Unsupervised Generative Video Dubbing

Jimin Tan, Chenqin Yang, Yakun Wang, Yash Deshpande

Center for Data Science, New York University | Advisor: Prof. Kyunghyun Cho, Bloomberg, NYC Media Lab

Introduction

Goal The project aims to modify a video of a person speaking in one language so that the person is perceived as speaking the same content in another language.

Method We build a generative model to modify an input video, and use pre-trained Visual Speech Recognition (VSR) models as proxies to teach the generator different lip shape representations.

Dataset

Lip Reading in the Wild (LRW)

- Around 1000 utterances of 500 different words from BBC newscasts

Newscast video (courtesy of Bloomberg)

- Narrated in English & Spanish with transcripts

Figure 1: LRW Dataset

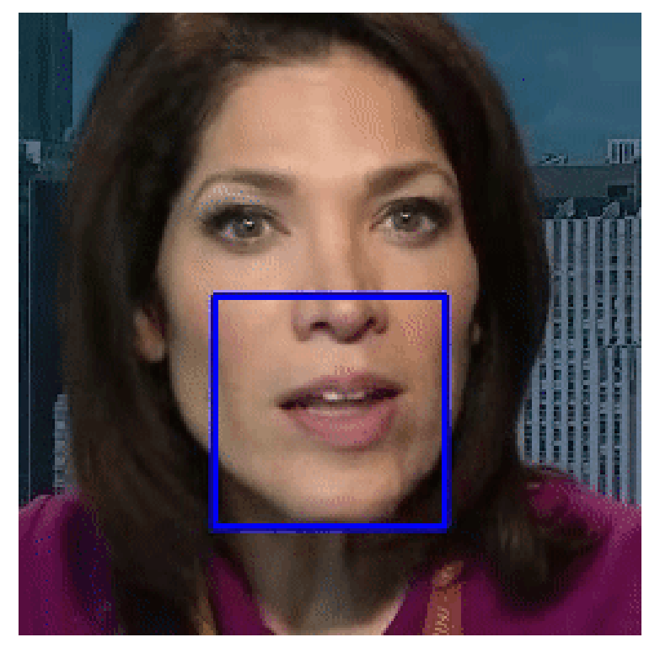

Figure 2: Newscast Video

Generative Model for Lip Modification

Our generative model **G** takes in as input a video clip ($\mathbf{x}$) of a person speaking a single word ($\mathbf{s}$), a target word ($\mathbf{t}$) and noise ($\mathbf{h}$) and outputs a video that will be classified by a VSR as ($\mathbf{t}$). With discriminator **D**, the loss function for the network can be written as:

$$\mathcal{L} = \mathbf{E}_x \left[\log D(x)\right] + \mathbf{E}_h \left[\log\left(1 - D(G(h))\right)\right] \qquad (1)$$

