MASTER THESIS IN ARTIFICIAL INTELLIGENCE

A Photo-realistic Voice-bot

Author: Jorge Alexander
Supervisor: Núria Castell Ariño, Computer Science Department, UPC

Facultat d'Informàtica de Barcelona (FIB)
Universitat Politècnica de Catalunya (UPC) - Barcelona Tech

Facultat de Matemàtiques de Barcelona
Universitat de Barcelona (UB)

Escola Tècnica Superior d'Enginyeria
Universitat Rovira i Virgili (URV)

July 1st, 2019

Contents

Acknowledgement

I would like to express my sincere gratitude to:

- My supervisor, Núria Castell Ariño, for always being able to make time for me.
- Josep Ramon Morros, for all the well explained technical advice.
- My family, for putting up with me being back at home.

Abstract

Technology is at the point where systems are capable of synthesizing video of human actors indistinguishably from ones in which the actor is present [1].

This research investigates whether or not it is possible to use this technology in order to create a system which, allows video generation of a human actor, that is able to interact with a user through speech in real-time, whilst also remaining indistinguishable from a real human actor. In other words, a photo-realistic voice-bot.

The work discusses the motivations and ethics, but also presents and tests a prototype system. The prototype aims to take advantage of the latest in real-time video manipulation software to create a natural sounding conversation with an artificially synthesized video.

Keywords: voice; bot; photo-realistic; real-time; video generation; speech synthesis; pix2pix; dialogflow.

Chapter 1

Introduction

1.1 Context

The idea of synthesizing features of a human being as an interface for a computer system, so that it is indistinguishable from a person, is not a novel idea.

In science-fiction, Isaac Asimov presents the physical synthesis of a human being in his 1954 book "Caves of Steel" as a robot that is indistinguishable from a human, allowing it to infiltrate undetected into anti-robot societies [1]. In 2013 the movie "Her" presented an artificial intelligence that could synthesize human-like voice and personality, leading to a romance with the human protagonist [2].

In non-fiction, presently, there are already household devices that begin to achieve the same goal, such as voice assistants which synthesize speech, Amazon Alexa, Google Assistant or Apple's Siri for example. However, it is only in the last few years, that these systems are becoming truly indistinguishable from real human interactions. In 2018 Google demonstrated a system called Duplex, which can make dinner reservations by pretending to be a human being [3]. This level of human-like synthesis has led to the rise of terms such as "Fake News" to indicate the artificial generation of text indistinguishable from human-written text and "Deep Fakes" which describe artificially generated images of human characteristics such as faces, that are indistinguishable from real ones [4].

It is important not to confuse the idea of a system that can synthesize a human for interaction, indistinguishably from another human, as intelligence. This is one of the flaws of the Turing test as a test for intelligence [5]. In order to pass the test, the system just has to be indistinguishable from a human when questioned, but does not have to show any internal reasoning.

This work concerns itself with three main objectives. The first, an attempt to cover the current state of human-realistic audio-visual generation as a form of human-computer interaction. The second, to present a prototype that can synthesize a photo-realistic human-looking avatar, which the user can interact with through speech. Finally, to discuss the ethical implications of realising this technology.

1.2 Motivations

The motivation for synthesizing interactions like a human include increasing user attention, engagement and entertainment [6]. Subjective tests have also showed that websites with a photo-realistic avatar head receive higher ratings than the same website without one [7].

However it is difficult to justify that these interactions be indistinguishable from a human one. To understand this, let us consider the case of a navigation system in a car that interacts with a driver through speech. In order to interact with the driver, the navigation system must be able to capture the driver's audio, interpret any speech in the audio correctly, and convert the speech into some internal processing structure. Then, after internal processing, in order to respond back to the driver, it must synthesize the response into natural language, and synthesize the natural language into audio which should contain understandable speech. It seems like a lot of effort on behalf of the system to adapt to a human's way of interacting, particularly when there may be a more optimised way to interact for both human and machine, for example by showing a map on a screen with a visual line of the route the driver must take. However, in this case, the advantage is that the system can inform the driver with audio, so that the driver is not distracted from driving.

