CS766 Mid-Term Report

LIPGAN: SPEECH TO LIP-SYNC GENERATION

Abhay Kumar, Elizabeth Murphy, Maryam Vazirabad
{kumar95, emurphy7, mvazirabad}@wisc.edu

March 24, 2021

1 Introduction

We explore the problem of lip-syncing a talking face video to match the target speech segment to the lip and facial expression of the person in the video. The primary task is to achieve accurate audio-video synchronization given a person's face in a static image or video and a target audio clip. The goal is to produce dubbed videos which are dynamic and unconstrained. In this report[1], we discuss our current progress with the model.

2 Methodology & Progress

2.1 Current state-of-the-art of LipGAN

Prajwal et. al.[1] proposed a "Face-to-Face Translation" system, which incorporates the LipGAN model for synthesizing realistic talking faces in still images and videos from the target translated audio. It takes a clip of a person speaking in a source language and output a video of the same speaker speaking in a target language such that the voice style and lip movements justify the target language.

2.2 Model architecture overview

We have implemented LipGAN [1] model. Here is a brief summary of the Generator and Discriminator networks.

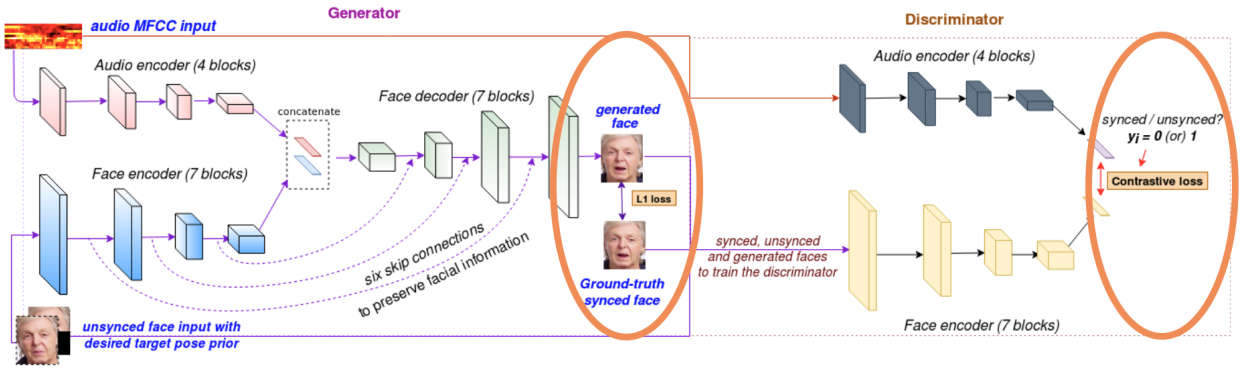

Figure 1: Overview of the LipGAN architecture [1]. We are focusing on modifying the *face reconstruction loss module* and *discriminator network* (both blocks encircled in orange) for better supervision around the lip region.

[1] https://abhayk1201.github.io/CS766_Project/midterm.html

Generator network

- **The Face Encoder**: This CNN module encodes the face features, including identity and pose.
- **Audio Encoder**: The audio encoder is a standard CNN model that takes a Mel-frequency cepstral coefficient (MFCC) heatmap and creates an audio embedding
- **Face Decoder**: This module synthesizes a lip-synchronized face from the joint audio-visual embedding by inpainting the masked region of the input image with an appropriate mouth shape.

Discriminator network

Contrastive loss between the encoded audio and encoded face is used to supervise the generator module to learn robust, accurate phoneme-viseme mappings to produce satisfactory talking faces with more natural facial movements.

2.3 Dataset

The LRS2 dataset [2] consists of thousands of spoken sentences from BBC television. Each sentences is up to 100 characters in length. The training, validation, and test sets are divided according to broadcast date. The train set contains 45,839 utterances with 329,180 words instances and 17,660 vocab size.

2.4 Pre-processing Result

We extracted the frames at 25fps using `opencv` and trimmed the face coordinates of the speaker. We used `dlib` `get_frontal_face_detector` to detect the speaker face from the video frames. Each detected face is resized to 96×96. The LipGan requires that audio files are in the form of a Mel-frquency cepstrum features (MFCCs) so we converted the audio files using the python `audio` library to match this format. Pre-processed outputs for a video is shown in Figure 2.

Figure 2: Pre-processeding output for the utterance- *"I Liked the radio podcast"*

2.5 Generating Talking Face from Image and Audio

After training the model with the pre-processed data set, we are able to generate talking face videos from static images and audio files. Our current progress does not yet include dubbing onto an original video. Sample generated videos from inputted static images and audio files are shown in the 'Video Links' of Figures 3 and 4.