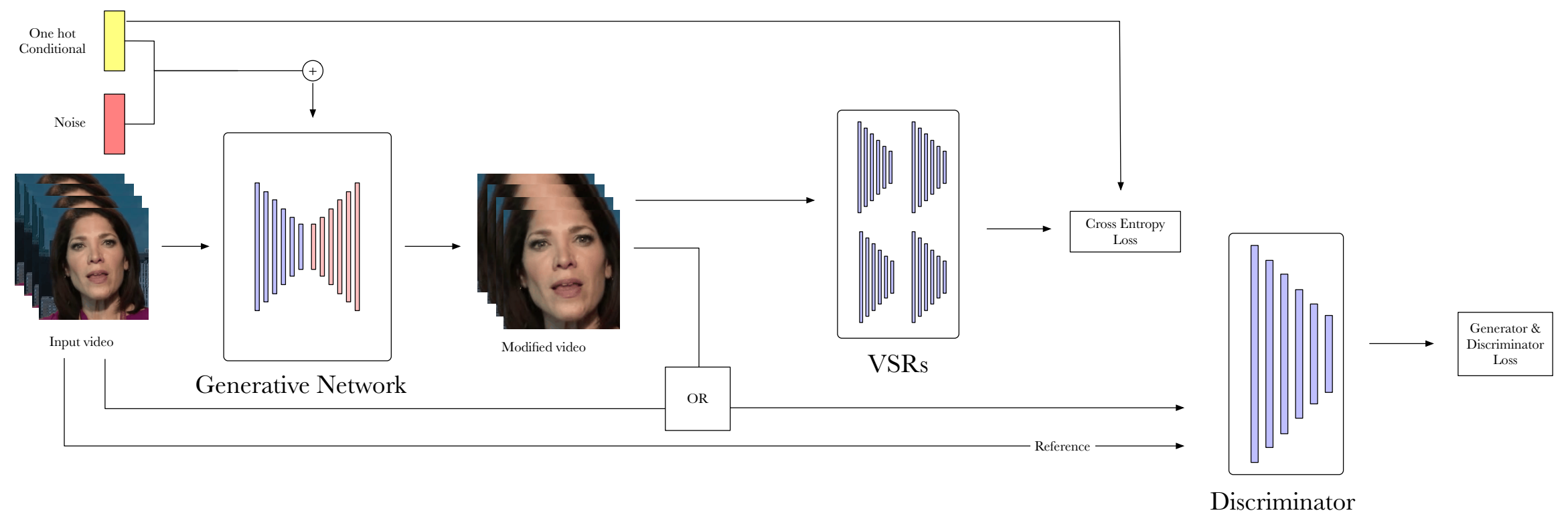

Figure 3: Overall Structure

Generator & Discriminator Structure

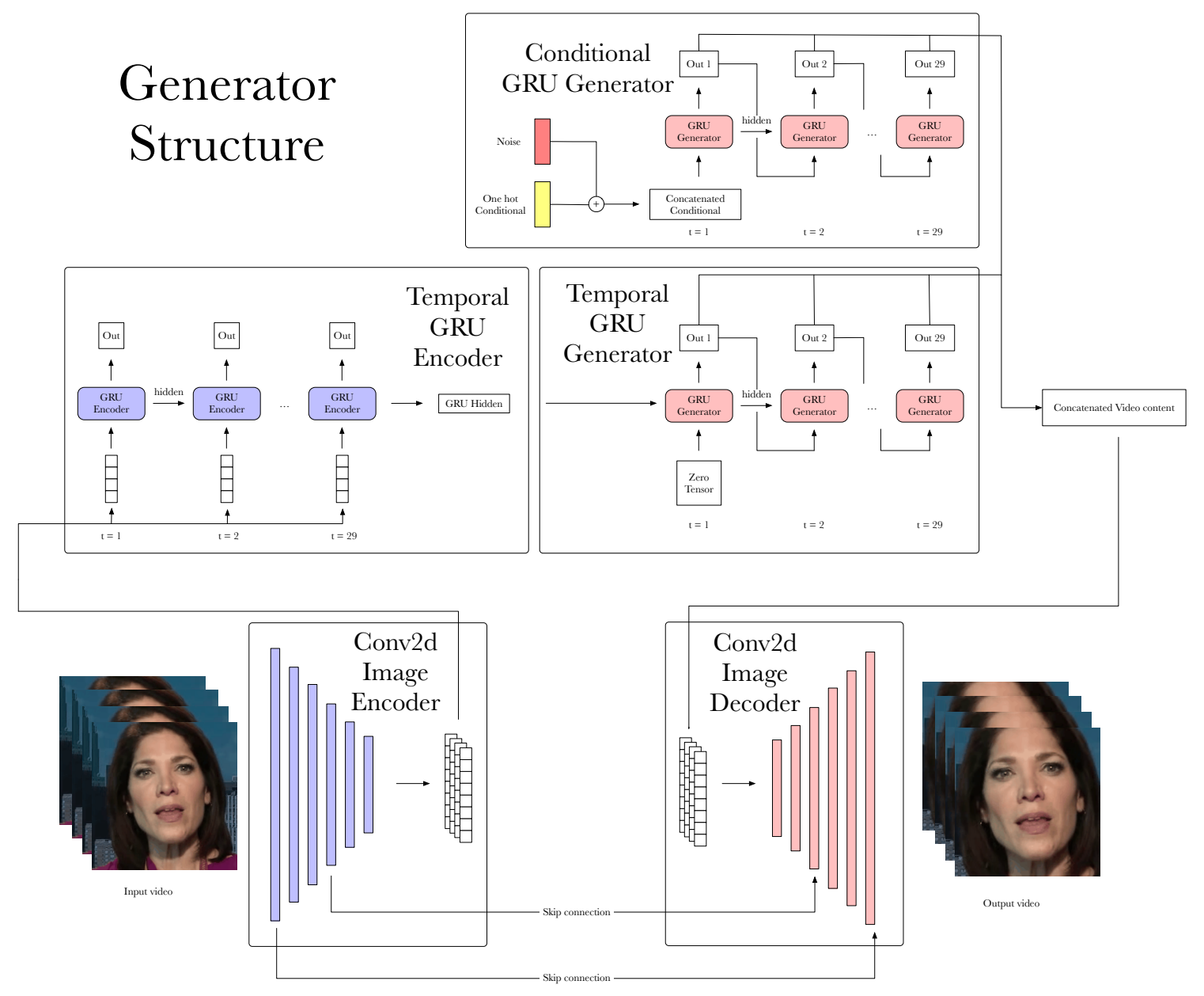

Figure 4: Generator

The generative network is an autoencoder that encodes image-level features with CNNs and temporal features with GRUs. Skip connections are added between the encoder and decoder for better reconstruction quality.