Hence this is a case where a restriction on the senses of a human has meant that a system has had to be adapted in order to interact with them. The motivation of helping those with restrictions on their senses is another motivation and aligns nicely with all the research in the area of assistive technology [8].

However, ultimately the synthesis of speech in a navigation system does not have to be perfectly human, it just has to be good enough for the human to understand the route to follow. So although this is a motivation for synthesis of human-similar interaction, it is not a motivation for indistinguishable human-like interaction. This is a subtle but most important point. In order to understand the motivations for indistinguishable (both photo and audio) human-like interaction, we need to dig a bit deeper.

Up until recently (2018) it could be assumed that you were interacting with another human when interacting in a human-like way, however this assumption is now being broken. Technology such as Google duplex allows human-like phone calls to be made by a machine, meaning that a human can no longer assume that they are speaking with another person on the phone anymore [9]. This is occurring across written and visual media also and may have profound consequences [4]. It is difficult at the present time to understand how the breakdown of this assumption will impact society, for better or worst, but a further motivation of this work is to try to understand how people react to indistinguishable human-like synthesis as a way of interaction.

Finally, having the ability to synthesize oneself digitally can have several motivations. A simple way in which this has already been done is in the form of answering machines, where a personalised message is often created by the user in order to communicate to callers later on that they are not able to answer a call. One can imagine that being able to synthesize a digital clone of yourself could provide useful when attempting to interact on a scale that one does not have the time for, if your presence were required.

1.3 Approach

There are several technologies out there that allow the synthesis of different aspects of a human being. For example, ones that have received media attention in 2018 include Fake News and Deep Fakes [4], as well as Google Duplex [9], which have demonstrated that we are able to generate human-like text, videos and speech respectively.

One way to unify all these techniques into one would be to create a digital photo-realistic voice-bot, i.e. a system that presents a user with a synthesized video of an actor, which then adequately interacts with the user via speech, as a human would be expected to do over a video-call.

In order to make the voice-bot indistinguishable from a human counterpart, the separation of the system into three main components is suggested: generation of a text response with semantic meaning from a user query (a dialogue system); synthesis of natural speech from text; and real-time, photo-realistic, video synthesis from speech. An additional constraint for the system is the ability to respond to the user with no longer a delay than a natural pause.

Chapter 2

Literature Review

2.1 Photo-realistic Voice-bot

Although there is a plethora of research regarding dialog systems, speech synthesis and video generation from speech, there are only a handful of attempts which unify them to create a photo-realistic voice-bot. The most similar work, appears to have stemmed from research involving photo-realistic talking heads. The techniques used to create these heads can be generally classified into model-based and image-based, however the first photo-realistic example was image-based.

Research on photo-realistic talking heads emerge around the late 1990s and beginning of the millennium. In 2000 Eric Cosatto and Hans Peter Graf create a system to generate a talking head from image samples [10]. Motivated to increase user attention, engagement and entertainment, they were able to create a system which can generate a photo-realistic model of a human head, that can be animated and lip-synced, and that closely resembled real people. To achieve this they record a talking person and apply image recognition to extract facial features as bitmaps. These bitmaps are stored in a database and accessed during the synthesis phase. The synthesis phase is driven by a text-to-speech (TTS) model that retrieves the corresponding bitmap associated with a particular phoneme, before blending it onto an image of the whole head using pose information. The system appears to create

photo-realistic images and also improve intelligibility in noisy environments compared to a simple 3D head model [10], however informal tests revealed issues with lip synchronisation and over articulation, and blending artefacts. Also notably missing is an integration of a dialog system.

In 2001-2004 Microsoft Research develop a system they name E-Partner; a photo-realistic conversational agent. Motivated to improve human-computer interaction, subjective tests had showed that an E-commerce website with a talking head gets higher ratings than the same website without a talking head [7]. The video generation from phonemes was also sample based, building upon Cosatto's work. The system integrated a rule-based informative and adaptive dialogue model that would draw from templates for responses. It used a Hidden Markov Model (HMM) for prediction of video from audio features. It used an interesting technique whereby they create an infinitely long video from a short video sequence by jumping back to an earlier frame whenever there is a smooth transition from the current frame (using L2 distance between frames to define smoothness) and also achieved video continuity so that the bot appeared to be listening, could be interrupted, and transitioned smoothly into an active talking state. It used pre-recorded speech to communicate with users, (but could use Microsoft's TTS for lower quality) and could also draw on additional multimedia material. The system was tested as a virtual tour-guide (Figure 1). Although the work [11] does not elaborate on any weaknesses, one limitation that can be inferred is the dependency of the system on templates, both for dialog and speech, limiting the generation of new content and the ability to scale the system to other use cases. Upon watching a demo of the system [1], it shows unnatural movement of jaw, an unnatural lack of movement of the upper part of the face, and a clearly digital TTS audio.

