CS766 Project Proposal

LIP-GAN: SPEECH TO LIP-SYNC GENERATION

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

February 24, 2021

1 Introduction

1.1 Problem Description

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 synchronisation given a person's face and target audio clip. We can extend it to be speaker-independent model to produce lip-sync in the "wild", where videos feature faces that are dynamic and unconstrained.

1.2 Motivation

With the increasing amount of multi-media content, there has been multiple research on synthesizing accurate, realistic talking face videos. This technique is broadly applicable to many scenarios such as realistic dubbing in the movie industry, conversational agents, virtual anchors, and gaming. There has become an increasingly large market for dubbing a foreign language onto videos; as the rate of video content creation increases, so does the need for accessibility for viewers across the globe. Providing a natural lip movement and facial expression generation improves the user's experience in these applications. Despite the recent advances [1, 2] and its wide applicability, synthesizing a clear, accurate and human-like performance is still a challenging task. We want to explore the state-of-the-art techniques and their limitations.

2 Methodology

2.1 Current state-of-the-art

Recently, Prajwal et. al.[3] proposed a "Face-to-Face Translation" system, which incorporates **Lib-GAN** for synthesizing realistic talking faces in still images and videos from the target translated audio. They propose a speech-to-speech translation pipeline that can take 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. This significantly improves the user experience by providing a more realistic and holistic translation system. [2] improved upon this work by producing significantly more accurate lip-synchronization in dynamic, unconstrained talking face videos and designed new evaluation benchmarks to measure lip-sync performance.

2.2 Project Approach

We are planning on implementing Lib-GAN [3] model and improve upon the work (refer 2.3). The Lip-GAN model comprises of Generator and Discriminator networks. Here is the brief summary of the modules.

[*]Sorted alphabetically

Generator network

- **The Face Encoder**: This 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 Existing Approach Limitations & Possible directions

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 should try 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 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).

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 Performance Evaluation

We will use the LRS 2 dataset [4] which contains over 29 hours of talking faces in the provided train split in the dataset.

3.1 Ablation studies & Comparison

Different standard quantitative metrics [5] have been used in the literature to compare the performance of Lip-sync task.

- Peak signal-to-noise ratio (PSNR): Measures the quality of reconstruction
- Structural SIMilarity (SSIM) Index [3]: For image/video quality assessment
- Landmark distance [6]: Evaluates whether the synthesized video corresponds to accurate lip movements based on the input audio. It use Dlib, a HOG-based facial landmarks detector.
- Lip-sync error Distance (LSE-D)[7]: Average error measure calculated in terms of the distance between the lip and audio representations. A lower LSE-D denotes a higher audio-visual match, i.e., the speech and lip movements are in sync.

Most importantly, we need to verify the performance qualitatively too, as the quantitative metrics could not substitute human evaluation. The end goal is to produce a video that is visually pleasing for the human viewer. Therefore, we will consider using real people to act as judges and evaluate the lip-synchronization based on performance metrics discussed in previous works.

4 Timeline

Till Feb 24:	Come up with a project proposal document and create web page Discussion about possible techniques and directions to be explored.
Feb 25 - Mar 07:	Set up the running environment for the state-of-the-art methods. Dataset collection Start implementing an existing approach as a baseline.
Mar 08 - Mar 20:	Have one working implementation of an existing approach (Lip-GAN).
Mar 21 - Mar 24:	Write Mid-term report.
Mar 25 - Apr 11:	Try experimenting with different loss formulations, multi-task setting in discriminator networks. Explore alternate methods available in literature.
Apr 12 - Apr 23:	Complete all experiments. Summarize result and ablation studies. Prepare presentation slides.
Apr 25 - May 5:	Complete course project web page.

Table 1: Tentative Timeline

5 Project Website

Project Website: `https://abhayk1201.github.io/CS766_Project/`
[2]

References

[1] Ruobing Zheng, Zhou Zhu, Bo Song, and Changjiang Ji. Photorealistic lip sync with adversarial temporal convolutional networks. *arXiv preprint arXiv:2002.08700*, 2020.

[2] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

[3] 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.

[4] 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.

[5] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[6] 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.

[7] 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.

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