University of Hertfordshire
School of Computer Science

7COM1086 Artificial Intelligence and Robotics Masters Project

September 2021

Re-TTAC: Realistic Text-driven Talking-face with Audio-visual Cloning

Aron Samuel Georgekutty
19001035
Supervisor: Dr. Manal Helal

Abstract

This project explores the possibility of developing a face reenactment model from a text input with realistic audio-visual reproduction capabilities. The recent advancements in voice cloning and talking-face generation are studied, and the existing state-of-the-art models are borrowed to construct a pipeline that enables us to do so.

Acknowledgement

I would like to express my profound sense of gratitude to Dr. Manal Helal for her constant support and invaluable guidance. I would like to thank the University of Hertfordshire and the School of Computer Science for the opportunity to explore deep learning as my final year project.

I would also like to thank my parents for their patient proof reading.

Table of Contents

1. Introduction

Over the last two decades, deep learning has been making headlines in both industrial and domestic sector, but gone are the days when people were amazed by the preview of a state of the art model without a demonstration of its immediate real-life applications. Manipulation of audio-visual data using machine learning is not something new as there are plenty of associated statistical methods that we have been using for decades, but the deep learning models have opened doors to some of the techniques we never thought were possible. According to Zhu et al. (2020), there are four main categories of tasks in audio-visual learning, and they are, separation & localization, correspondence learning, audio-visual generation and representation learning. Although we will be indirectly using all of these four categories, our focus will be on audio-visual generation and correspondence learning. The most popular models that have made waves from these categories are Deepfakes (Nguyen et al. 2021) and Face2Face (Thies et al. 2019), which leverage powerful techniques from Generative Adversarial Networks (GANs) to automate realistic facial reanimation. Deepfakes are often exploited with their ability to make realistic fake videos that have given them an infamous reputation in social media, which has, in turn, pushed the researchers to develop better detection models (Pan et al. 2020). However, these methods have a major limitation, which is that they have to be continuously driven by another video, which could either be a live webcam capture, or a stored media. There are methods that uses audio to drive a talking face (Yi et al. 2020) (Wu et al. 2021) (Eskimez, Zhang, and Duan 2021), but all of them use a pre-recorded voice to drive the talking face. Since speech recording requires human intervention and necessary equipment, it complicates the process of generating a talking face even after having the script ready.

The need for a completely text-driven video reenactment model is the chief motivation of this project. This model finds a wide range of applications where a similar scene has to be shot or played back several times with different dialogues, like news reading, virtual anchors or role-playing video games. For this we would require a state of the art Text to Speech (TTS) system that is able to learn the voice of the source from the input video, reproduce the input text in their own voice and finally drive the same video with the new voice, which should look and sound as natural as the input video. The first part requires a voice cloning model that works in real-time. Although Jia et al. (2018) were able to make a model using their Speaker Verification to multispeaker TTS (SV2TTS), there were no publicly available implementation of the same, until Jemine (2019) has come up with one. The second part is an existing model (Yi et al. 2020) that uses a pre-recorded voice to drive a talking face, with their implementation using PyTorch available on GitHub. We take these two models and try to concatenate them to make a unified pipeline that achieves our goals.

In this project, we aim at recreating a movie scene (or any audio-visual data with a human subject and their voice) by providing a new dialogue as text input. The resulting video should not only look natural, but also should be able to mimic the audio-visual nuances of the subject from the input video. As we have real-time implementations of both the voice cloning and talking-face generation systems, our unified model should also ideally work in real-time.

As our model depends on voice cloning and talking face generating, we take a look at the state of the art papers in these areas, one by Jia et al. (2018) and another by Yi et al. (2020). The first one is a TTS system that uses a speaker verification model for transferring the learning. It uses Google's WaveNet and Tacotron, which will be explained in detail later on in this report. The second one is a talking face generator that takes an audio signal from a source person and drives a target person with personalized head pose and lip-sync. Based on these papers, our objectives are defined as follows:

1. Study the advancements made in the field of TTS and realistic talking face generation.

2. Find an implementation of SV2TTS by Jia et al. (2018) to clone the voice of the person in the input video.

3. Refine the audio-driven approach by Yi et al. (2020) to work with a short audio-visual input.

4. Construct a unified pipeline that seamlessly concatenates the above two methods to achieve a realistic text-driven audio-visual model.

2. Literature review

In this chapter, we look at the recent advancements in voice cloning and talking face generation, along with their background knowledge. As it is machine learning and signal processing that paved way for synthesizing realistic voices and images with ease, we will go through the basic concepts behind image processing and audio processing, followed by some of the audio-visual processing tools and techniques that we will be using for this project. We also take a look at two data sets

2.1 Background

2.1.1 Image processing

As images as stored as matrices of variable dimensions in computers, and videos are nothing but collection of image frames that are played with a specific frame rate, a variety of matrix manipulating techniques combined with ANNs can pave way to advanced image processing and computer vision techniques. Conventional neural networks are cursed with the dimensions when it comes to image processing as image matrices exponentially increases the calculations required when they are directly used inside the network. Hence the birth of Convolutional Neural Networks (CNN), which looks similar to ANNs, but they are optimized to work with images. They use filters (or kernels) to perform convolutions on the images, which not only makes them compact, but retains only the crucial information. Once trained, an optimized model can work so fast that any computer or modern smartphones with an average specifications (1 Ghz CPU and 2GB RAM) can do object detection tasks in real-time. COCO (Lin et al. 2015), Pascal VOC (Everingham et al. 2015), ImageNet (Deng et al. 2009) and CIFAR-10 (Krizhevsky 2012) are some of the most popular datasets that are used for object recognition tasks. MobileNets (Howard et al. 2017) and YOLO (‘You Only Look Once’) (Redmon et al. 2016) are two of the lightest models for object detection, and because of their speed and reliability, their iterations are used in autonomous navigation and self-driving cars.
Although image processing tasks have reached a reputable standard, video processing on the other hand has a lot more to advance, as it requires spatio-temporal analysis along with observing individual frames. It also has some additional scopes like object tracking and action classification, which would

not be possible with single image frames. Object tracking mainly uses optical flow (Fischer et al. 2015), which is the computing of pixel shift between frames after locking on to a target's coordinates using object detection. Action classification, in addition to optical flow, also uses pose estimation (Zheng et al. 2021), which includes the detection of key points on the body, much like identifying the facial landmarks, and their classification and tracking.

Machine learning models have further advanced now that we are not limited to the processing of existing audio and videos, but are also capable of generating images and sounds that never existed before. Introduction to Generative Adversarial Networks (GAN) and improvements in 3D modeling are the key reasons this has become possible. Let us look at how these two techniques work.

2.1.1 (a) Generative Adversarial Networks (GANs)

They are a deep-learning based generative models introduced by Goodfellow et al. (2014), which uses unsupervised learning to automatically discover and study the patterns in the training data and generate samples that 'look-like' as if they were from the training data. As shown in fig 1, it consists of two sub-models: a generator and a discriminator. A generator accepts a random vector from the input vector space using a Gaussian distribution and uses it as a seed to generate an output in a multidimensional vector space called latent space, and it corresponds to the input space. Upon training, the latent space converges more and more to the input space, and when it reaches an adequate accuracy, the training process is paused and the generator is stored, which will later be used to generate new samples. A discriminator on the other hand works like a classifier, where it takes input from the original data set or the ones generated by the generator, and classify them as real or fake (generated).

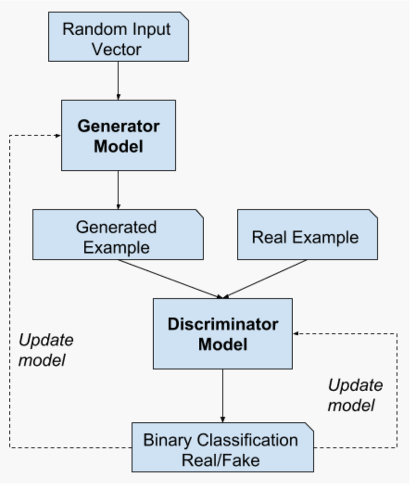


Figure 1: General Architecture of GANs

Although the machine knows if the input fed to the discriminator is generated or real, the discriminator is not informed about this as it takes the input data for classification randomly. This gives the generator an advantage of producing data that looks very similar to the input and hence 'fool' the discriminator. In GANs, the generator and the discriminator are trained together, where the former generates new samples and mixes them with some real samples and are fed to the latter. After classification, the discriminator is told about the true label of the data, which helps it to update itself. Likewise, the generator is also told about the classification done by the discriminator. This forces both the models to compete against each other in what is called a zero-sum game (Guan, Zhang, and Tsiotras 2021), which is a minimax algorithm in game theory. Either of them performing better means that the other is not performing well, and hence to attain an equilibrium, the discriminator should become totally ‘unsure’ about the input it gets. This would mean that the generator has started generating samples that are very close to the training data. Once this ideal threshold value is obtained with some approximation, the training is stopped and the generator model is saved for data generation tasks.

2.1.1 (b) 3D Morphable model (3DMM)

It is used to recover the 3D facial geometry using a 2D image of the face. According to Ferrari and Berreti (2018, p. 326), the appropriate training data should have a sufficient variability in terms of ethnicity, gender, age, so as to enable the model to include a large variance in the data. Even though optical-flow methods were used to provide a dense vertex by vertex alignments between each scanned 3D images in the original work by Blanz and Vetter (2003), it did not have the capability to accommodate facial expressions, which is crucial for a natural appearance of the reconstructed face. To accommodate this feature, Ferrari and Berreti (2018) has come up with Dictionary Learning based 3DMM (DL-3DMM) as shown in fig 2, which was obtained by the landmarks detection on the face their partitioning into a set of non-overlapping regions. This method was used in the Binghamton University 3D facial Expression (BU-3DFE) database (Yin et al. 2006) and exhibited a robust framework with large expression variations.