3 Problems Faced & Possible directions

3.1 Technical Challenges

- It's hard to have a quantitative measure of the model's performance which could substitute human evaluation. The literature has shown many quantitative measures like Peak signal-to-noise ratio (PSNR), Structural SIMilarity (SSIM) Index [1], Landmark distance [3], Lip-sync error Distance (LSE-D)[4], audio-visual match distance, etc. However, these quantitative measures do not clearly help us identify the shortcomings of the LipGAN model. So, we have been trying to verify the performance qualitatively primarily, as the quantitative metrics could not substitute human evaluation. We are performing the human evaluation ourselves and are not using other non-bias judges (which is what was used in the research of the original LIPGAN model). The end goal is to produce a video that is visually pleasing for the human viewer. We have shown few shortcomings that we identified during human qualitative evaluation in the next few points (Note these shortcomings are not explicitly mentioned in the LipGAN paper).

- Limitations in profile face: LipGAN overcompensates one side lip movement, i.e the generated lip movement is more than natural one. This happens primarily for profile view of the face.

- Limitations in Lip Movement and Audio Syncing: Sometimes during pauses in audio, the lips in the generated dubbing will continue to move. This may be due to the model picking up on small background noises in the audio that are not noticeable to the average listener.

- Limitations due to spurious lip region detection: We have noticed that lip-sync generation has spurious movements on non-lip region, like lower chin or side chin as shown in Figure 3. We have observed this when the face detection fails to correctly localize the lip region. Profile view of the detected face usually faces this limitation. We need to observe few more such cases to generalize the limitation and come up with a possible modification. We are yet to see how focusing [refer *Ongoing progress*] on lip-region will help improve such cases.

- Teeth Region Deformation: We observed that the LipGAN model generates image frames which smoothed out teeth and lip region. Lower teeth is merged with upper lip and smoothed out. This deformation is shown in Figure 4.

- Limitations due to Facial Expression: We observed that with certain facial expressions that the LipGAN model preforms worse. One example of this is when the person in the image has a deep frown. The model has a harder time detecting the location of the lips. (Often missing the bottom corners of the frown)

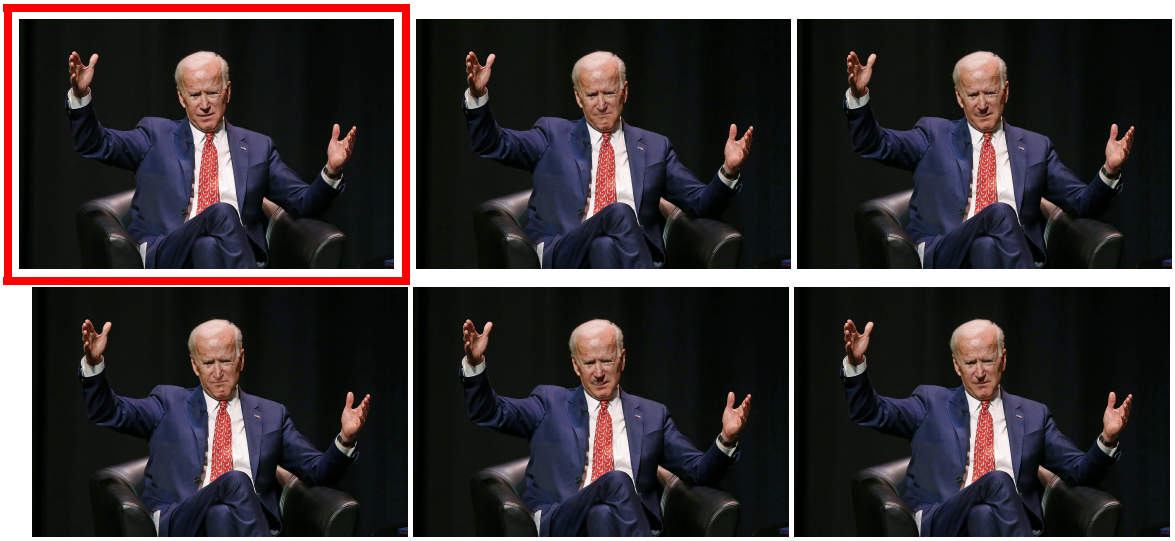

Figure 3: Generated image frames from LipGAN showing spurious lip region movements on lower chin. Input image is inside the red box. The generated image frames have smoothed out the teeth and lip pixels. Video Link

