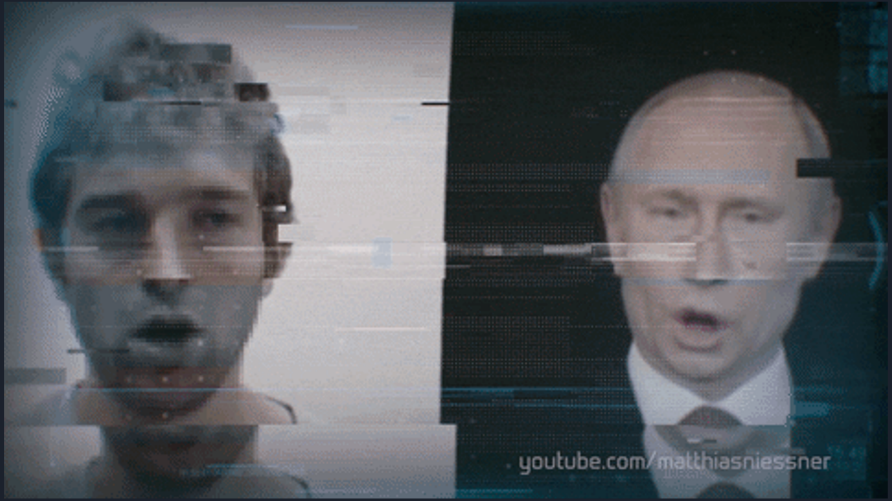
CS 766 - SPEECH TO LIP SYNC GENERATION

Elizabeth Murphy, Abhay Kumar, Maryam Vazirabad

Speech to Lip Sync Generation

- Purpose:
- lip-syncing a talking face video to match the target speech segment to the lip and facial expression of the person in the video
- Applications:
- realistic dubbing in movies, CGI animations in movies, and gaming.
- Deepfake Technology
- emerging form of synthetic media
- state of the art models are so convincing that they have the potential to deceive