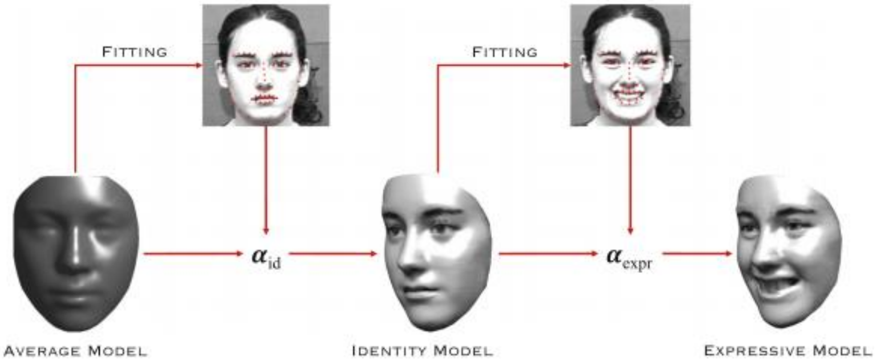


Figure 2: Workflow of the proposed DL-3DMM method with expression-specific deformation (Ferrari and Berreti, 2018, p. 329)

2.1.2 Audio Processing

Audio signals, as opposed to image data are analog in nature, which needs to be converted to digital format to be stored and understood by the computer. Although most of their analysis are done after applying necessary Fourier transformations (Settel et al. 1994), which converts them to wave functions, so that we can apply existing signal processing methods to study them. Of the many possible applications of machine learning in audio processing, voice recognition and speech generation are the the most explored ones. According to Purwins (2019), Mel-frequency Cepstral Coefficients (MFCC) are the most common representation in traditional audio signal processing, because it is modeled based on how a human ear perceives voice. We take a look in detail how MFCCs work.

2.1.2 (a) Mel-Frequency Cepstral Coefficients (MFCC)

The first step is to convert the analogue audio signals to digital format with a specific sampling frequency. After this, a pre-emphasis is done to amplify the magnitude of the higher frequency range, as its energy is considerably low when compared to the energy of the lower frequencies. The next step is windowing, as MFCC is used to detect phones in the speech, the audio signal should be sliced into segments of 25ms width and 10ms step size. This is done with the assumption that an average person speaks about 3 words per second, with each of the word containing 4 phones, and each of them having 3 states, summing up to 36 states per second or 28ms per state, which approximates to our window of 25ms. Now, to do spectral analysis on the audio slices, a discrete Fourier transform is applied on them (Jurafsky and Martin, 2008).

Human ears can differentiate sounds at lower frequencies better than those at higher frequencies, whereas the machine has a constant resolution at all frequencies. Since it is observed that mimicing the human ear property will yield better results when analyzing speech, a mel scale is used to map the frequency to a logarithmic curve as given below (Taylor, 2009).

$$mel(f) = 1127\ln(1+\frac{f}{700})$$

This is called the Mel-filter bank, and the resulting frequency goes through an inverse Fourier transform. This is because in human voice, each sound is attributed to the unique position of the tongue and other articulators, which selectively dampens and amplifies the frequencies of the vibrations produced by the glottis in the wind pipe, which are basically multiples of a fundamental frequency.
Ladefoged and Maddieson (1996) have studied the transfer function of the vocal cavity for different phones and observed that inverting the time and frequency domain after the transformations will result in the fundamental frequency having the highest frequency in the time domain (fig 3).

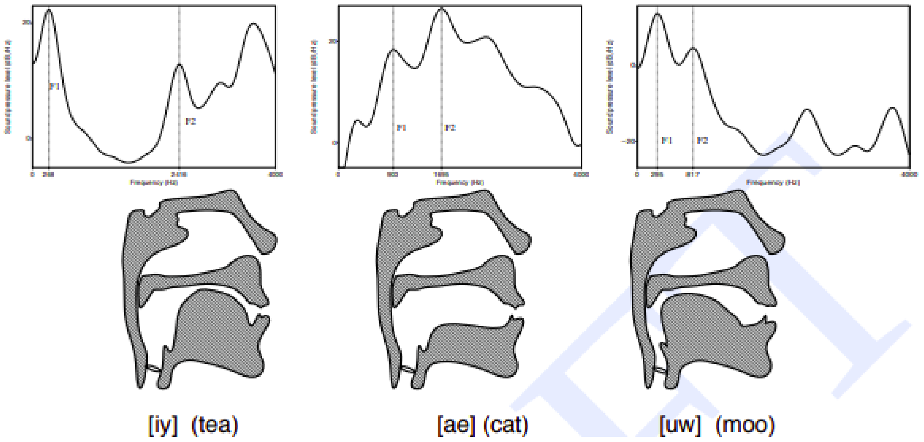


Figure 3: Positions of the tongue for three English vowels, high front [iy], low front [ae] and high back [uw]; tongue positions modeled after Ladefoged (1996)

This inverse Fourier transform of the log of a signal is called a cepstrum, and the first 12 coefficients of signal, along with the energy of the signal sample are taken as the 13 primary features of MFCC. The secondary features includes the first and second order derivatives of these features, which adds another 26 features. This makes a total of 39 features that MFCC technique generates for each audio sample.

2.1.2 (b) Mean Opinion Score (MOS) for voice quality

MOS score is the most popular metric used to measure the voice quality, be it the one during a voice call or a synthesized voice. Standardized by the International Telecommunications Union, MOI was originally developed for traditional calls, but now it is widely used for Voice over IPs (VoIPs) too.
It is calculated as follows.

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

Where R(i) represents an individual rating and N is the total number of subjects who have given the rating. Each individual rating R(i) has a maximum of 5 (good) and a minimum if 1 (bad), and depends on three metrics, listening quality, transmission quality and conversational quality. It is also affected by physical factors such as the hardware, bandwidth of transmission, jitter, latency, packet loss and codec version. Codec version has the most impact, as it decides how well the audio is reproduced by the Digital to Analog Converter (DAC), depending upon its compression ratio. For our project, we will be using Waveform Audio File Format (WAV) for all the audio files, as it does not compress the original analog audio. Developed by IBM and Microsoft, WAV delivers lossless audio with very high sample rate and bit depth.

2.2 Overview of the tools and techniques used

Our project aims at developing a talking face with voice cloning, for which we require a background knowledge about the conventional and state of the art tools in practice. We look at the following topics for an understanding about the same:

2.2.1. Text-to-Speech (TTS)

Speech synthesis, which is the reproduction of human speech by artificial methods, mainly from text, has become increasingly popular with their wide range of applications including voice assistants, like Amazon’s Alexa or Apple’s Siri. Although their voice sounds natural, the race for creating a realistic text-to-speech model is still evolving. Text-to-Speech methods are broadly classified into two, Statistical Parametric Speech Synthesis (SPSS) based and Concatenative synthesis based. In SPSS (Aroon and Dhonde 2015), a statistical generative model learns the relationship between the graphemes of the input text and the acoustic features in their respective phonemes, and it is called the acoustic model. A vocoder then synthesizes the speech using these acoustic features by reconstructing an audio waveform from them. Fig 4 shows the general SPSS pipeline.

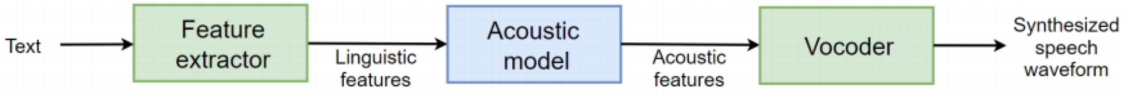


Figure 4: General SPSS pipeline.

Meanwhile, concatenative speech synthesis (Oloko-oba, T.S, and Samuel 2016) works by concatenating actual instances of speech from a database. It is based on a time-domain joining algorithm that selects pre-recorded human utterances and smoothly concatenates them. Although it produces very natural sounding speech, it also requires a large database of voice recordings and is limited to the user in those recordings. Apple's Siri uses concatenative speech synthesis and Fig 5 shows the concatenative synthesis of an activation command.

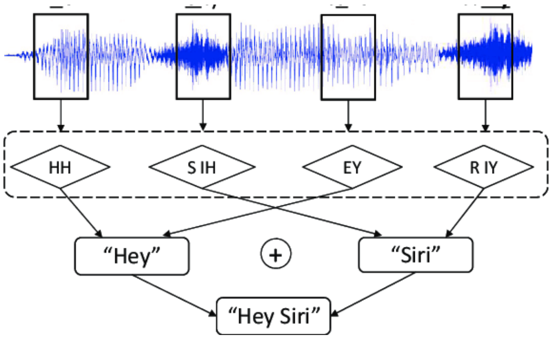


Figure 5: How concatenative speech works

A breakthrough in the field of TTS was made by Google's WaveNet (Oord et al. 2016), which is a fully probabilistic and an autoregressive model, and it was able to produce state of the art TTS in both English and Mandarin. Being an audio generative model based on PixelCNN (Oord et al. 2016), it was capable of producing a very human-like voice. The authors have compared this to state of the art TTS methods using SPSS (Zen et al. 2016), Concatenative method (Gonzalvo et al. 2016) and human speech as shown in Fig 6. The mean opinion scores were obtained from blind tests, taken from over 500 human subjects and 100 test sentences.

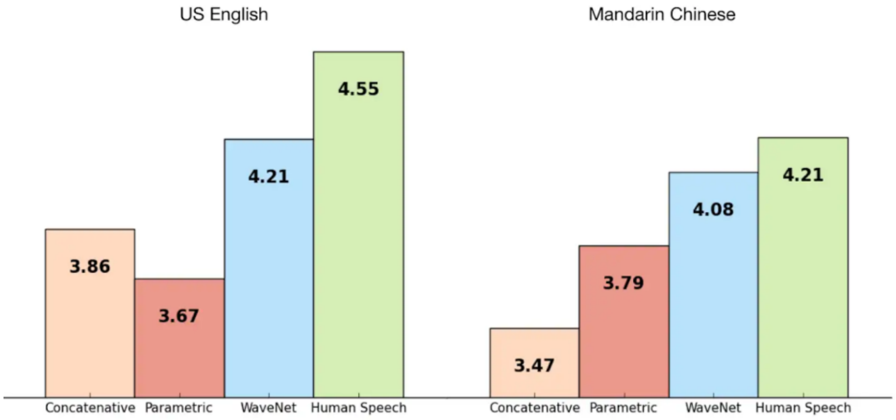


Figure 6: MOS scores of WaveNet (Oord et al. 2016)

A year later, Google produced Tacotron (Wang et al. 2017), an end-to-end generative TTS model which synthesizes speech from text and audio pairs using a sequence-to-sequence model, comprising of an encoder, attention based decoder, and a post-processing network. A modified version called Tacotron 2 (Shen et al., 2017) was released the same year, which used a modified WaveNet as vocoder. It was able to achieve an impressive MOS of 4.53, breaking all the records in the history of TTS. In fact, it was on par with human speech according to a study done by the authors invloving 800 human subjects with 100 different test sentences and the results are shown in Fig 7.

