CS 753: Course Project

Final Presentation

Facial Emotion Synthesis from Speech

Vamsi Krishna Reddy | Vighnesh Reddy | Yaswanth Kumar

160050064 160050090 160050066

Problem Statement

- Input:
  - A person's face as an image - called *Input A*
  - Speech that has an emotion (emotion is not known) - called *Input B*
- Output:
  - Generate the (independent) person's face with the emotion detected from the speech.

- Regressing the Action Units (AU) corresponding to the face gives a more fine-grained control over emotion.

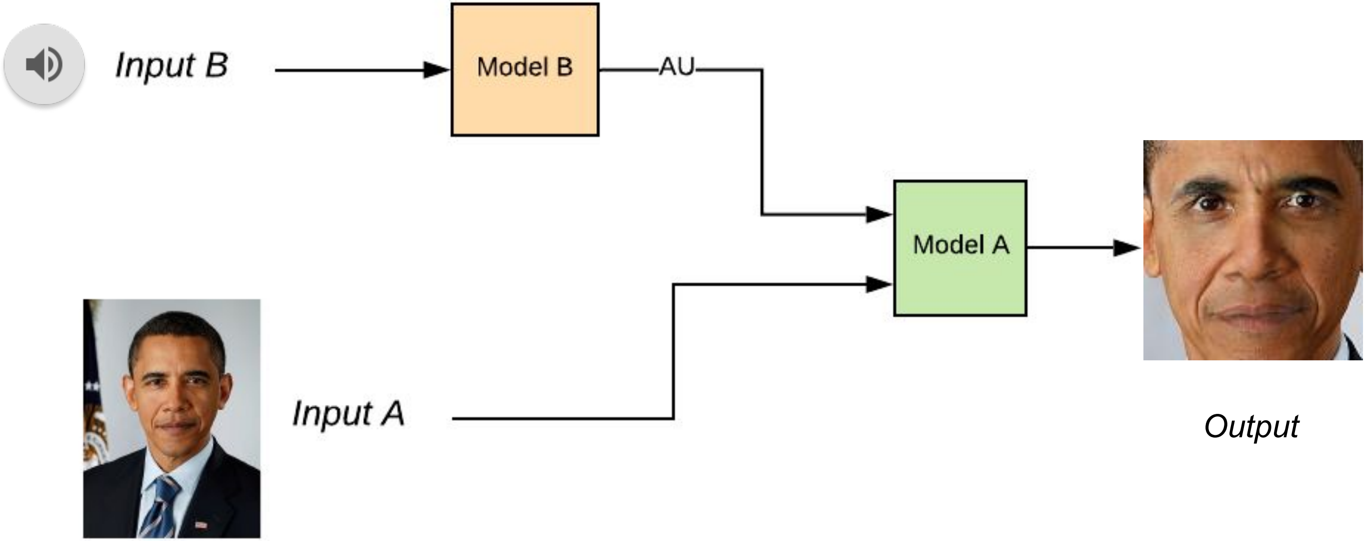
Action Units

- 17 Action Units determine facial coding.

