Generating Talking Head Models From Text With Synced Voice

Chenrong Lu, Shane Mulligan
BSC in Computer Science, minor Mathematics, DipGrad Computer Science
University of Otago, Dunedin, NZ
luch5015@student.otago.ac.nz,mullikine@gmail.com

1 Introduction

In this report I introduce a pipeline to animate a chatbot. From just one single photo, and optionally, a reference voice, the animation pipeline is able to synthesize a realistic chatting head which can be dynamically prompted with text to generate talking sequences where the audio is synced to lip movements.

Although I have formatted this report like a research paper, it is by no means trying to be one. There are no new research being done, or anything to contribute to science through this report. This is just the summary and account of a hobby project, which demonstrates the possibility of using deep learning to create realistic chatting heads.

There are many ways to make the pipeline much, much better, without fundamentally changing the stages of the pipeline. For example, getting better lip sync, reducing jittering, and rendering higher quality faces. More on that later.

Figure 1: Fast synthesis of talking head from text prompt: Joi from Bladerunner 2049.

2 Overview

The motivation for creating this project is the observation that deep learning based implementations of parts of this pipeline are becoming more robust and accessible with the computation of commodity hardware.

These parts of the pipeline are:
Computer Chat Program(Chatbot)
Text to speech synthesis(TTS),
speech to facial land mark estimation(STL),
facial landmark to facial animation(LTF).

The pipeline is constructed with modularity in mind. So each of the part of the pipeline could be substituted freely. I have also implemented scripts for ”skipping in”, to any part of the pipeline if wanted. So that the final animation can be done via: Chatbot Driven, Text Prompt Driven, Audio Prompt Driven, or Video Prompt Driven.

Text to speech synthesis via unit selection synthesis has been widely applied, they are robust, realistic, and predictable while being able to customise to a desired human voice. Although in this pipeline I have included a selection of three different deep learning based text to speech modules.

The last part of the pipeline has been only previously accessible for animations which are designed to be manipulated dynamically. It hasn’t been possible, until recent advances in GPU hardware, such that we could use deep learning based approaches to learn realistic, so called ”talking head videos”.

3 Chatbot

In this current pipeline, the openAI gpt-2 was used. But of course any chatbot can do.(Radford et al., 2019)

4 TTS

A common approach to synthesize voice from text using deep learning has been to split the process into two stages. First, predict the mel-spectralgram, or any intermediate latent representation, of the desired resulting audio from text. Second, use a spectralgram to audio

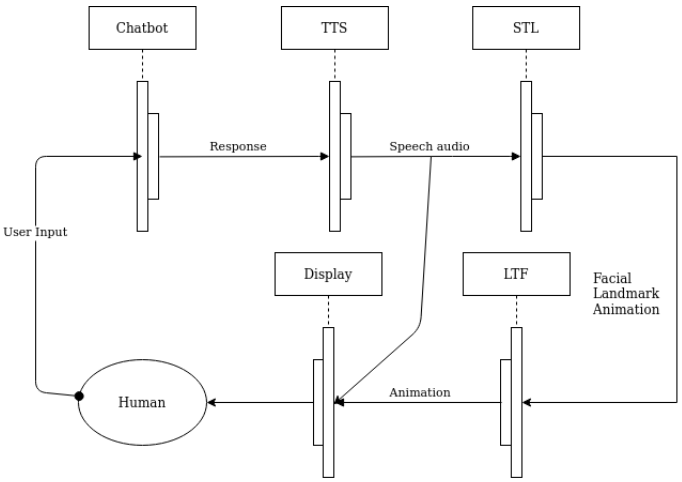

Figure 2: Diagram of the pipeline.

model, such as Wavenet, to construct the voice.(Shen et al., 2017)

This approach has the advantages of being modular, having high quality of sound, and fast inference time.

The current state of the art models of the first stages of the TTS process include Tacotron2(Shen et al., 2017), and Fastspeech2(Ren et al., 2020), and the second stage include Wavenet(van den Oord et al., 2016), Melgan(Kumar et al., 2019), and their variants. For this pipeline I chose Tacotron2 and Multi-Band MelGan(melgan-stft)(Yang et al., 2020).

4.1 Voice Cloning

Synthesizing specific voices is important. So there are two ways of doing that with the current pipeline. One is to fine-tune the vocoder (text to mel-spectralgram) to a specific voice. As an example I scraped 15 minutes worth of Elon Musk speaking, and trained the vocoder(as well as the synthesizer, actually) of the dctts model. The dctts model(Tachibana et al., 2018) is convolution based, and much cheaper to train than the Tacotron, giving less time and sample needed, but suffering in output quality.

But perhaps a much more efficient way is to invest in few-shot cloning of people's voices. There are luckily also many open source projects available currently which can provide a demo of what a desired outcome might look like. Specifically I am using 5 second voice cloning(Jia et al., 2018).