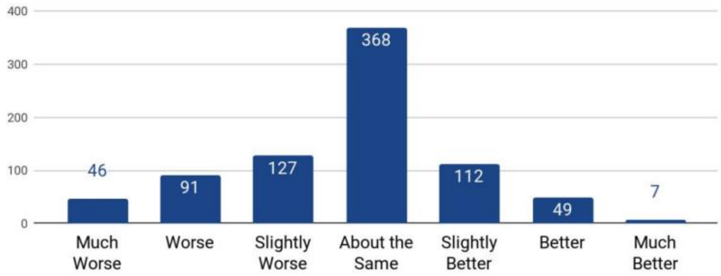


Figure 7: MOS evaluation of Taccotron 2 (Shen et al., 2017)

Some of the parallel works in TTS were Baidu’s three Deep Voice models, Deep Voice 1; Real-time Neural Text-to-Speech (Arik et al. 2017), Deep Voice 2: Multi-Speaker Neural Text-to-Speech (Arik et al. 2017) and Deep Voice 3: Scaling Text-to-speech with Convolutional Sequence Learning (Ping et al. 2018). Researchers from Facebook AI also caught up on this race with their Voiceloop (Taigman et al. 2018), which is inspired from a phonological loop that holds verbal information for a short time. Tacotron 2, still being the state of the art in TTS, we will be using it for the construction of synthesizer and vocoder in our project.

2.2.2 Voice cloning

Mimicking the voice of an existing user still stands as a challenge to machine learning enthusiasts, however, with the advent of advanced deep learning methods, there has been various projects that have made waves with their text-to-speech models. One such example is the one by Baidu Research, where they have have put forward a Neural Voice cloning model (Arik et al. 2018), which is based on Deep Voice 3 and uses only a few audio samples to synthesize a person’s voice. Speaker adaptation and speaker encoding were the two methods that were adopted in the paper. The former fine-tunes a pre-trained multispeaker model with a few auidio-text pairs from an unseen speaker, and the latter directly estimates the speaker embedding from an unseen speaker’s audio samples. The current state of the art is Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (Jia et al. 2019), which we will be using in this project. It uses a framework for zero-shot voice cloning that only requires five seconds of reference speech. The noises in the resultant speech audio is rectified using WaveRNN (Kalchbrenner et al. 2018) and the training is made seamless using generalized end-to-end loss (Wan et al. 2020).

2.2.3. Talking head generation

Since the advent of deepfakes, talking heads were a default go to tool used to showcase the capabilities of GANs (Goodfellow et al. 2014) , which were either video driven (Thies et al. 2019) or audio driven (Zhang & Weng 2020). Among them, the former found great success in the social media platforms, as they were widely used in the face filters on various apps. This success can be attributed to the fact that the model used the facial landmarks from the user’s face to drive the talking head, which looks natural with corresponding head movements and lip syncing. The audio driven however, had to overcome a lot of challenges, mainly the 3D face reconstruction, pose estimations and lip syncing. Chung, Jamaludin, and Zisserman (2017), in their paper ‘You said that’, discusses a novel algorithm that uses a voice sample to model a talking face with lip-syncing was done using a single image. Although this yielded promising results, due to its lack of head movements and expressions, it was not so popular.

2.2.4. 3D face reconstruction and generation using GANs

3D face reconstruction aims at rendering a 3D model of the face using a single (Guo et al. 2021) or multiple 2D face images (Li et al. 2017). Sometimes an added feature of depth is also used to create 3D face models from the RGB-D data (Bouaziz et al. 2013). Most of these methods a backed up by 3D Morphable Models (3DMM), which is a learning-based method that uses PCA on the face data to represent them as Eigenvectors in the newly formed “face-space”(Eigenfaces). Due to lack of sufficient real face data, some methods use synthetic training data (Richardson, Sela, and Kimmel 2016). Video driven face generation GANs mostly use Face2Face (Thies et al. 2019), which is an adaptation of Pix2Pix (Isola et al. 2018), that uses conditional adversarial networks for image to image translation.

Kim et al. (2018) presented a novel approach to generate photo-realistic videos from 3D rendered face images using GANs, but it was limited to specific target persons and required thousands of samples for training. Additionally, face reenactment methods requires a person on the other end to continuously drive the rendered face and hence the reconstructed video can only be as long as the original input video, provided there is no looping. Olszewski et al. (2017) used a multi-linear PCA model to drive a face from a single RBG image using source video sequence. Memory networks are widely popular for their ability to augment neural networks using external memory. It makes an effective few-shot or even one-shot learning as it can remember crucial information selectively. As Yi et al. (2020) achieved realistic audio-driven talking head using this method, we will be using the same in this project.

2.3 Similar works

Our project looks at two papers and combines them to develop the model. One by Jia et al. (2019), which is multi-speaker speech synthesis model that uses a network-based system for TTS synthesis even from a speaker who is unseen during the training phase and another one by Yi et al. (2020), which is on generating a realistic talking face from just the audio input.

2.3.1 Transfer Learning from Speaker Verification to Multi-speaker Text-To-Speech Synthesis (SV2TTS)

We look at the state of the art method in voice cloning (Jia et al. 2019), which is a multi-speaker speech synthesis model that uses a network-based system for TTS synthesis even from a speaker who is unseen during the training phase.

2.3.1(a) Model architecture

The model consists of three components as shown in fig 8, a recurrent speaker encoder , a seq2seq synthesizer, and an auto-regressive vocoder, and are based on three previous works from Google, the G2E loss (Wan et al., 2017), Tacotron 2 (Wang et al., 2017) and WaveNet (Oord et al., 2016) respectively. The encoder is fed a speech signal and it computes a vector of fixed dimension and the corresponding grapheme or phoneme is given as the input to the synthesizer. The synthesizer predicts a mel spectrogram based on this input and is fed to the vocoder, which synthesizes a speech waveform.

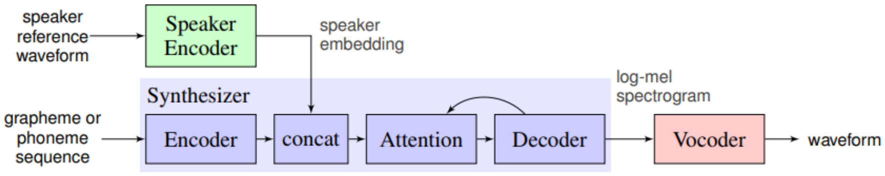


Figure 8: Model overview of the SVTTS. (Jia et al. 2019, p. 3).

(i) Speaker Encoder: It maps the audio of the speaker into a low dimensional vector embedding. As it is independent of the context of speech and background noise, it is a good generalization of the features in a speech utterance. The authors of this paper prefers G2E loss (Wan et al., 2017), which is a highly scalable and accurate framework for speaker verification. It takes a speech utterance as input, computes its log-mel spectrogram frames, and then maps it to an embedding vector of fixed dimension. G2E forces the speech utterances from the same speaker to have a high cosine similarity and those from different speakers to be distant in the embedding space.

The model used for encoder is a 3-layer LSTM with 768 hidden nods, which are followed by a projection layer of 256 units. An L_2-normalization of the output layer of the final frame gives the final embedding. An arbitrary length of utterance, broken down into 800ms windows is overlapped by 50%

during the inference. Although the network does not learn to capture relevant speaker characteristics directly, the training during speaker discrimination task enables it to achieve speaker verification.

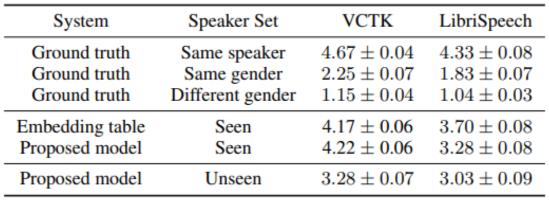
(ii) Synthesizer: The authors have extended the sequence-to-sequence model of Google's Tacotron 2 (Wang et al., 2017) to accommodate multiple speakers following a scheme similar to Deep Voice 2 (Arik et al. 2017). At each time step, an embedding vector for the target speaker is concatenated with the synthesizer encoder output. The authors compare two variants of Deep Voice 2, one that uses the speaker encoder to compute the embedding, and a baseline that optimizes a fixed embedding for each speaker from the training set. They have trained the synthesizer on pairs of text transcript and target audio. They have observed that mapping the text to a sequence of phonemes would lead to a faster convergence and hence improved pronunciation of proper nouns and rare words. A pre-trained speaker encoder is used for training the network in a transfer learning configuration to extract a speaker embedding from the target audio signal. That is, the speaker reference signal and target speech are kept the same during training.
Using a 50ms window and a step size of 12.5ms, the target spectrogram features are computed. This is then passed through Mel-filter bank of 80 channels and compressed with a log dynamic range. The authors have extended the Tacotron 2 architecture by augmenting the L_2 loss on the predicted with an additional L_1 loss. They have also found that this combined loss was more robust on a noisy training data. Unlike Kaliouby and Robinson (2005), they have not used an additional loss for speaker embedding.