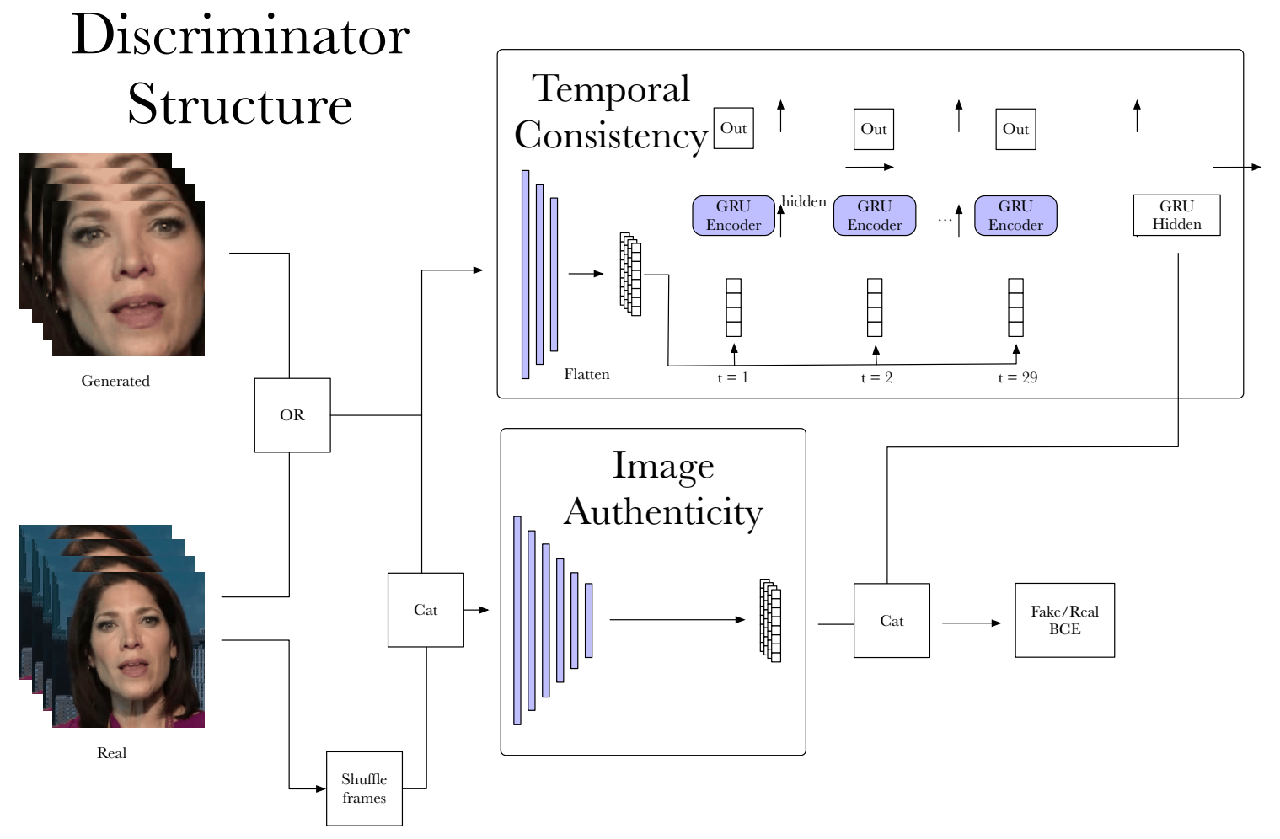

Figure 5: Discriminator

The discriminator consists of a temporal module for penalizing inconsistency between frames and an image quality module for evaluating individual frames.

Results

A separate set of hold-out VSR systems is used to evaluate our generative model. The generative model achieves a *top-1* accuracy of 83% and a *top-3* accuracy of 94% during evaluation.