Microsoft extended on this research in 2011. By incorporating aspects of image based approaches with model based approaches they were able to improve on the E-Partner system. With twenty minutes of 2-D video from a person, the system is able to create a photo-realistic 3-D talking head, including facial expressions and

[1] https://www.microsoft.com/en-us/research/project/photo-real-talking-head

Figure 1: Microsoft's E-Partner

text-to-speech synthesis. It renders a 3D head by mapping or wrapping 2D video images around a simple, 3D mesh model, and uses Hidden Markov Model to predict images from a given phoneme. The advantage of combining a model-based approach, meant that they were able to control aspects such as the pose of the head. However in [6], there is some notable lack of texture when wrapping a 2D image around a simple mesh model (see Figure 2).

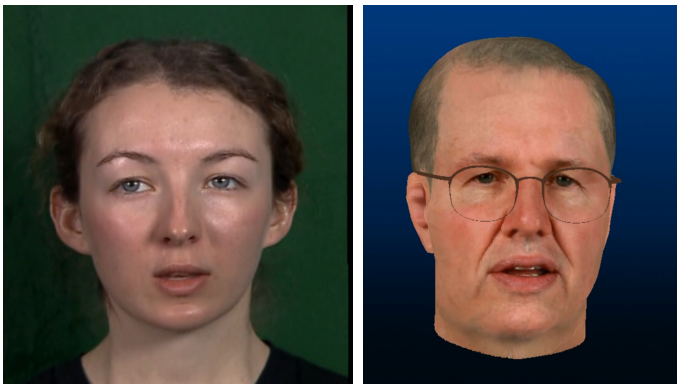

Figure 2: Microsoft's E-Partner (left) vs Microsoft's 3D Talking Head (right)

Further work by Microsoft research in 2015 [12], used bi-directional long short-term memory (LSTM) network to predict visual sequences from audio, instead of an HMM one. The user study conducted suggested preference of the new LSTM one.

However, again, since Microsoft's photo-realistic conversational agent, we've failed to find research that combines the dialog system, text-to-speech synthesis and photo-realistic video synthesis. The approaches that include all three, such as the Virtual

Human Toolkit [13], use a model-based approach. The advantage of model-based approaches include having the flexibility of pose change, environment change and lighting changes, often at the sacrifice of photo-realism. However recent advances in using deep neural networks to synthesize photo-realistic images suggests that advantages such as pose control can now be performed by image based approaches, for example the X2Face model in [14]. Furthermore, recent approaches to generate facial animations from speech, such as by Vougioukas et al. [15], are now able to generate videos of previously unseen subjects, meaning that these models can easily be adapted to synthesize new people without the extra modelling overhead often required by model-based approaches.

The following sections describe recent development in the areas from which we can choose components from in order to improve on Microsoft's photo-realistic conversational agent.

2.2 Realtime Video Synthesis

Work by Anderson et al in 2013 [16], included an image based-approach using active appearance models (AAM) in 2D (and HMM prediction). Interesting results from the work include a large-scale user study (100 users) of the realism of different Video-Text-To-Speech systems. It revealed that image based talking heads were achieving high levels of perceived realism. The paper reports that users were rating an average of 3.8/5, for Anderson et al's system (which included emotional expressions), and 4.3/5 for Liu et al's system [17], (which did not include emotional expressions) where a score of 5 corresponded to completely real, indistinguishable video. This indicates that by 2013, research was getting close to achieving near indistinguishable synthesis of human talking heads. The user-studies conducted also indicated that recognition of emotional expression was comparable, even slightly higher than that from real footage. However, these systems were still limited by a text to speech model that sounded digital, (they used non-synthesized audio samples for the user studies) as well as the exclusion of a dialogue model