Dataset / Preprocessing

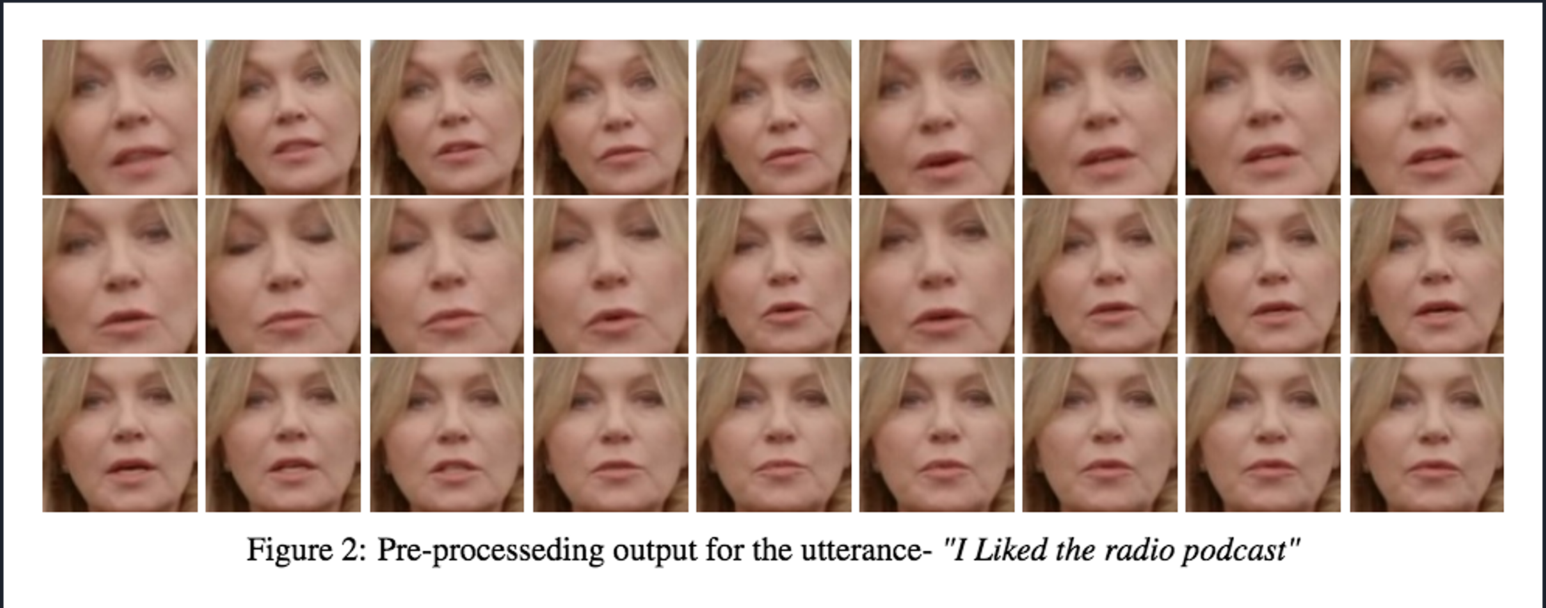
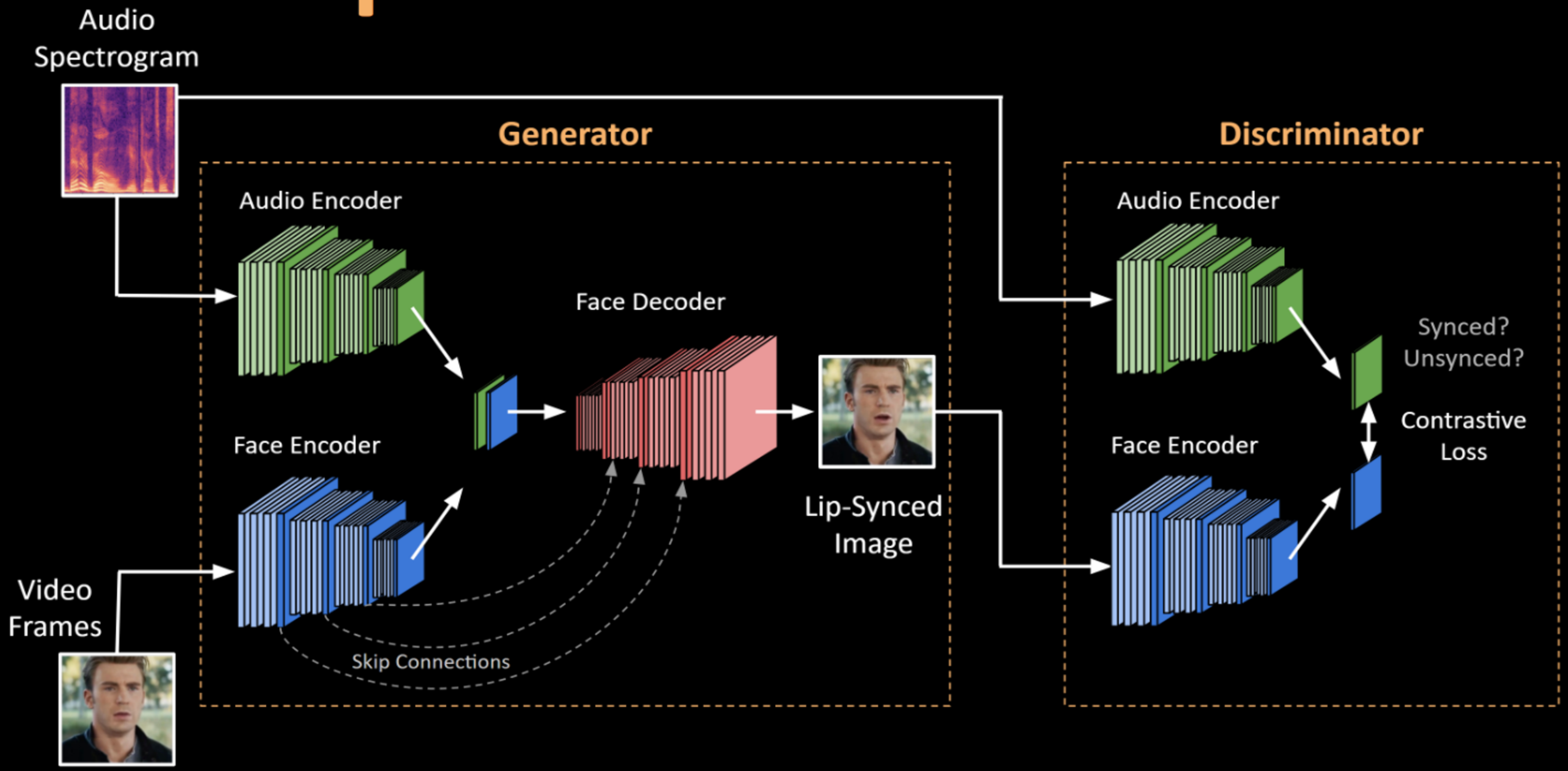