Figure 3: End to End demo Demo Video

5 Speech to facial landmark

Speech to facial landmark could be reduced to the problem of: from speech data (2 dimensions), to mouth landmark (Around 20 points). I think it has been shown that a vanilla Recurrent Neuronetwork or a LSTM can do just fine.
I haven't yet trained any models to do that, since I just wanted to put the pipeline together first, so currently I am using a whole face facial landmark prediction thing, and then just using the mouth movements.

For this pipeline I am using either 'End-to-End Speech-Driven Facial Animation using Temporal GANs' (Vougioukas et al., 2018) or 'Generating Talking Face Landmarks from Speech' (Eskimez et al., 2018). The approach from 'End-to-End Speech-Driven Facial Animation' could potentially replace the next part of the pipeline. But I am not aware of any current approaches which produces both high fidelity as well as speaker identity preserving results. So I am extracting a facial landmark movement from one which produces high quality generated results and then using it to animate my faces as a driving video.

Figure 4: From facial landmark to face animation(First Face: An anonymous contributor to the GRID dataset, Second Face: Shane Mulligan, Third Face: Elon Musk)

6 Facial landmark to rendered Face

For this I am using the first order model(Siarohin et al., 2019), which generates a driving video and a single reference image a facial animation. Perhaps using something like pix2pix(face2face) which goes directly from facial landmarks to face movements would have been better, but I was able to get realistic blinking and pupil movements from this way.

6.1 Pupil Movement

I was able to control to the pupil movement via manipulating an overlay of a fake eye over the actor via facial key points. This causes a lot of jitter and can be improved.

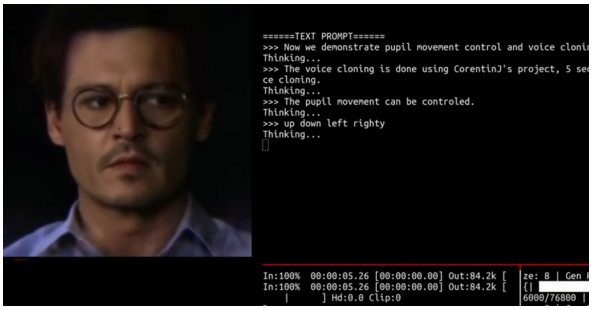

Figure 5: Demo for pupil movement. Demo Video

7 Plans For Improvement

1.The Text To Speech part of the pipeline ideally should be a faster, and more reliable version of 5 second voice cloning(Jia et al., 2018).
2.The speech to land mark part of the pipeline should be just a simple recurrent neuronetwork which predicts only the mouth landmark movements.This will significantly reduce inference time and quality.
3.The face rendering should be less jittery, and support higher resolutions. A better interpolatable latent space should also be constructed to enable more capability to manipulate the face. Such as face turning, nodding etc. 4. Connecting the pipeline as a end to end model. At least eliminate the speech to landmark part and train speech to latent space manipulation directly.

8 Speed of the pipeline and Hardware requirement

Testing using a nvidia RTX 2080ti GPU. Since the rendering is realtime, the part of the pipeline which the speed should be measured is the text to speech and speech to landmark module.

Performance on Nvidia RTX 2080ti GPU		
Utterance	Words	Time taken(sec)
Test test test test test test test test test test test test test test test test.Test test test test test test test test test test test test test test.	26	3.26356959342956
Test test test test test test test test test test test test test.	13	2.04077553749084
Test test test test test	5	1.40281891822814
Hello there.	2	1.05777335166931

9 Deploying on the web

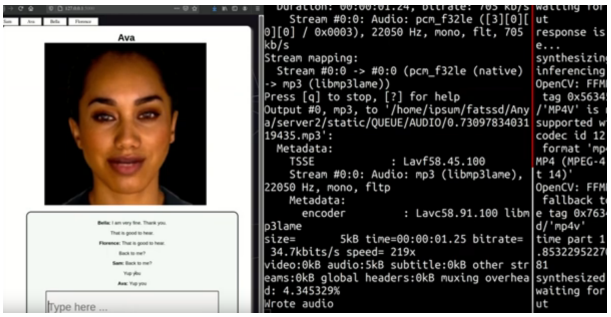

Figure 6: Text prompted speech and talking sequence generation demo. Demo Video

10 My interest

I'm mostly interested in general artificial intelligence research.The approach from Soul Machines to study the way infants learn interests me very much. I believe that understanding the process of which humans take the first step and obtain the base semantics of the world is critical to approaching general intelligence.

References

Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In *Latent Variable Analysis and Signal Separation*, pages 372–381, Cham. Springer International Publishing.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2018. End-to-end speech-driven facial animation with temporal gans.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2020. Multi-band melgan: Faster waveform generation for high-quality text-to-speech.