(iii) Neural Vocoder: The authors have adopted the sample-by-sample auto-regressive WaveNet (Oord et al., 2016) to use it as their vocoder. It works by inverting the synthesized Mel spectrograms that are emitted by the synthesis network to be converted into time-domain waveforms. The architecture remains the same as in Tacotron 2, with 30 dilated convolution layers. The network however is not directly conditioned on the output from the speaker encoder.

2.3.1 (b) Results

Speaker similarity and naturalness of speech were the two parameters the authors used to evaluate their model. They have used the Mean Opinion Score (MOS) as the metric since most voice quality testing methods use the same. They have used LibriSpeech and VCTK as the datasets for their training.

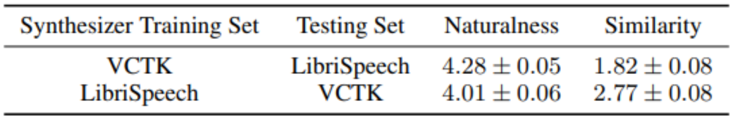
LibriSpeech is corpus of more than 1000hrs of 16Hz read English speech (Panayatov and Povey), and VCTK is a speech data corpus with utterances by 110 English speakers with each speaker speaking about 400 sentences.



System	Speaker Set	VCTK	LibriSpeech
Ground truth	Same speaker	4.67 ± 0.04	4.33 ± 0.08
Ground truth	Same gender	2.25 ± 0.07	1.83 ± 0.07
Ground truth	Different gender	1.15 ± 0.04	1.04 ± 0.03
Embedding table	Seen	4.17 ± 0.06	3.70 ± 0.08
Proposed model	Seen	4.22 ± 0.06	3.28 ± 0.08
Proposed model	Unseen	3.28 ± 0.07	3.03 ± 0.09

Figure 9: MOS scores on speaker similarity

For the similarity test, an evaluation set of 100 phrases that are unseen during the training were used and the results are shown in fig 9: and a cross data evaluation was also done for both the similarity and naturalness as shown in fig 10:



Synthesizer Training Set	Testing Set	Naturalness	Similarity
VCTK	LibriSpeech	4.28 ± 0.05	1.82 ± 0.08
LibriSpeech	VCTK	4.01 ± 0.06	2.77 ± 0.08

Figure 10: MOS values of Naturalness and similarity after cross data evaluation

2.3.2 Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose

This paper aims at generating a realistic talking face that transfers an audio signal from a source person to the visual information of a target person. Since most of the existent methods consider only lip movement with fixed head pose, this paper focuses on making the generated face look as natural as a real-world talking face with accompanying head movement. Wang, Enescu and Sahli (2013) observed that the natural head movement also has a significant impact on the quality of communication. Humans are sensitive to subtle head movements in videos and hence get easily uncomfortable while talking with a fixed head pose. Although Busso et al. (2007) were able to establish a correlation between head pose and speech, it was not until recently that we were able to predict head pose correctly from a given speech signal (Greenwood, Laycock, and Matthews 2017), but this solution however was not so practical. A year later, the same authors have developed a joint learning method (Greenwood, Laycock, and Matthews 2018) that infers the facial activity from the audio first and then models the head pose from those facial features. This paper however simultaneously infers facial expression and corresponding head pose from the speech.

A major challenge in synthesizing a high-resolution and realistic talking face with a smooth background transition is the in-plane and out-of-plane head rotations that happens naturally while talking. This paper takes care of this problem by including a short talking face video of the target person along with the associated audio. Then, both this information is used to lean the personalized talking behavior of the target person, which includes the movements of the lips, eyes, brows and the overall head pose. In order to synthesize a high quality video output, the model reconstructs 3D face animation from slicing and cropping the short input video into multiple image frames of the target person's face, and re-renders it in high resolution video frames. However, since this is not done using a dedicated 3D engine, and only limited image information is available, the rendered images are far from realistic ones. To remedy this problem, the paper proposes a novel memory-augmented GAN module, which uses the target person's face from input video frames to refine the rendered frames into more realistic ones with smooth transitions. The authors claim this to be the first ever model that can transfer audio from an arbitrary source to an arbitrary target person, as opposed to a similar work in the past (Suwajanakorn, Seitz, and Kemelmacher 2017) that could only work with a specific person (Obama).

2.3.2(a) Model architecture

The model has two stages as mentioned in Fig 11:

1. 3D facial animation from audio-visual information: For training they have used the LRW dataset, which is a corpus of 1000 utterances of 500 different words, which are spoken by different speakers, with each video of 29 frames length (1.16s). Using the LRW video dataset, a general mapping is trained between the audio speech and the facial expression including common head pose. After the 3D face reconstruction, this mapping is fine-tuned with the input audio signal and short video to create a personalized talking behavior of the target person.
2. Generating realistic talking face from the 3D facial animation: As the texture and lighting information from the input video cannot do much during the rough rendering by the light-weight graphics engine, a memory-augmented GAN, which was also trained using the LRW dataset is used to refine the rendered frames.

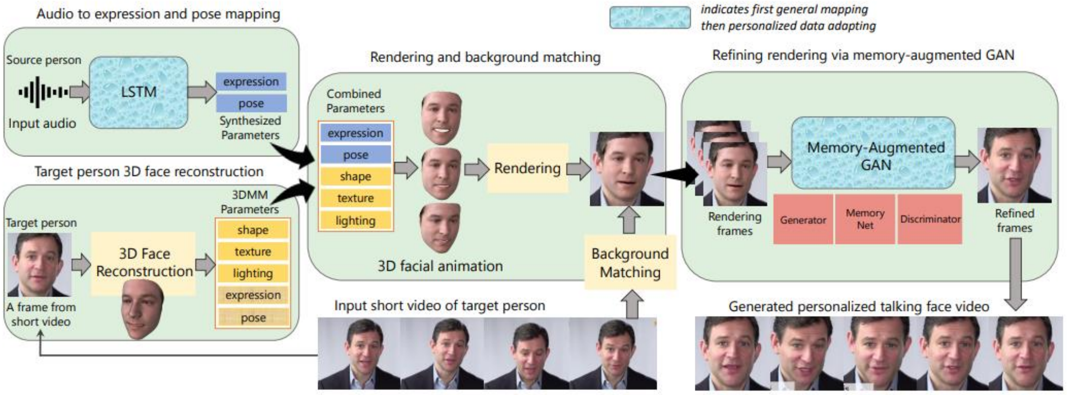


Figure 11: Fig: Model architecture by (Yi et al. 2020)

Now, let us look at each of these steps in detail.

3D face reconstruction

The method proposed by Deng et al. (2020) is used for the 3D face reconstruction. It takes a 2D face photo as input and deploys a parametric model of the 3D face Geometry, which is later fit in using a CNN. It also retains the original texture and the illumination conditions from the input photo after the 3D reconstruction.

With the input face photo as I, the method reconstructs the 3DMM coefficients X(I), which is given by

$$X(I) = \{\alpha, \beta, \delta, \gamma, \rho\} \ \varepsilon \ R^{(257)},$$

where,

$\alpha \ \varepsilon \ R^{(80)}$ is the coefficient vector for face identity,

$B \ \varepsilon \ R^{(64)}$ is for the facial expression,

$\delta \ \varepsilon$ R^(80) is for the facial texture,

$\gamma \ \varepsilon$ R^(27) is the coefficient vector for illumination, and

$\rho \ \varepsilon$ R^(6) is the pose vector for rotation and translation.

Now, the face shape S, is represented as,

$$S = S' + B(id)\ \alpha + B(exp)\ \beta,$$

and the face texture T, is represented as,

$$T = T' + B(tex)\ \delta,$$

where,

S' and T' are average shape and texture,

B(id), B(exp) and B(tex) are the principal components of the shape, expression, and texture respectively. B(id) and B(tex) uses the Basel Face Model (Paysan et al. 2009) and B(exp) uses FaceWareHouse (Cao et al. 2014). Lambertian surface assumption is used for the computation of illumination and spherical harmonics basis functions are used to approximate it to the original frame (Ramamoorthi and Hanrahan 2002).

Audio to pose and expression mapping

The existing talking head models only focus on the lips and the jaw, which looks unnatural as a realistic talking face involves head movements and changes in eyes and brows. The input video is analysed to map changes to the overall dynamic elements on the face by syncing the audio information with the 3D geometry. As observed by Yi et al. (2020), the speaking style in a short period of time is consistent with the audio and pose. The first step is to extract MFCC features from the cloned voice and use it to model the expression and pose using 3DMM coefficients. An LSTM network is designed to establish a mapping between then as follows:

Let S={s(1),…,s(T)} be the MFCC features the the cloned audio sequence, B={b(1),…,b(T)} be the corresponding ground-truth expression coefficient sequence, and P={p(1),…,p(T)} be the ground-truth pose vector sequnce. A predicted expression coefficient sequence and pose vector sequence are generated as B'={b'(1),…,b'(T)} and P'={p'(1),…,p'(T)} respectively. The mapping from audio to pose and expression can now be formulated as:

$$[\ B'(t), p'(t), h(t), c(t)\ ] = R(\ E(\ s(t)\ ), h(t-1), c(t-1)\ ),$$

where,

R is the LSTM network,

E is an additional audio encoder,

h(t) is the hidden state of the LSTM, and

c(t) is the cell state of the LSTM.

Rendering and background matching

With the extracted 3DMM coefficients and audio to pose-expression mapping, a face image sequence is rendered using the rendering engine by Genova et al. (2018). Albedo, a measure of overall brightness of the rendered object by calculating the proportion of incident light that is reflected away from a surface, is used by the authors to analyse the rendered 3D face. The albedos computed from the 3DMM coefficients were too smooth and of low-frequency, which made the rendered face not look visually similar to the source. So they devised an alternative, in which a detailed albedo is computed from the source and the rendered face is then projected onto this image plane, after which each mesh vertex is assigned the pixel colour. This way, they were able to compute the albedo using a dividing illumination technique, and the albedo from the frame with least rotation angles and most neutral expression was set as the final albedo of the video. This scheme was used in both general mapping and fine-tuning, for personalized mapping.