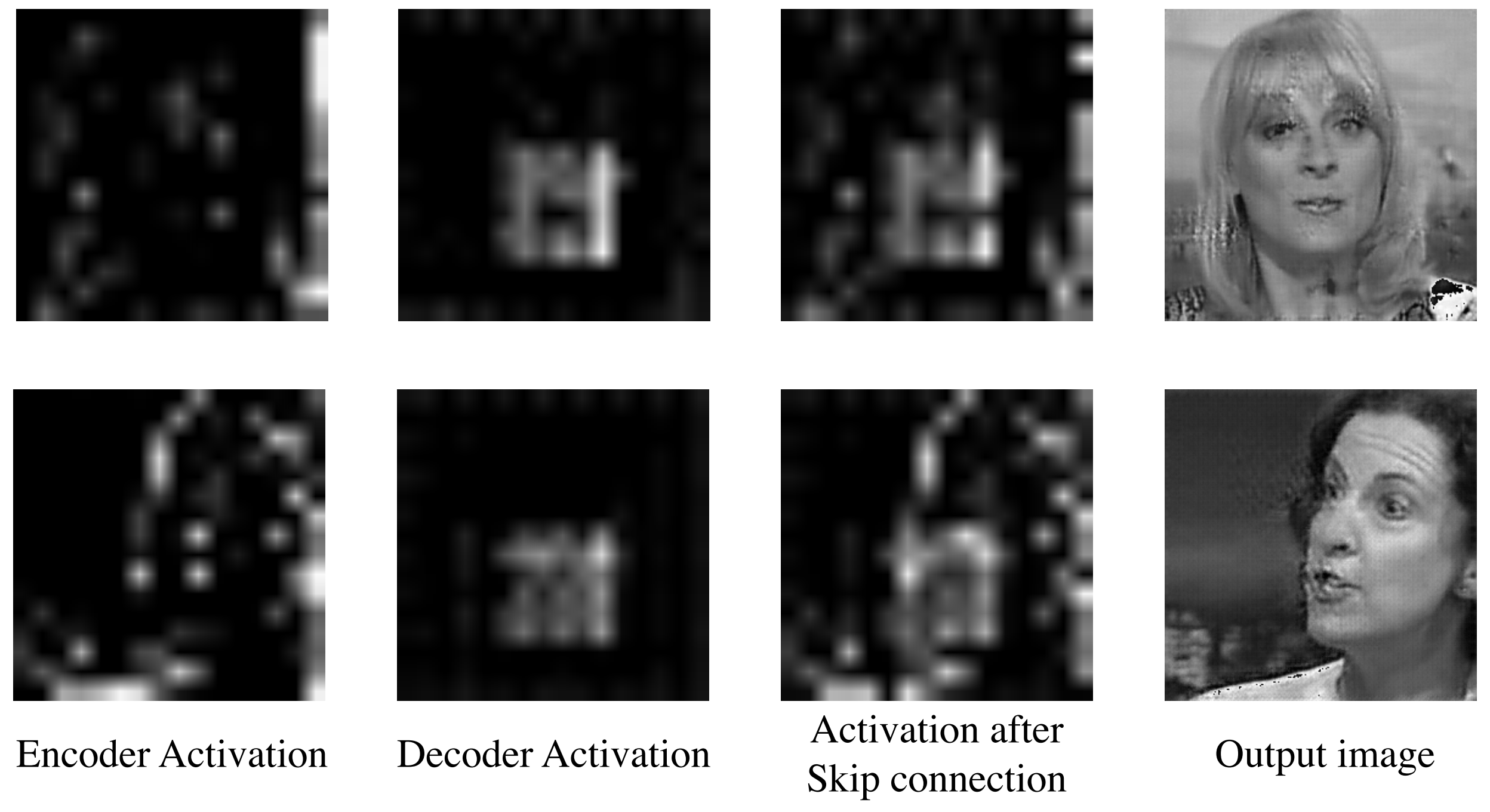

Figure 6: Activations in the generator

Results (contd.)