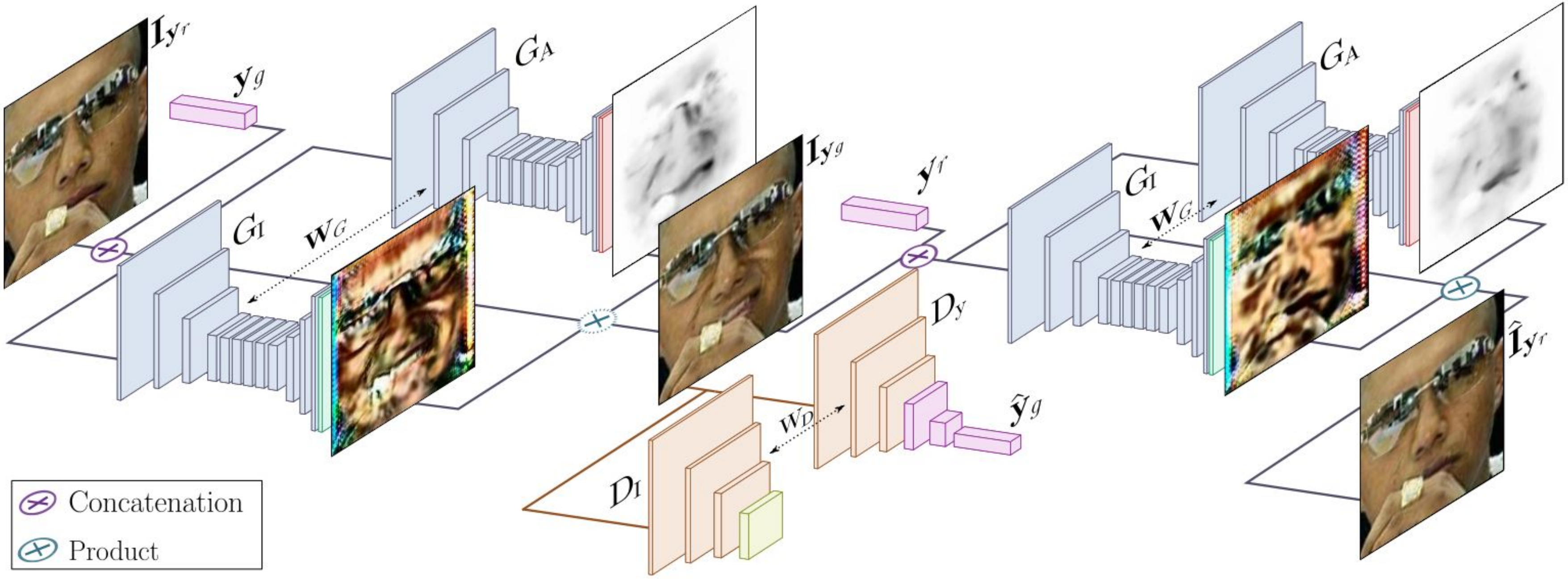
Emotion	**Action units**
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+16

Source: https://en.wikipedia.org/wiki/Facial_Action_Coding_System

Architecture

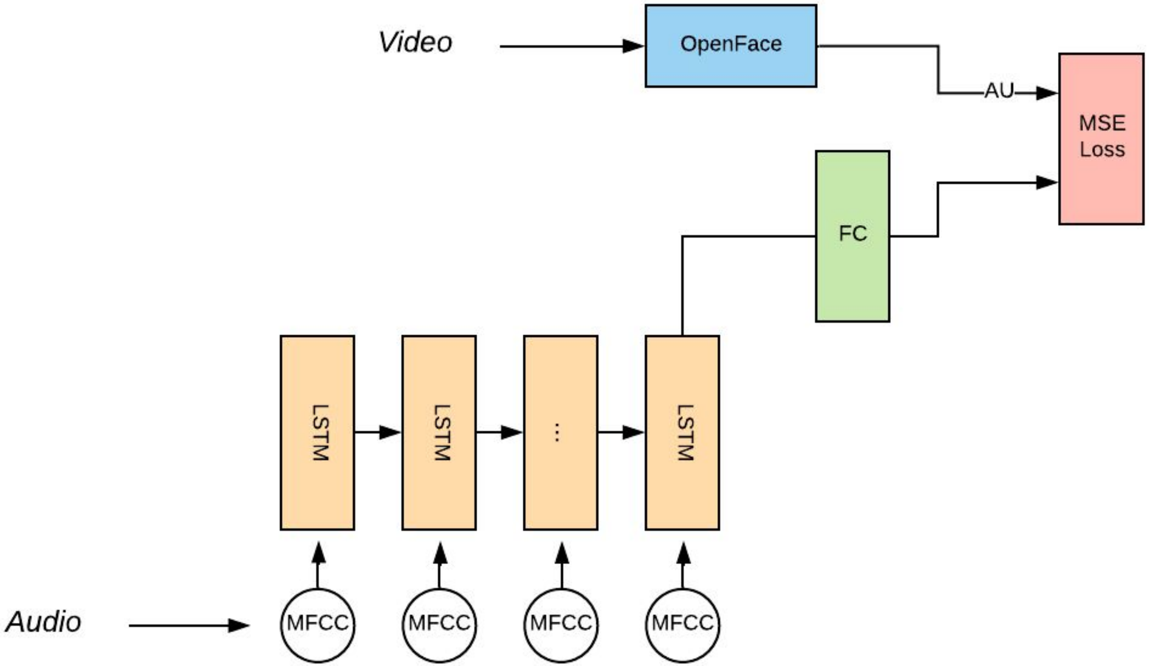


Model A



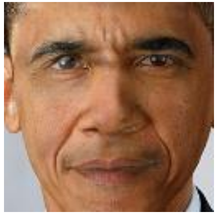
Pumerala et al. [1]