After the facial part is done, the hair and background regions also has to be mapped onto the rendered frames to generate a realistic talking head. The authors initially proceeded with the intuitive solution of matching a suitable background from the input with a similar head pose, but this did not yield them good results as the length of input videos were usually very short, about 10 seconds, which is less than 300 frames. So they proposed a new method, where some key-frames corresponding to crucial head movements from the synthesized pose sequence were extracted, and the key-frames with the largest head orientation in one-axis (either right or left) were chose to match the backgrounds. They called these matched backgrounds as key backgrounds and used linear interpolation to determine the frames between two adjacent key-frames. The poses are also modified in accordance with these backgrounds. In case the input source is a single image instead of a video, the authors suggest rotating the image using a face profiling method (Sela, Richardson, and Kimmel 2017) to obtain a matching background by pose prediction.

Refining frames using GANs

As previously mentioned, the synthesized frames need to be fine-tuned using a memory-augmented GAN to look more realistic. This paper does not use the traditional GAN-based face reenactment methods, which is what most deepfakes and Face2Face models use, and reason the authors state are:

1. Face reenactment is limited to single, specific face identities, whereas their model can generate different frame refinement effects for different identity features of target faces, keeping the GAN model fixed.
2. Face reenactment requires thousands of frames for training a single, specific face identity, whereas their model requires only a few frames for each individual identity for learning the general mapping. After this, the network is fine-tuned using a small number of frames from the input video.

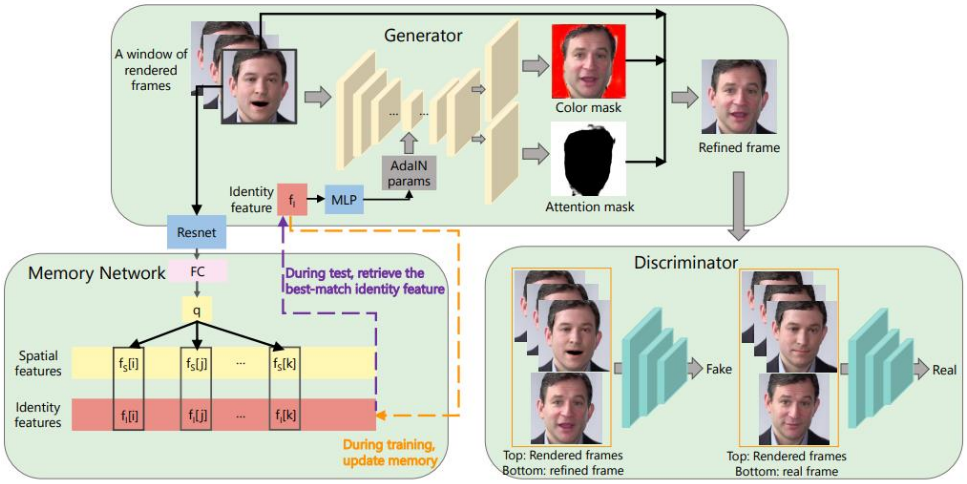
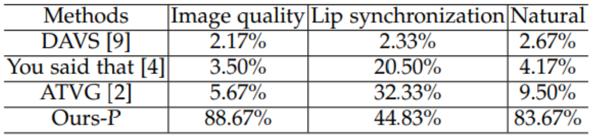


Figure 12: Memory-augmented GAN (Yi et al. 2020)

The process of refining the frames is modeled as a function that maps the rendered synthetic frames to the frame domain of the real face (original input data). Fig 12 shows the GAN network used in this paper, which consists of a conditional generator, a conditional discriminator, and an additional memory network for handling this multiple-identity refinement. The spatial features and identity features are stored in the memory network and are updated during the training process, and it can retrieve the best-match identity features including rare instances from the training set during the test. The conditional generator uses U-Net (Ronneberger, Fischer, and Brox 2015) and AdaIN (Huang and Belongie 2017) to synthesize a more refined frame after taking an identity feature and a window of three frames as input. The conditional discriminator is fed a window of three frames in random. It could be either a refined frame or a real frame. Based on the accuracy of classification, the discriminator is updated accordingly.

2.3.2 (iii) Results

As evaluating the naturalness and visual quality of synthesized videos is not standardized, the authors of this paper have designed a user study that evaluates their model and the state of the art methods based on a subjective score. 15 real world talking face videos were collected and the first 300 frames (12s for videos with a frame rate of 25fps) were used to fine-tune the network. For each of the video, two sets of audio were used, one from the remaining sections of the original video and the other from either VoxCeleb (Nagrani, Chung, and Zisserman 2017), which is an audio-visual database of over 7000 celebrities. 30 comparison groups were created, with each group containing 5 videos, where one was the original, and the other four were generated using four different methods. A total of 20 participants attended the study and each one of them were given all the 30 groups for comparison, with 3 criteria to assess for each group, that is, the image quality, lip sync and naturalness. The results are given in the fig 13, which shows that this paper has beat the state of the art methods with a huge margin.



Methods	Image quality	Lip synchronization	Natural
DAVS [9]	2.17%	2.33%	2.67%
You said that [4]	3.50%	20.50%	4.17%
ATVG [2]	5.67%	32.33%	9.50%
Ours-P	88.67%	44.83%	83.67%

Figure 13 Visual quality of the output by Yi et al. 2020

3. Methods

3.1 Our Model

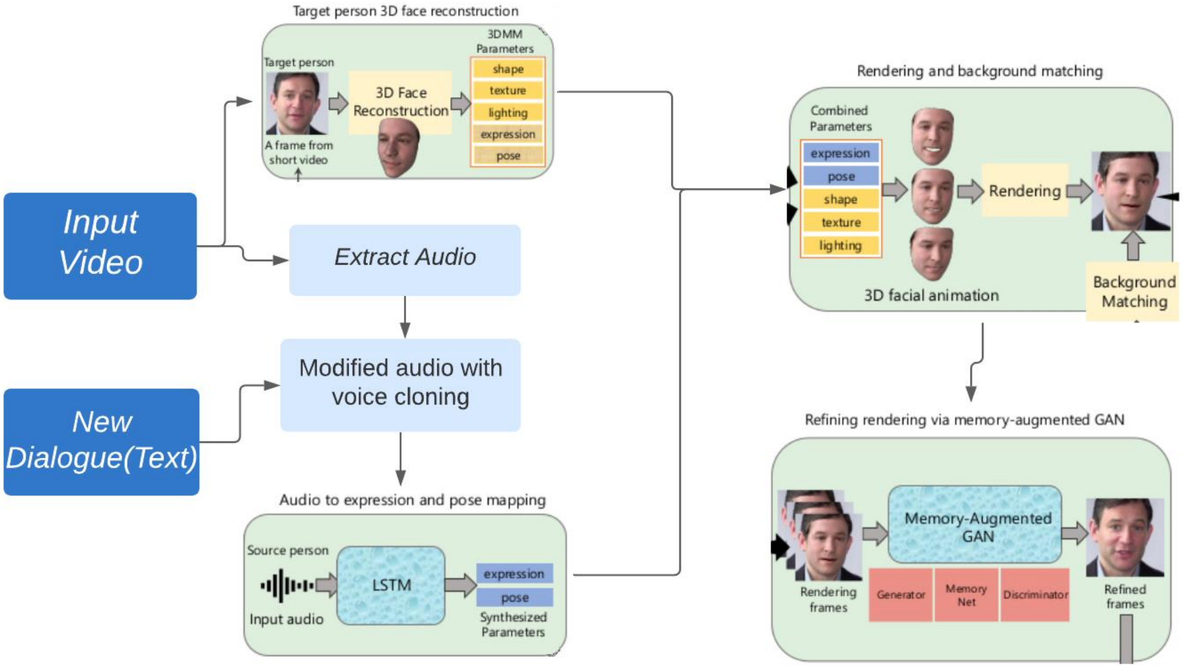


Figure 14: Proposed framework of the model used in our project. The figure is taken from Yi et al. (2020) and modified.

Based on the works by Jia et al. (2018) and Yi et al. (2020), we were able to combine the both and make a unified pipeline as shown in Fig 14. From the input video, the audio and video are first separated and the audio is used for mapping the vocal features of the subject to the SV2TTS system. A new dialogue is provided by the user as text format also to this system. The voice-cloning model uses this text input and the vocal features of the subject from the video to generate a new audio waveform. The generated audio is used by the LSTM network in the talking-face model to extract the expression and pose parameters. The extracted video is sent to the 3D face reconstruction system, which reconstructs a 3D model of the face and also extracts the 3DMM parameters like the shape, texture and lighting from the input video. These combined with the extracted pose and expression from the audio is sent for a final rendering of the face after applying all the parameters for the lip-sync and personalization of the video. The rendered face is fine-tuned using memory-augmented GAN for the final video output.

3.2 Experiments:

The methods were implemented in PyTorch and run on a Google Compute Engine equipped with a Tesla P100 GPU with 16GB VRAM and 3584 CUDA cores.

3.2.1 Experimental setup

For our experiment, we have taken a short clip (15s) from the TV series: The Office ("The Injury," Season 2, Episode 12), starring Steve Carell. The input video has a resolution of 1280x720 pixels, with a frame rate of 24fps and H.264 codec for video and AAC-LC codec for audio.

3.2.2 Voice Cloning

Though Jia et al. (2019) provides the best model out there to clone a voice using transfer learning, the authors have not made any implementations public. For his Master’s thesis, Jemine (2019) have implemented this model and made a real-time voice cloning tool publicly available (Appendix 1). As this tool is written in Python and Qt4 GUI, which makes it cross-platform, we use this tool to clone the voice from the input video. The toolkit is pre-trained using the LibriSpeech dataset and can accommodate any additional voice to be cloned. Audio clips of Steve Carell from both this scene and his other appearances are used for the voice cloning. As the audio codec AAC is not a lossless compression, we first convert it to WAV format. It is then loaded into the toolkit, which computes the embedding and updates the UMAP (McInnes and Healy, 2018) projections. This embedding is used to generate a spectrogram. Now, a new custom dialogue in text form can be provided by the user, which will be sythesized to generate a corresponding spectrogram. Additionally a mel-spectrogram of the audio sample and a heatmap of the embedding vector is also plotted for reference. The tool however does not support punctuation as of now. The synthesized spectrogram can now be converted to an audio waveform using the vocoder.