Some of the most recent state-of-the-art work has stemmed from the paper *Face2Face: Real-time Face Capture and Reenactment of RGB Videos* [18]. This allowed transferal of head motions as well as expressions onto a photo-realistic fake video, in real-time, driven by a source video of an actor. Thies et al. later improved on their own work to include facial expression control [19]. All that would be left to create an improvement on Microsoft's photo-realistic chatbot would be to drive the model (with emotion and lip-sync) from text or speech, and to connect a suitable dialog model to generate it. Although Thies et al. work is not publicly available, there are severela publicly available implementations, for example the Pix2Pix image-translation model [20], which uses Generateive Adversarial Networks (GANs), to learn how to map from a source image to a target image, and has shown success in transferring photo-realistic images from source to target.

Other work that might at first seem suited to creation of a video chatbot include Obamanet [21], which attempts a complete map from text -> speech -> facial expression -> output video, although the clarity of the mouth-region remains blurry without the pre-processing explained in [22].

2.3 Facial Expression Synthesis

In 2019, [23] developed text based editing of a talking-head. Which would allow for creation of a photo-realistic voice-bot if connected to a dialogue system. However although this work appears realistic, it lacks emotional expression and takes around 110ms per frame, which would lead to a slow frame rate of around 9 frames per second. Generating complete facial expressions remains a challengeing task, although lip-syncing is nearing human-like synthesis.

For example, in 2017 Zisserman et al. created the Speech2Vid model [24], to predicting lip-synced mouth features in a subject independent way. His work also includes Syncnet [25], which creates realistic lip-syncing of a video. However, the work lacks generation of emotional facial expression. The X2Face model [14], a self-supervised model for driving video from other video or audio, partly addresses the issue, but

produces warped results when driven by audio. One of the reasons why Zisserman et al. models are not successful in generating facial expressions may be that the audio is fed into the models as mel-frequency cepstral coefficients (MFCC), which are conjectured to lose the information needed to infer emotion from audio [26].

In 2017, Karras et al at Nvidia publish *Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion* [27]. They propose a framework to generate facial animation, including emotion, from speech-audio. A challenging task, as it has to deal with understanding the lexical stress of a phrase, interaction between facial muscles and skin. By using a convolutional neural network architecture that generalised over different speakers, they are able to generate facial expression landmarks from a speech file. However, the emotional state was not inferred automatically from the audio, it was passed as an accompanying emotional state.

Other successful work in generating speech-driven facial animation include Imperial College's end to end model *End-to-End Speech-Driven Facial Animation with Temporal GANs*. According to them, it is the first ever method which is capable of generating subject independent, photo-realistic videos directly from raw speech-audio, including lip movements and natural facial expressions, without relying on manual handcrafted intermediate features. It works in real-time and would also be a good candidate to create a photo-realistic voice-bot [15].

Another interesting piece of relevant research was by Pham et al. in 2017 [26]. They were able to create emotional facial expressions from audio, by instead of using MFCC coefficients, working directly off of the Fourier transform of the speech-audio signal. This provides a potentially useful technique for generating emotional facial landmarks from speech.

2.4 Realtime Speech synthesis

Historically, there have been many attempts to synthesize human speech, however an important step was the development of the Vocoder and Voder by Bell Labs in the 1930s. The Vocoder allowed a way to encode speech into linguistic features,

the Voder allowed synthesizing of the fundamental tones and resonances into speech. However, the quality of the synthesis was not perfect and clearly synthetic [28]. More recently however, synthesis of speech is nearing indistinguishable from human levels of sound. Research suggests [29] that users prefer the combination of synthetic sound and synthetic video than photo-realistic video and synthetic sound, and therefore achieving human levels in both aspects is important.

Currently, state of the art speech synthesis models are still based on the two step process that Bell Labs used in the 1930s, however now they are powered by neural-networks and a lot more computing power. They generally have one step to encode the speech from text, for example creating a time-aligned mel-frequency spectrogram, and a second step to transform the encoding into audio-speech [30]. However, there is now also success shown by end-to-end synthesis approaches, which do not depend on linguistic knowledge to engineer an internal representation, such as Char2Wav (Sotelo et al 2017) [31], (although according to [32], acoustic parameters are still learned as an intermediate step).