Figure 2: Pre-processeding output for the utterance- *"I Liked the radio podcast"*

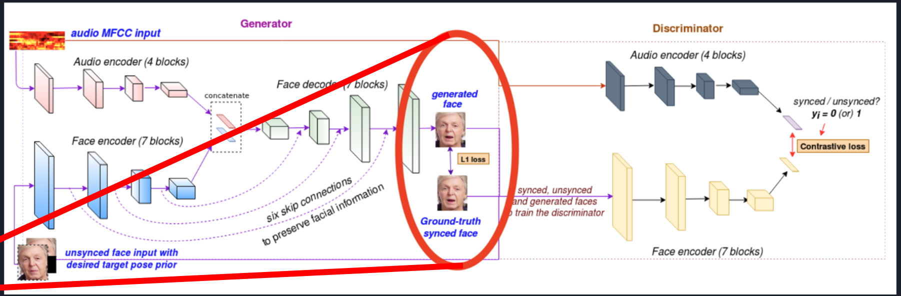
LipGAN FRAMEWORK



[1]

Results

Experiments

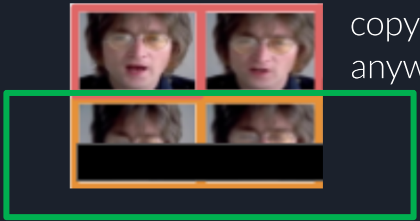


Weighted L1 reconstruction loss

- Idea: Lip reason contributes to less than 4% of the total reconstruction loss. Can we improve by **focusing** on the lip region.

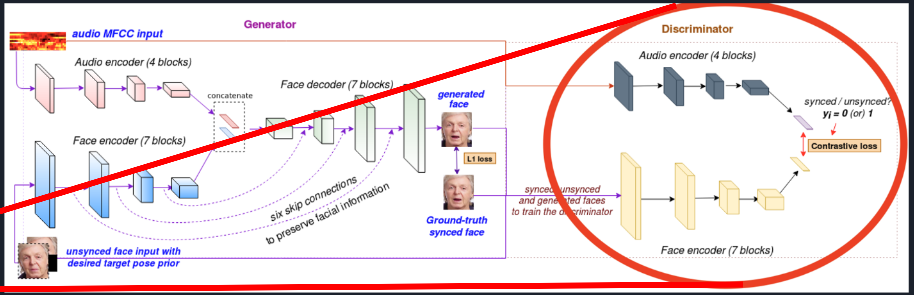
- Early epochs vs late epochs reconstruction loss.

- Target pose prior(actual frame with masked lower portion): Generative model learns to copy the remaining portion anyways. So, focusing on lip region is anyways redundant after few initial epochs.



- Pixel level reconstruction is not a strong judge of lip-sync.

Experiments



Discriminator Network

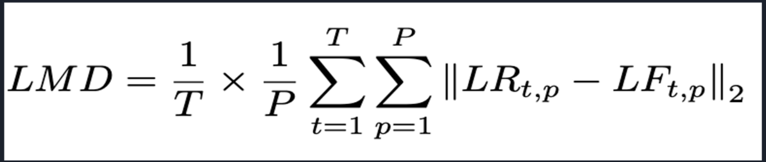
- Idea: Can we add a additional discriminator in Multi-task setting.

- In literature: Researcher have tried with **Expert Lip-sync discriminator,** which is not fine-tuned during model training. Having a Lip-sync expert (pre-trained) as an additional supervision helps to accurately discriminate and enforce lip-sync in generated images.

- Testing discriminator network : ~63% on 1000 randomly generated lip sync face images (due to lot of artifacts due to large scale and pose variations)

Limitations of LipGAN Model

- **Difficult to <u>quantitatively</u> measure shortcomings**
  - Landmark Distance
    (the lower the better)



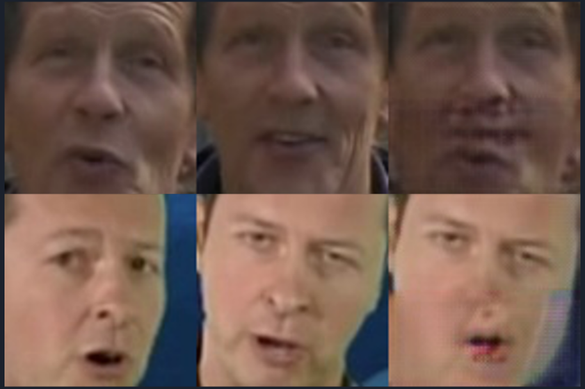
$$LMD = \frac{1}{T} \times \frac{1}{P} \sum_{t=1}^{T} \sum_{p=1}^{P} \|LR_{t,p} - LF_{t,p}\|_2$$

Shortcoming: Just reducing lip movement globally (as in mumbling) will satisfy this.