3.2.3 Audio-driven talking face

The training part involves two steps, a general mapping trained using the LRW dataset, and fine tuning to accommodate a new user. Since the first step is already done by the author, we only need to fine-tune the network with our cloned audio and the extracted video from the audio-visual input, and it is done as follows:

1. The input video is converted to 25 frames per second, and the first 300 frames are extracted.
2. 3D face reconstruction of each frame is done using MATLAB, but since we use Google Colab, Octave is used instead. Appendix 2 shows the 3D reconstruction of a sample of frames. Fig 15 shows how a masking layer is created to wrap the 3D face around the original frame and match the background.

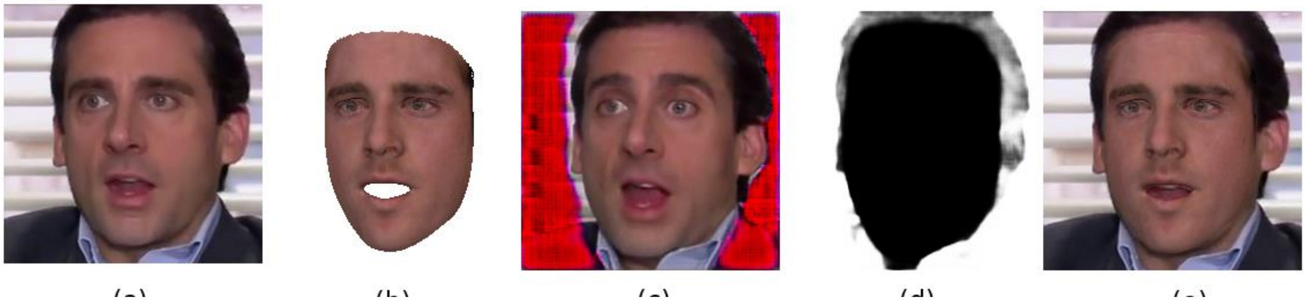


Figure 15: (a) Shows the original cropped face from the input video, (b) 3DMM extracted from it, (c) Color mask around the face, (d) Attention mask on the face and (e) Generated image for fine-tuning using GAN.

3. Fine tuning of the video with expression and pose correction is done using memory-augmented GAN. Fig 16: shows variations of the same frame with real and generated images back to back. A random sample from this collection is sent to the discriminator for classification and the weights of the generator and the discriminator and updated accordingly.

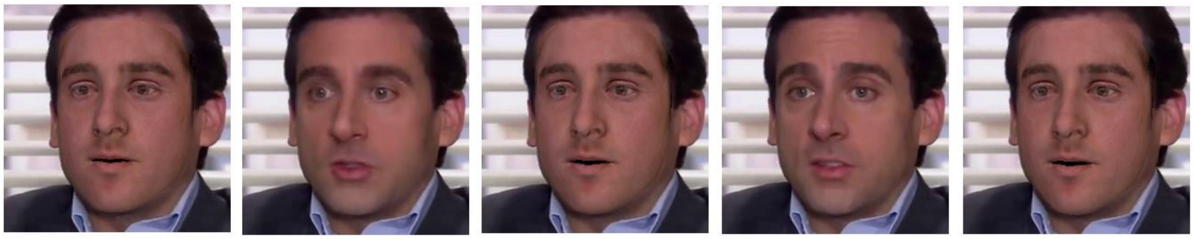


Figure 16: Variations of the same frame with real and generated images used for training GAN

4. The cloned voice is used as a test input. The 3DMM parameters, combined with lip-sync, pose and expression parameters from the LSTM network are used to refine the individual frames.
5. The generated frames are then attached back to the body and stitched together to form the final output video as shown in fig 17:

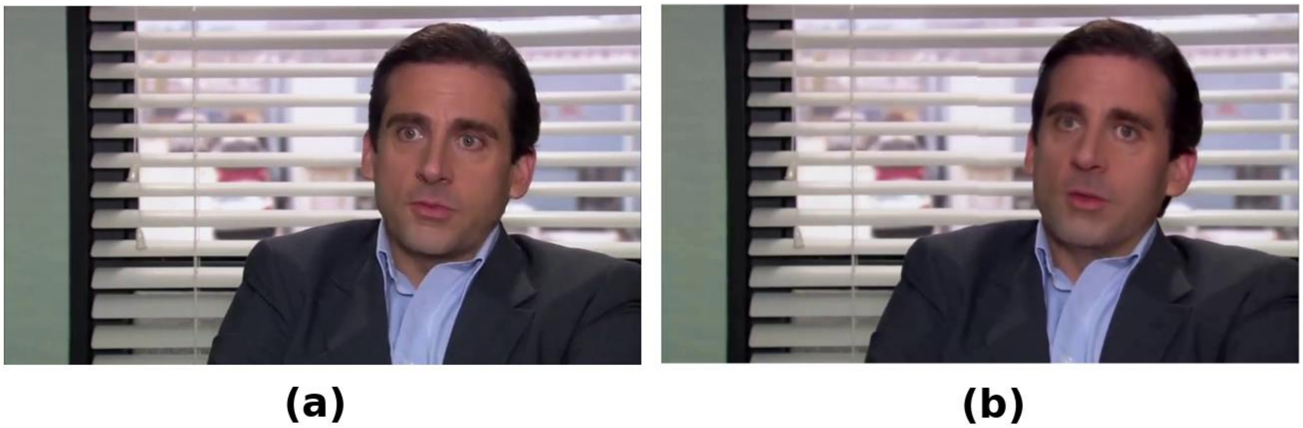


Figure 17: (a) Shows the input video sequence with full body, (b) shows the final output with background matching and fully scaled back to original aspect ratio.

4. Results

The quality of output audio-visual data is measured using two metrics, one for the audio quality and the other for the video quality, and both requires a survey from human evaluators. At the time of writing this report, the evaluation of the output that we have generated could not be completed, but the metric used for the individual components of this project by their respective authors and their results are used as a reference for an evaluation in the future.

For the audio, MOS being the standard for voice quality measure, is used as the metric to check the similarity and the naturalness of the speech, and is already explained in the literature review. For the video however, Yi et al. (2020) in addition to the subjective evaluation for the overall quality, as mentioned in their review in this report, have also proposed an innovative metric called HS for analyzing the head pose behavior.

The metric HS is used to measure the similarity of the head pose between the input video and the generated one. Three Euler angles are used to model the head movements, pitch for the head nod, yaw for the head shake/turn and roll for the head tilt (). A histogram of the pose angles in the real personalized video and the generated video are computed as $P_{(real)}$ and $P_{(gen)}$ respectively. Then the normalized Wasserstein distance (Villani, 2016) W_1 between them is computed. As W_1 is inversely proportional to the similarity between P(real) and P(gen), the HS metric is formulated as,

$$\text{HS} = 1 - W_1\ (P_{(real)}\ , P_{(gen)}\ ),$$

Where, HS ranges from 0 to 1, and higher the better. The authors were able to achieve an average score of 0.859 for 15 pairs of personalized videos, with a maximum of 0.956 and a minimum of 0.702. As we have used the model by the authors without any modifications to the visual part, we assume to have gained their same accuracy for head pose behavior, which is commendable. However, we expect to produce an overall score for the audio-visual output by the time of the demonstration of this project (Refer to Appendix 3).

6. Conclusion and future prospects

In this project, we looked at two state-of-the-art models, one in voice cloning and another in audio-driven talking-face, and were able to successfully concatenate them to create a pipeline for a realistic text-driven face reenactment. Although the results were satisfactory, we could not achieve a real-time implementation as the fine-tuning for a new face takes time. However, once trained on a face, the overall running time given a new text input is less than 10 minutes.

Our current project is limited to only one subject in the video as of now but has the potential to accommodate multiple subjects since face recognition and pose-tracking methods have reached an improved standard.

7. Bibliography

Arik, O. et al. (2017).Deep Voice 2: Multi-Speaker Neural Text-to-Speech.arXiv:1705.08947 [cs.CL].

Arik, O. et al. (2017).Deep Voice: Real-time Neural Text-to-Speech.arXiv:1702.07825 [cs.CL].

Arik, Sercan O. et al. (2018).Neural Voice Cloning with a Few Samples. arXiv:1802.06006 [cs.CL].

Aroon, Athira and S.B Dhonde (2015). "Statistical Parametric Speech Synthe-sis: A review". In:2015 IEEE 9th International Conference on IntelligentSystems and Control (ISCO), pp. 1–5.doi:10.1109/ISCO.2015.7282379.

Beard, M 2019, A Gentle Introduction to Generative Adversarial Networks, viewed 15 August 2021, <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/>

Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. 25(9), 1063–1074 (2003)

Bouaziz et al., (2013), Online modeling for real-time facial animation. ACM Trans. Graph.
Busso, Carlos et al. (2007). "Rigid Head Motion in Expressive Speech Anima-tion: Analysis and Synthesis". In:IEEE Transactions on Audio, Speech, andLanguage Processing15.3, pp. 1075–1086.doi:10.1109/TASL.2006.885910.

Cao, Chen et al. (Mar. 2014). "FaceWarehouse: A 3D Facial Expression Database for Visual Computing". In:IEEE transactions on visualization and computergraphics20, pp. 413–25.doi:10.1109/TVCG.2013.249.

Deng, J. et al. (2009). "ImageNet: A large-scale hierarchical image database".In:2009 IEEE Conference on Computer Vision and Pattern Recognition,pp. 248–255.doi:10.1109/CVPR.2009.5206848.

Deng, Yu et al. (2020).Accurate 3D Face Reconstruction with Weakly-SupervisedLearning: From Single Image to Image Set. arXiv:1903.08527 [cs.CV]

Eskimez, Sefik Emre, You Zhang, and Zhiyao Duan (2021).Speech Driven Talk-ing Face Generation from a Single Image and an Emotion Condition. arXiv:2008.03592 [eess.AS].