The current state-of-the-art for speech-synthesis would be the Tacotron2 model [33], which builds upon Wavenet [34]. Wavenet is a previous model that could generate almost human like speech and Tacotron uses it as a vocoder. It then uses a recurrent sequence-to-sequence character embedding for mel-spectrogram prediction. It is able to achieve a mean opinion score (MOS) of 4.53. Human speech achieves MOS scores of 4.58 according to [34].

Work by Baidu Research [35] (2018), demonstrated two models capable of synthesizing a person's voice. Not only this, but one of the approaches was able to synthesize a persons voice with only seconds of audio data. It was able to achieve a maximum MOS score of 3.16Âś0.08.

Finally work by Adobe [36], allows editing of user speech to say previously unuttered sentences, whilst maintaining the narration style of the user speech being edited. However, it is intended for editing of audio-clips rather than real-time speech synthesis.

2.5 Dialogue Models

State-of-the-art Dialogue models are still not able to have indistinguishable human-like conversations with humans [37].

In advancing the field, researchers focus on several different approaches, not necessarily evaluated as being human-like. For these researchers, an evaluation method such as the Turing test, which tests whether a dialogue model is able to convince the interlocutor that they are a human, is not a good evaluation mechanism [38]. In the case of the work presented here, whose goal is to synthesize human-like AI, the Turing test remains valid.

For state-of-the-art dialogue models evaluated via the Turing test, the Loebner competition is a good place to start. The Loebner prize is an annual prize given at the competition for the most human-like chatbot, according to evaluation by a panel of judges implementing a Turing type test. There are also two grand prizes, the first for a text chatbot that pass the Turing test and another one for a chatbot that can do this whilst processing not only text, but visual, and auditory input as well; known as passing the total-Turing-test [5]. As of the competition held int 2018, no chatbot has managed to win either of the grand prizes.

The winner of the last three Loebner prizes (2016, 2017, 2018) was Mitsuku, a rule-based chatbot. One of the reasons it one is its ability to reason about objects mentioned in dialogue. It personifies an 18 year old female from Leeds, England.

Another way of developing a dialogue model is through generative approaches. For adapting to the style of a particular person, generative dialogue models would yield more success, as they forgo having to create expert rules specifically for that person. In fact, in 2016, Stanford, Microsoft Research and University College London described a persona-based neural conversation model, which trained on conversational data could imitate the style of a particular speaker [39].

Replika AI[2], a company based on generating dialogue with a persona based neural

[2]https://replika.ai/

network, is an example of where this technology is making it into the commercial space. Other commercial applications include Dialogflow, which allow the user to create a series of request-response rules which they then use machine learning algorithms to select optimally. However, they also provide Wavenet speech synthesis and recently have incorporated information retrieval queries from documents.

Chapter 3

Methodology

3.1 Overview

The system was approached in a modular fashion, so that it could be worked on and upgraded incrementally. Furthermore, in order for the system to run in real-time, the speech processing was delegated to a background thread process, which communicated with a concurrent real-time, video playing thread via a duplex message queue (the Python Pipe connection object was used). Figure 3 shows an overview of the system and how a user would interact with the it.