Model B

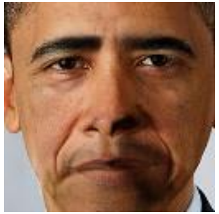

Video
OpenFace
AU
MSE
Loss
FC
LSTM
LSTM
...
LSTM
Audio
MFCC
MFCC
MFCC
MFCC

Results

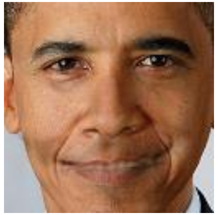
Angry 🔊



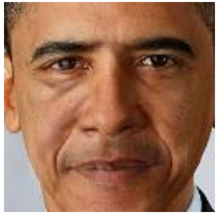
Sad 🔊



Happy 🔊

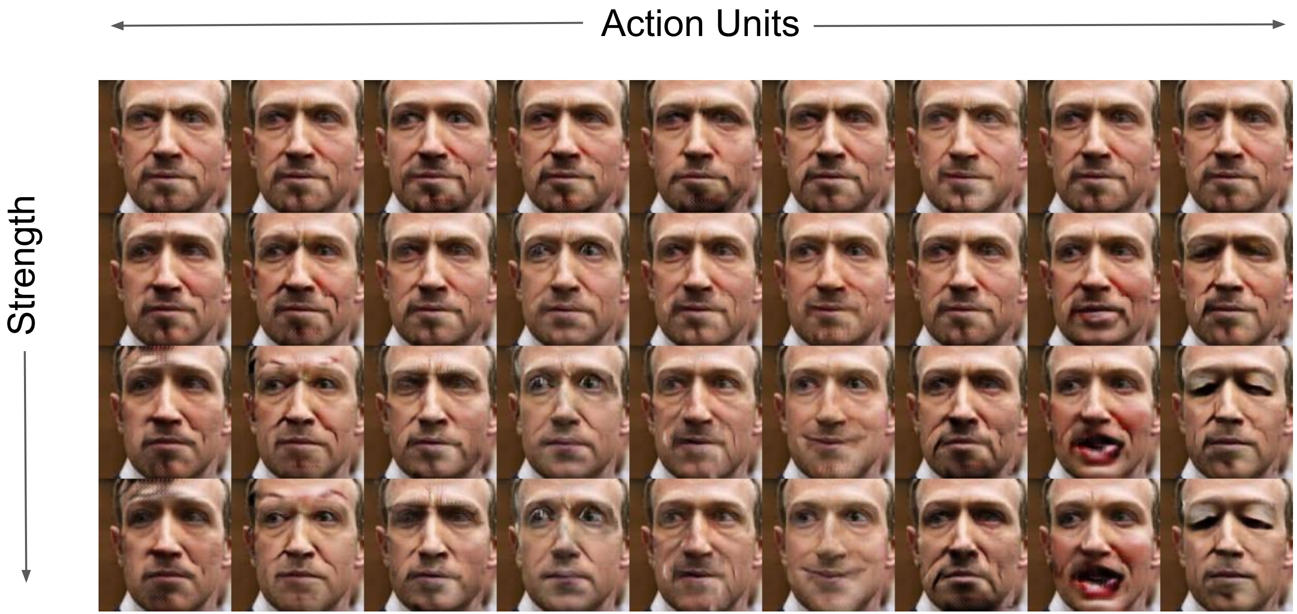


Disgust 🔊



Results: Model A

Original Image



Results

- We use the RAVDESS [3] dataset to 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).
- The regressed model though is able to capture the emotion, doesn't precisely do well in encoding the level of emotion as expected by doing regression instead of classification.

References

[1] GANimation: Anatomically-aware Facial Animation from a Single Image: https://arxiv.org/pdf/1807.09251.pdf

[2] Multimodal Speech Emotion Recognition Using Audio and Text: https://arxiv.org/pdf/1810.04635.pdf

[3] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): https://smartlaboratory.org/ravdess

[4] OpenFace: https://github.com/TadasBaltrusaitis/OpenFace/

Thank You!