Everingham, M. et al. (Jan. 2015). "The Pascal Visual Object Classes Chal-lenge: A Retrospective". In:International Journal of Computer Vision111.1,pp. 98–136.

Ferrari, Claudio & Berretti, Stefano & Pala, Pietro & Bimbo, Alberto. (2018). Learning 3DMM Deformation Coefficients for Rendering Realistic Expression Images: First International Conference, ICSM 2018, Toulon, France, August 24–26, 2018, Revised Selected Papers. 10.1007/978-3-030-04375-9_27.

Fischer, P. et al. (2015).FlowNet: Learning Optical Flow with Convolu-tional Networks. arXiv:1504.06852 [cs.CV]

Genova, Kyle et al. (2018).Unsupervised Training for 3D Morphable ModelRegression. arXiv:1806.06098 [cs.CV].

Gonzalvo, Xavi et al., eds. (2016).Recent Advances in Google Real-time HMM-driven Unit Selection Synthesizer. Sep 8–12, San Francisco, USA, pp. 2238–2242.url:http://www.isca-speech.org/archive/Interspeech_2016/pdfs/0264.PDF.

Goodfellow, Ian J. et al. (2014).Generative Adversarial Networks. arXiv:1406.2661 [stat.ML].

Greenwood, David, Iain Matthews, and Stephen Laycock (Sept. 2018). "JointLearning of Facial Expression and Head Pose from Speech". In: pp. 2484–2488.doi:10.21437/Interspeech.2018-2587.

Greenwood, David, Stephen Laycock, and Iain Matthews (Aug. 2017). "Predicting Head Pose from Speech with a Conditional Variational Autoencoder".In: pp. 3991–3995.doi:10.21437/Interspeech.2017-894.

Guan, Yue, Qifan Zhang, and Panagiotis Tsiotras (2021). Learning Nash Equi-libria in Zero-Sum Stochastic Games via Entropy-Regularized Policy Approx-imation. arXiv:2009.00162 [cs.LG]

Guo, Jianzhu et al. (2021).Towards Fast, Accurate and Stable 3D Dense Face Alignment. arXiv:2009.09960 [cs.CV].

Guoming, Zhang et al. (Aug. 2017). "DolphinAtack: Inaudible Voice Com-mands". In:doi:10.1145/3133956.3134052.

Howard, Andrew G. et al. (2017).MobileNets: Efficient Convolutional NeuralNetworks for Mobile Vision Applications. arXiv:1704.04861 [cs.CV].

Huang, Xun and Serge Belongie (2017).Arbitrary Style Transfer in Real-timewith Adaptive Instance Normalization. arXiv:1703.06868 [cs.CV].

Isola, P. et al. (2018).Image-to-Image Translation with Conditional Adversarial Networks. arXiv:1611.07004 [cs.CV].
Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman (2017).You said that?arXiv:1705.02966 [cs.CV].

Jemine, C. (2019). *Master thesis : Real-Time Voice Cloning.* (Unpublished master's thesis). Université de Liège, Liège, Belgique. Retrieved from https://matheo.uliege.be/handle/2268.2/6801

Jia, Ye et al. (2019).Transfer Learning from Speaker Verification to MultispeakerText-To-Speech Synthesis. arXiv:1806.04558 [cs.CL]

Jurafsky D. and Martin J. (2008). Speech and Language Processing, Pearson Education (2nd edition).
Kalchbrenner, Nal et al. (2018), Efficient Neural Audio Synthesis. arXiv:1802.08435 [cs.SD].

Kaliouby, R. and Robinson P. (2005). "Generalization of a Vision-BasedComputational Model of Mind-Reading". In:Affective Computing and Intel-ligent Interaction. Ed. by Jianhua Tao, Tieniu Tan, and Rosalind W. Picard.Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 582–589.isbn: 978-3-540-32273-3

Kim, H. et al. (2018).Deep Video Portraits. arXiv:1805.11714 [cs.CV].

Krizhevsky, Alex (May 2012). "Learning Multiple Layers of Features from TinyImages". In:University of Toronto.

L. Peter and M. Ian (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Li eat al., (2017), Learning a model of facial shape and expression from 4d scans. ACM

Li J., Yu D., Jui-Ting Huang, and Yifan Gong, “Improving Wideband Speech Recognition Using Mixed-Bandwidth Training Data In CD-DNN-HMM”, 2012 IEEE Workshop in Spoken Language Technology (SLT2012).

Lin, Tsung-Yi et al. (2015).Microsoft COCO: Common Objects in Context.arXiv:1405.0312 [cs.CV].

McInnes, Leland, John Healy, and James Melville (2020).UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML].

McInnes, Leland, John Healy, and James Melville (2020).UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML].

Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman (2017). “Vox-Celeb: A Large-Scale Speaker Identification Dataset”. In:Interspeech doi:10.21437/interspeech.2017-950.url:http://dx.doi.org/10.21437/Interspeech.2017-950

Nguyen, Thanh Thi et al. (2021).Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv:1909.11573 [cs.CV].

Oloko-oba, Mustapha, Ibiyemi T.S, and Osagie Samuel (Oct. 2016). “Text-to-Speech Synthesis Using Concatenative Approach”. In: International Journalof Trend in Research and Development3, pp. 559–462.

Olszewski, Kyle et al. (2017). “Realistic Dynamic Facial Textures from a Single Image Using GANs”. In:2017 IEEE International Conference on Computer Vision (ICCV), pp. 5439–5448.doi:10.1109/ICCV.2017.580.

Oord, A. and Dieleman, S. 2016, WaveNet: A generative model for raw audio, viewed 15 August 2021, <https://deepmind.com/blog/article/wavenet-generative-model-raw-audio>

Oord, Aaron van den et al. (2016).Conditional Image Generation with Pixel-CNN Decoders. arXiv:1606.05328 [cs.CV].

Oord, Aaron van den et al. (2016).WaveNet: A Generative Model for RawAudio. arXiv:1609.03499 [cs.SD]

Pan, Deng et al. (2020). “Deepfake Detection through Deep Learning”. In:2020IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), pp. 134–143.doi:10.1109/BDCAT50828.2020.00001

Paysan, P. et al. (2009). “A 3D Face Model for Pose and Illumination In-variant Face Recognition”. In:2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301.doi:10.1109/AVSS.2009.58.

Ping, Wei et al. (2018).Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. arXiv:1710.07654 [cs.SD].

Purwins, Hendrik et al. (May 2019). “Deep Learning for Audio Signal Pro-cessing”. In:IEEE Journal of Selected Topics in Signal Processing13.2,pp. 206–219.issn: 1941-0484.doi:10.1109/jstsp.2019.2908700.url:http://dx.doi.org/10.1109/JSTSP.2019.2908700.

Ramamoorthi, Ravi and Pat Hanrahan (Sept. 2002). “An Efficient Representation for Irradiance Environment Maps”. In: Computer Graphics (ACMSIGGRAPH ’01 Proceedings)01.doi:10.1145/383259.383317

Redmon, Joseph et al. (2016).You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV].

Richardson, Elad, Matan Sela, and Ron Kimmel (2016).3D Face Reconstruction by Learning from Synthetic Data. arXiv:1609.04387 [cs.CV].

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015).U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597[cs.CV]

Sela, Matan, Elad Richardson, and Ron Kimmel (2017).Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. arXiv:1703.10131 [cs.CV].

Settel, Zack et al. (Jan. 1994). “Real-Time Timbral Transformation: FFT-basedResynthesis”.In:Contemporary Music Review 10, doi:10.1080/07494469400640401.

Shen, J. et al. (2018), Natural TTS Synthesis by Conditioning WaveNeton Mel Spectrogram Predictions. arXiv:1712.05884 [cs.CL].

Suwajanakorn, Supasorn, Steven Seitz, and Ira Kemelmacher (July 2017). “Syn-thesizing Obama: learning lip sync from audio”. In:ACM Transactions onGraphics36, pp. 1–13.doi:10.1145/3072959.3073640.

Taigman, Yaniv et al. (2018).VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. arXiv:1707.06588 [cs.LG].

Taylor P.(2009). Text-to-Speech Synthesis, Cambridge University Press.
Thies, J. et al. (2020).Face2Face: Real-time Face Capture and Reenactment of RGB Videos. arXiv:2007.14808 [cs.CV]
Trans. Graph.
V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In SIGGRAPH, volume 99, pages 187–194, 1999

Villani, C., 2016. *Topics in optimal transportation*. Providence, Rhode Island: American mathematical Society.

Wan, Li et al. (2020), Generalized End-to-End Loss for Speaker Verification.arXiv:1710.10467 [eess.AS].

Wan, Li et al. (2020).Generalized End-to-End Loss for Speaker Verification.arXiv:1710.10467 [eess.AS].

Wang, Weiyi, Valentin Enescu, and Hichem Sahli (2013). “Towards Real-Time Continuous Emotion Recognition from Body Movements”. In: Human Behavior Understanding. Ed. by Albert Ali Salah et al. Cham: Springer Inter-national Publishing, pp. 235–245.isbn: 978-3-319-02714-2.

Wang, Yuxuan et al. (2017).Tacotron: Towards End-to-End Speech Synthesis.arXiv:1703.10135 [cs.CL].

Wu, Cho-Ying et al. (2021).Voice2Mesh: Cross-Modal 3D Face Model Genera-tion from Voices. arXiv:2104.10299 [cs.GR]

Yi, Ran et al. (2020).Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose. arXiv:2002.10137 [cs.CV]

Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3D facial expression database for facial behavior research. In: IEEE International Conference on Automatic Face and Gesture Recognition (2006)

Zen, Heiga et al. (2016). "Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices". In: Proc.Interspeech. San Francisco, CA, USA, pp. 2273–2277.