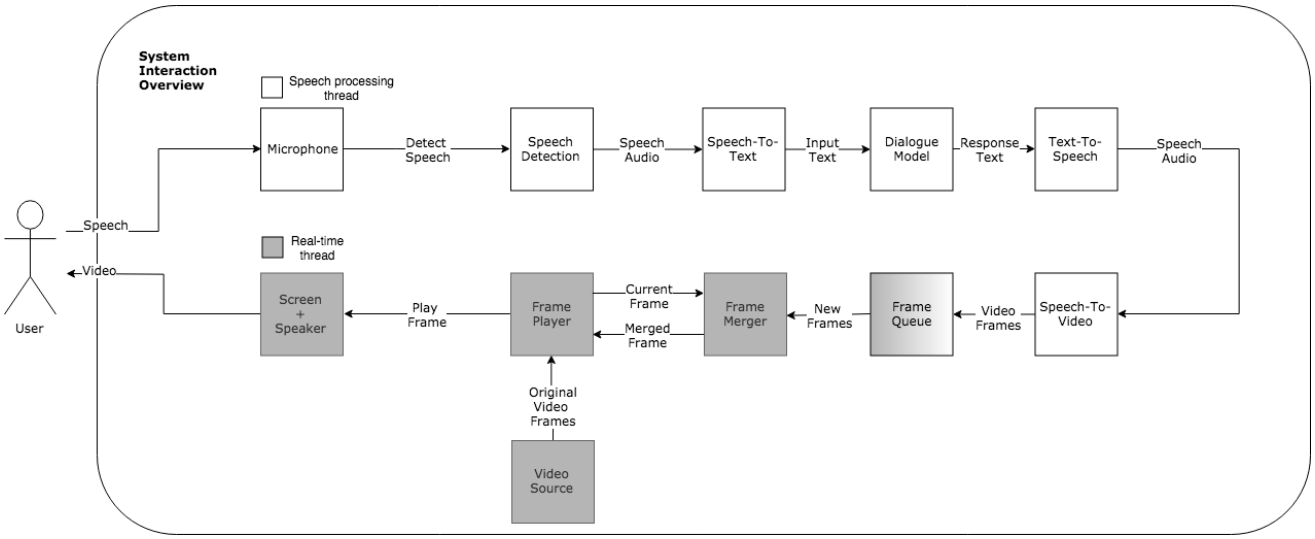

Figure 3: User System Interaction

The system was built on a computer with a webcam and microphone, an Intel Core i7-8750H CPU, NVIDIA RTX 2070 GPU, and 32 GB of RAM. The user interacts with the system by talking into the microphone and watching the video on the computer screen.

The following sections describe how each part of the system operates.

3.2 Dialogflow

Dialogflow is a web platform that allows the creation of conversational bots[1]. It provides an extensive Python API which allows audio streaming, automatic speech detection, speech-to-text conversion, dialogue model creation and text-to-speech conversion. Not only that, but it produces text-to-speech synthesis using a model trained with the Wavenet architecture [34]. Furthermore, as the platform is owned by Google, who are responsible for the Google Duplex system, there may be the opportunity to improve the text-to-speech system in the future. Therefore Dialogflow was used as a starting point to cover those respective parts of our system. Figure 4 shows the components that were handled on the Dialogflow platform.

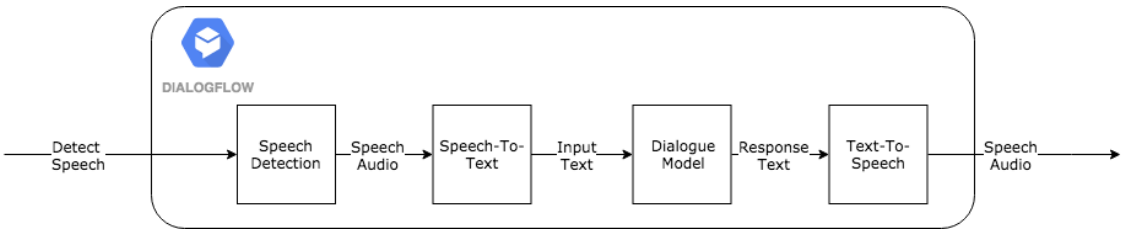

Figure 4: Dialogflow Integration

The system continuously streams audio to Dialogflow. If Dialogflow detects speech in the audio, then it begins to process it, converting it to text.

To process the text, Dialogflow provides the tools to build a dialogue system, but it does not build it for you, therefore the dialogue model had to be created. For testing, the dialogue model was built according to the rules of the original Eliza chatbot [40]. This gave a simple base of automatic reply functionality for the chatbot.

[1] https://dialogflow.com/

However, the dialogue model is easily customised to include recorded speech response (bypassing the TTS module) or could also be adapted for Wizard-Of-Oz based testing.

