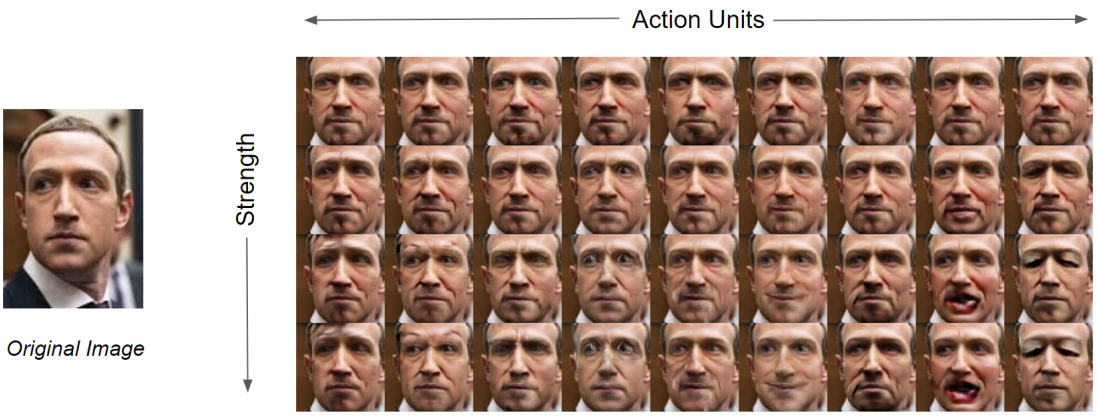

Figure 6: Architecture of Model B

5 Conclusion

We tackle the task of emotion recognition from speech as a regression task instead of a more common view as a classification task, thus having an advantage to encode various degrees and intensities of similar emotions. We note that the regressed model though is able to capture the emotion, does not precisely do well in encoding the level of emotion as expected by doing regression instead of classification. This may possibly be due to insufficient amount of data corresponding to the diversity in the training samples of RAVDESS [6] dataset. In general, our model is able to capture the emotion and embed it into a face given as input. Exploring more architectures for Model A or training over a larger dataset may lead to improved results.

References

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[2] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[6] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, 2018.