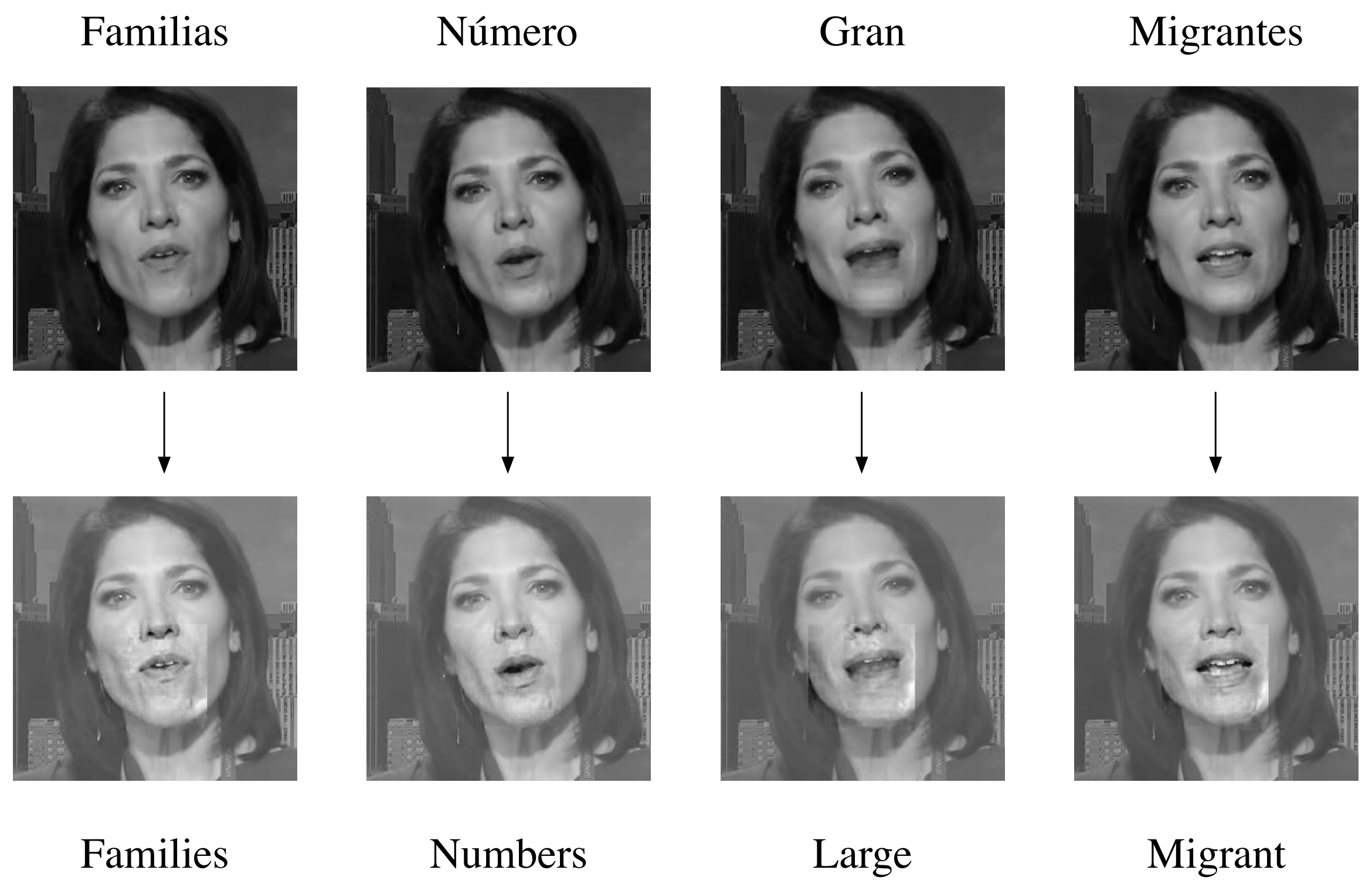

Figure 7:Spanish to English video modification

Future Work

- Update generative model for generating:
  - RGB videos
  - Sentence-level videos
- Unfreeze VSR and use it as a second discriminator
- Perform crowd-sourced human evaluation

References

[1] Chung, J. S., Zisserman, A. (2016, November). Lip reading in the wild. In *Asian Conference on Computer Vision* (pp. 87-103). Springer, Cham.

[2] Shrivastava, N., Saxena, A., Kumar, Y., Shah, R. R., Stent, A., Mahata, D., ... Zimmermann, R. (2019). MobiVSR: Efficient and Light-weight Neural Network for Visual Speech Recognition on Mobile Devices. *Proc. Interspeech 2019*, 2753-2757.

[3] Stafylakis, T., Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. *arXivpreprint arXiv:1703.04105.*

[4] Vougioukas, K., Petridis, S., Pantic, M. (2019).Realistic Speech-Driven Facial Animation with GANs. *arXivpreprint arXiv:1906.06337.*