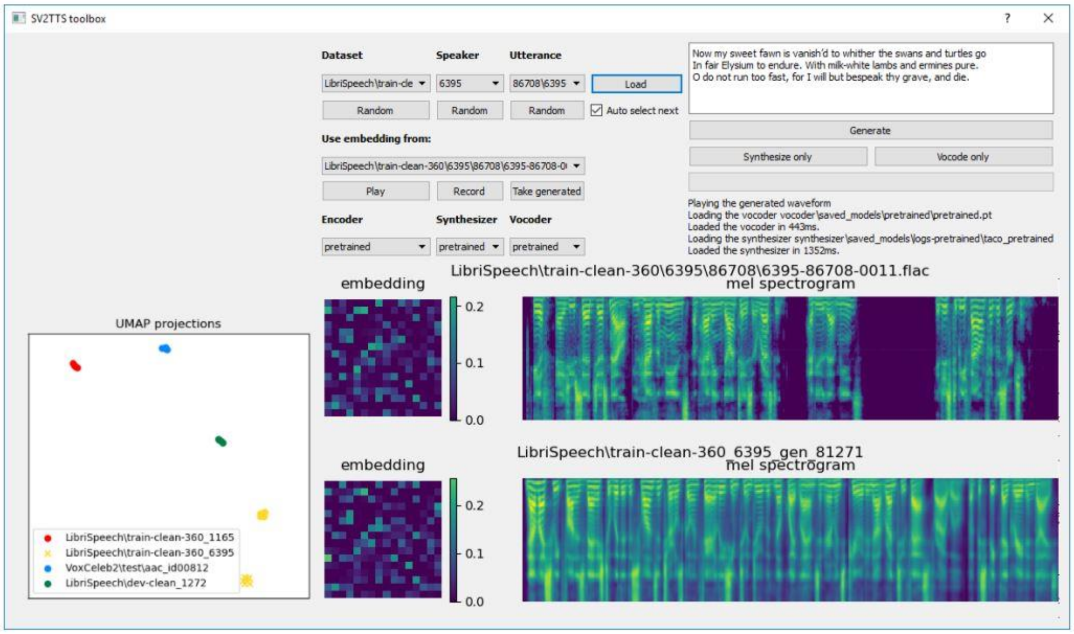
Zhang, X. and Weng, L., 2020. Realistic Speech-Driven Talking Video Generation with Personalized Pose. *Complexity*, 2020, pp.1-8.

Zheng, C. et al. (2021).Deep Learning-Based Human Pose Estimation: A Sur-vey. arXiv:2012.13392 [cs.CV].

Zhu, Hao et al. (2020).Deep Audio-Visual Learning: A Survey. arXiv:2001.04758 [cs.CV]

8. Appendices:

Appendix 1



SV2TTS toolbox

Dataset: LibriSpeech\train-cle

Speaker: 6395

Utterance: 86708\6395

☑ Auto select next

Use embedding from: LibriSpeech\train-clean-360\6395\86708\6395-86708-0

Encoder: pretrained

Synthesizer: pretrained

Vocoder: pretrained

Now my sweet fawn is vanish'd to whither the swans and turtles go
In fair Elysium to endure. With milk-white lambs and ermines pure.
O do not run too fast, for I will but bespeak thy grave, and die.

```
Playing the generated waveform
Loading the vocoder vocoder\saved_models\pretrained\pretrained.pt
Loaded the vocoder in 443ms.
Loading the synthesizer synthesizer\saved_models\logs-pretrained\taco_pretrained
Loaded the synthesizer in 1352ms.
```

?
Load
Random
Random
Random
Play
Record
Take generated
Generate
Synthesize only
Vocode only
UMAP projections
LibriSpeech\train-clean-360_1165
LibriSpeech\train-clean-360_6395
VoxCeleb2\test\aac_id00812
LibriSpeech\dev-clean_1272
embedding
0.2
0.1
0.0
LibriSpeech\train-clean-360\6395\86708\6395-86708-0011.flac
mel spectrogram
embedding
0.2
0.1
0.0
LibriSpeech\train-clean-360_6395_gen_81271
mel spectrogram

Figure 18: The Interface of theSV2TTS toolbox we used for real-time voice cloning (Jemine 2019 p. 31).

Appendix 2

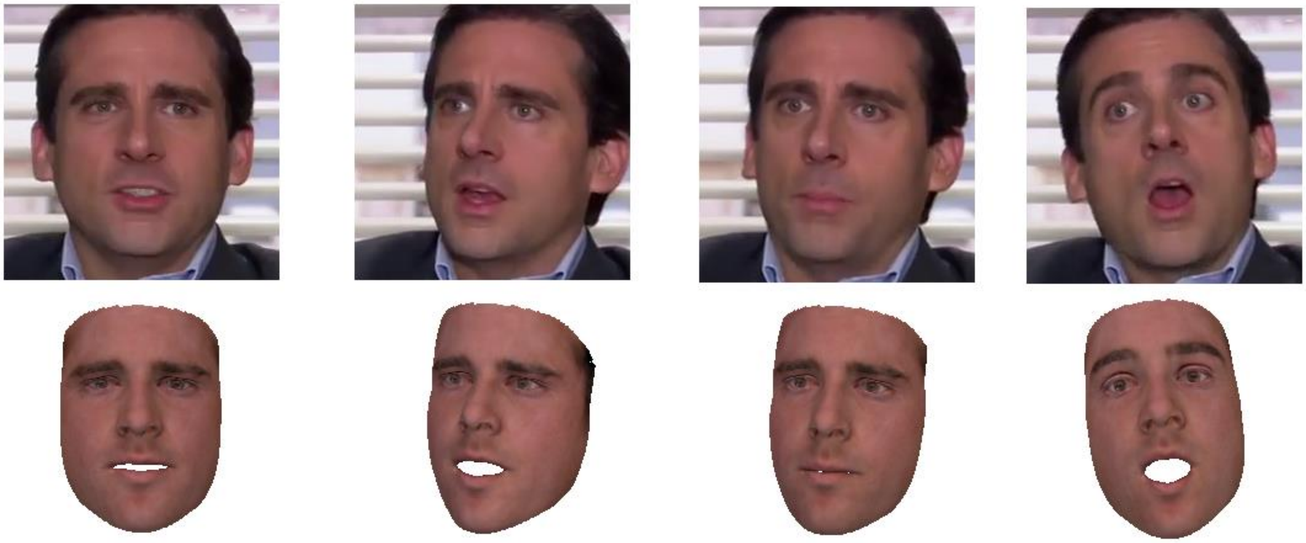
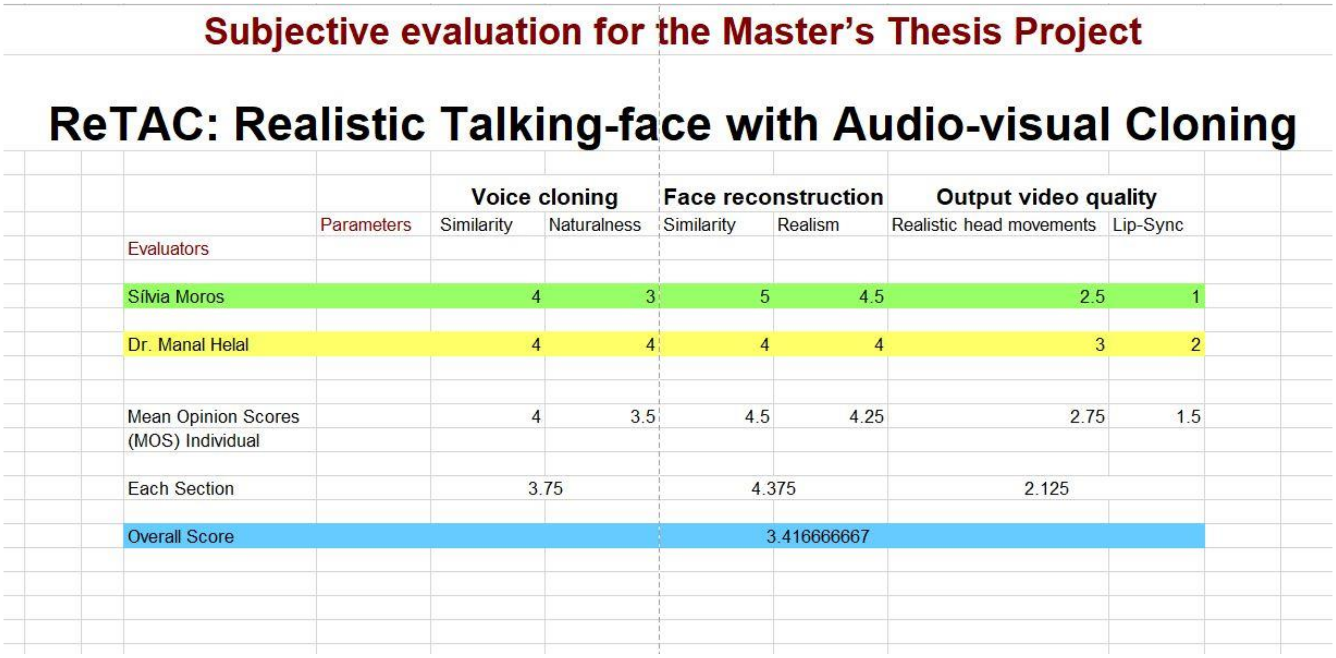


Figure 19: 3D face reconstruction using 3DMM on the frames extracted from the input video sequence.

Appendix 3



Subjective evaluation for the Master's Thesis Project

ReTAC: Realistic Talking-face with Audio-visual Cloning

		Voice cloning		Face reconstruction		Output video quality	
	Parameters	Similarity	Naturalness	Similarity	Realism	Realistic head movements	Lip-Sync
Evaluators							
Sílvia Moros		4	3	5	4.5	2.5	1
Dr. Manal Helal		4	4	4	4	3	2
Mean Opinion Scores (MOS) Individual		4	3.5	4.5	4.25	2.75	1.5
Each Section		3.75		4.375		2.125	
Overall Score		3.416666667					

Figure 20: Subjective evaluation during the presentation of the thesis by the supervisor and the evaluator. MOS Scale: 0 for the worst and 5 for the best.