  - PSNR: Developed to evaluate the overall image quality and not fine-grained lip sync error.
    Same for SSIM (Structural SIMilarity (SSIM) Index)

Limitations of LipGAN Model

- Difficult to quantitatively measure shortcomings
- **Spurious lip region detection**



Limitations of LipGAN Model

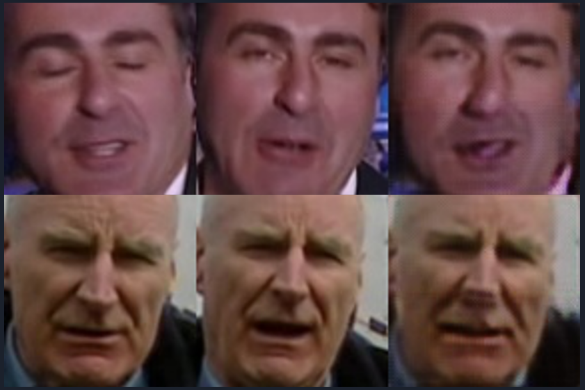
- Difficult to quantitatively measure shortcomings
- Spurious lip region detection
- **Profile face overcompensation / skewed lip sync**

Limitations of LipGAN Model

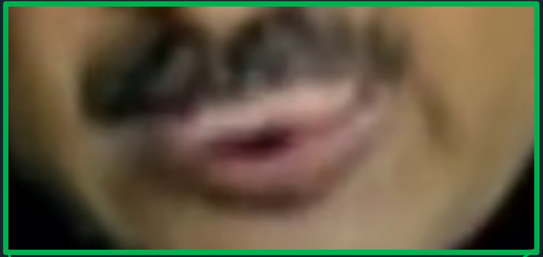
- Difficult to quantitatively measure shortcomings
- Spurious lip region detection
- Profile face overcompensation
- **Issues with lip movement and audio synchronization. Especially, background music leads to high murmuring lip movement.**

Limitations of LipGAN Model

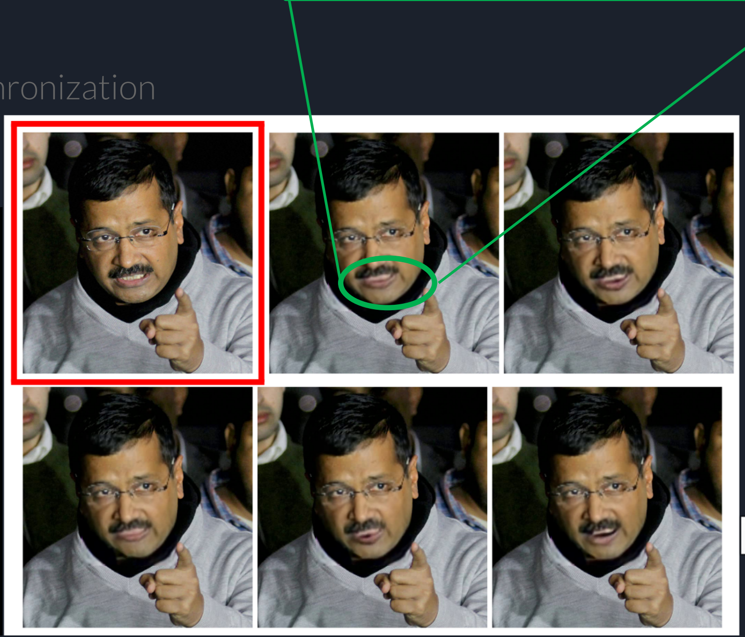
- Difficult to quantitatively measure shortcomings
- Spurious lip region detection
- Profile face overcompensation
- Issues with lip movement and audio synchronization
- **Teeth Deformation (crooked teeth) or no teeth at all**



Limitations of LipGAN Model



- Difficult to quantitatively measure shortcomings
- Spurious lip region detection
- Profile face overcompensation
- Issues with lip movement and audio synchronization
- Teeth Deformation
- **Limitations due to facial expression**