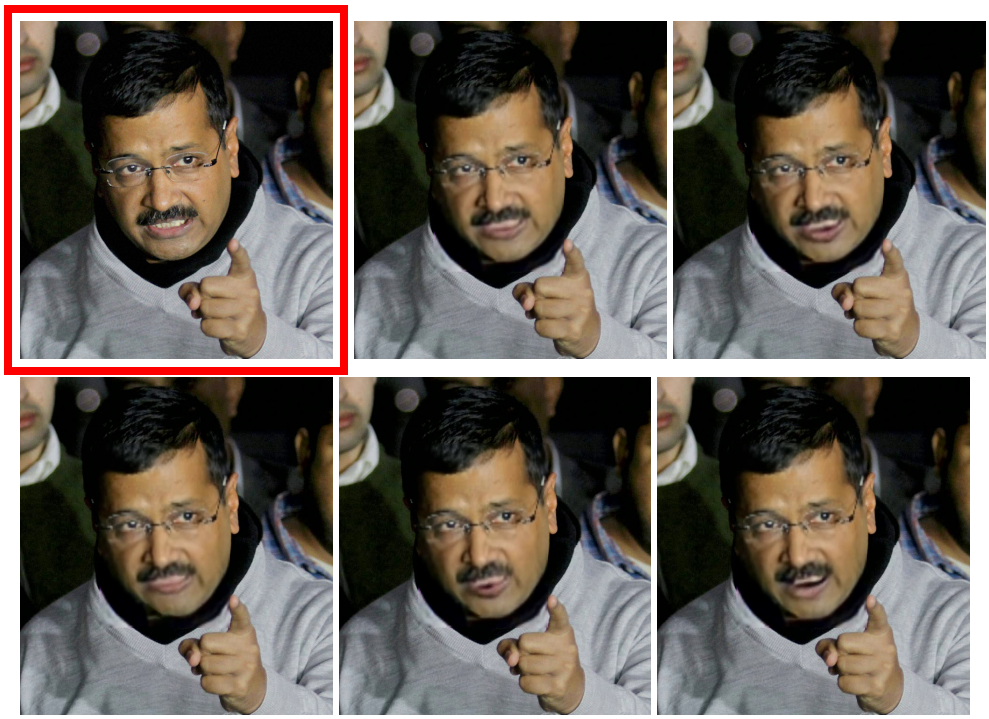

Figure 4: Generated image frames from LipGAN showing deformation of teeth region. Input image is inside the red box. The generated image frames have smoothed out the teeth and lip pixels. Video Link

Ongoing progress We are trying to modify the face reconstruction loss module. Currently, the face reconstruction loss is calculated for the whole image to ensure correct pose generation, background around the face, and preservation of the identity. The lip region contributes to less than 4% of the total reconstruction loss. However, we need to emphasize the reconstruction loss in lip region. We are planning to explore different techniques, like weighted reconstruction loss, or having a separate discriminator (as in a multi-task setting) to focus on the lip-sync only. We can jointly train the GAN framework with two discriminator networks (one for visual quality, and one for lip sync).
Therefore, we will consider using real people to act as judges and evaluate the lip-synchronization based on performance metrics discussed in previous works.

- Please note that the technical challenges that we have presented above will be hard to be solved fully. But, we will try to present an ablation studies and activation heatmaps if the *Ongoing progress* helps or not.

- *Optional Goal:* If time and computational power permit, we can experiment with different model architectures for each of the blocks mentioned in 2.2. For example, we can use state-of-the-art model architectures to extract richer and complex audio and face embedding.

3.2 Implementation Challenges

- LRS2 dataset [2] acquisition and pre-processing tasks were not easy given the huge size (roughly 50GB). The pre-processed files were not readily available for use because of the signed Data Sharing agreement [5] with BBC Research & Development. After several attempts, we were able to download the part files [6] and save it to shared Google Drive. Now, we can mount the shared drive on Google Colab VM.
- We are training the model on Google Colab [7]. Given the high space requirement and processing power, we have been facing some limitations like timeout of the running session, disk quota constraint, etc. (Note: we are facing these limitations despite having the Colab Pro subscription). We decided to randomly select a subset from the original data to train a separate model for that. Now, we are training two separate models: one for a subset of original dataset and one on the complete dataset.
- Although the original model described in the LipGAN paper utilizes MATLAB and was the one most extensively researched, due to issues working with the MATLAB implementation and our preference for using Python, we are using Python to implement the model. However, even with our preferred language of choice, we have had to resolve multiple python packages dependencies issues to have a working setup on Google Colab.

4 Project Website

Project Website: https://abhayk1201.github.io/CS766_Project/
Mid-Term Report: https://abhayk1201.github.io/CS766_Project/midterm.html
[2]

References

[1] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1428–1436, 2019.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[4] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[5] Lip reading sentences in the wild agreement (lrs2) document. https://www.bbc.co.uk/rd/projects/lip-reading-datasets.

[6] The oxford-bbc lip reading sentences 2 (lrs2) dataset. https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html.

[7] Google colaboratory. https://colab.research.google.com/.

[2] https://abhayk1201.github.io/CS766_Project/