Implementation Challenges

- Computation time for preprocessing (~**50GB raw data**; ~**30hr of video; ~10GB processed data**)
- Limitations due to use of **Google Colab PRO** (Session getting killed after certain time)

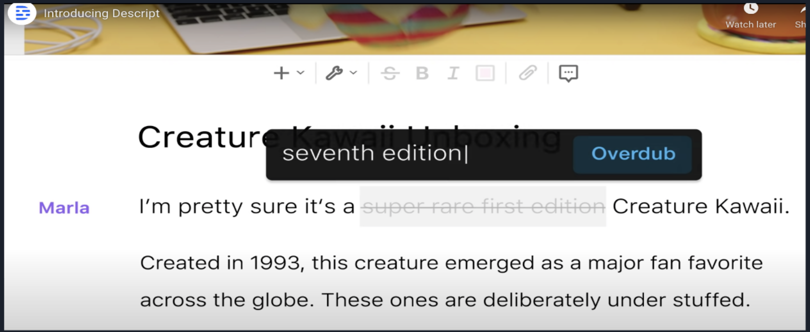
  → single gpu training took 3hr/epoch [Tesla P100]

  → ~71 M trainable parameters [Resnet50 had ~25M parameters]

- Storage size limitation on google colab.

Discussion & Future Work

- Similar model works best for editing/dubbing a small segment of a video. [**lyrebird.ai**]



Creature Kawaii Unboxing

seventh edition|

Marla I'm pretty sure it's a ~~super rare first edition~~ Creature Kawaii.

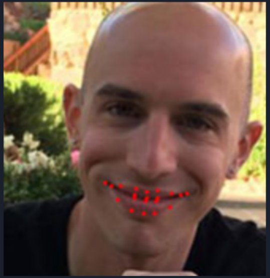
Created in 1993, this creature emerged as a major fan favorite across the globe. These ones are deliberately under stuffed.

- Speaker independent; Language independent; Pose invariant.

Discussion & Future Work

- Works well with animated cartoon as well, creating lip-synced bitmoji or AR emoji

Discussion & Future Work

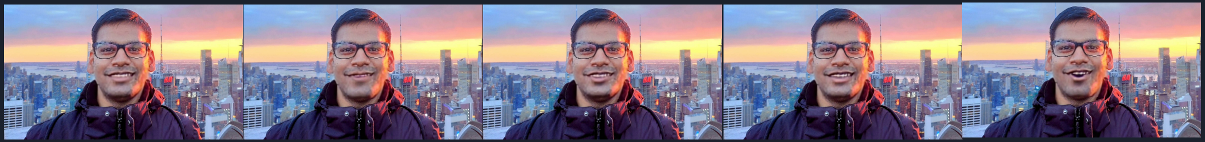


- Need more detailed keypoint detector for lip region.

- Worth trying out 3D representation of face, with mesh-like grid to have more structured and smooth lip movement along with side cheeks, jawline etc. (muscles getting pulled/pushed/squeezed)

  → will also enforce a sense of depth measure.

- Million dollar idea: Live lip-syncing in video call (even if speaker is not sharing his video feed, we just need one static face image)
  → Live lip-synced video + privacy



- Watch lip-synced dubbed movies/Tv series. I lip-synced 10 min of “Money Heist” (dubbed from Spanish to English) and I definitely liked the lip-synced version!!

Privacy Matters!

+

“This group did a great job.”

✗

But, we did NOT!

PLEASE Take prior permission before publishing Lip-synced videos!!!

One last demo!

- English Translation: “I have wanted you so much, so truly ... that the entire universe has conspired for me to get you”   (Hindi Movie dialogue)

References

1. https://medium.com/deepgamingai/automate-your-lip-sync-animations-with-this-ai-lipgan-ad35551ae62d
2. https://arxiv.org/pdf/1803.10404.pdf
3. http://cdn.iiit.ac.in/cdn/cvit.iiit.ac.in/images/Projects/facetoface_translation/paper.pdf
4. https://arxiv.org/abs/2103.00484
5. https://arxiv.org/pdf/2005.08209.pdf
6. https://arxiv.org/pdf/2008.10010.pdf
7. https://arxiv.org/pdf/1705.02966.pdf
8. https://github.com/Rudrabha/LipGAN/tree/fully_pythonic
9. https://www.bbc.co.uk/rd/projects/lip-reading-datasets [dataset]
10. https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
11. https://colab.research.google.com/
12. https://www.descript.com/lyrebird https://www.descript.com/video-editing

Questions?