Once a text response had been generated, Dialogflow converted the text to speech. To do this, the WaveNet TTS model was chosen, which has a mean opinion score (MOS) of 4.21 (human speech has an MOS of 4.55) [34]. The trained model that Dialogflow provides under the label *en-US-Wavenet-F* was used, with no changes in pitch or speaking rate. A 16-Bit output audio encoding was selected.

3.3 Speech To Video

Once the system had received the response audio bytes from Dialogflow, the next step was to synthesize video from the audio.

The lip-synced generation was focused on first, and several approaches had already been done by previous papers. The first that was tried was the one used by Obamanet [21]. Obamanet works by training a time-delayed LSTM network to predict the Dlib [41] facial landmark points of the lips, given an audio sample of Obama speaking. However, upon viewing examples [2], the lip-syncing seemed to be missing the pronunciation of certain phonemes, such as the 'th', which is due to the lack of tongue modelling [21].

The most convincing model found to predict lip movement from audio in real-time was the Speech2Vid model created by Chung et al [24]. The work proposes a model that, given still images of a face with varying mouth openness, and an audio speech segment, outputs a lip-synced video of the target face speaking. Also interestingly, the model generalises to unseen faces, and therefore does not need to be retrained for new identities.

A Speech2Vid model is available pre-trained[3] on the private LRS2 lip-reading dataset [42] and created using a Directed acyclic graph neural network (DagNN) in Mat-

[2]https://www.youtube.com/watch?v=9Yq67CjDqvw
[3]http://www.robots.ox.ac.uk/ vgg/software/yousaidthat/

convnet [43] with no compatible extraction tools. Therefore incorporating it into the system was a bit tricky, but the following solution worked. Figure 5 shows an overview of how it was incorporated.

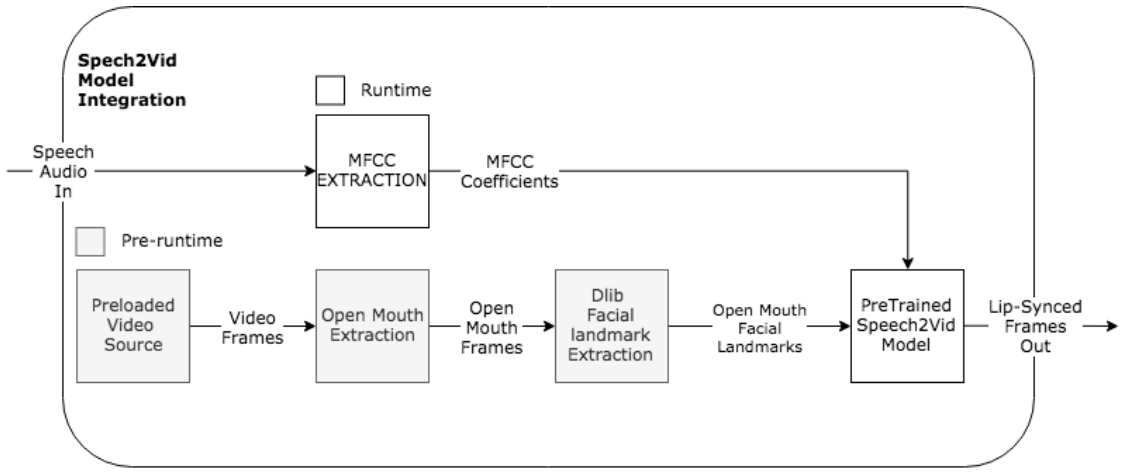

Figure 5: Speech2Vid Model Integration

First of all, in order to use the model, we needed to generate 5 images of varying mouth openness (10, 30, 50, 70, 90 is recommended). Therefore, pre-runtime, a short video of the actor speaking a few sentences was processed. To do this, a Dlib 68 point facial landmark predictor [41] was used to extract the frames with the correct mouth openness. Mouth openness was defined as the pixel distance from Dlib facial landmark 52 to 58. Finally, the Dlib facial landmarks of these images were extracted, rotated using a Procrustes transformation to fit the landmarks of an average face, as described in [24]. To generate a frame, mel-frequency cepstral coefficients (MFCC) were created from the Dialogflow output audio and processed according to [24]. The coefficients were passed to the model as input along with the facial landmarks of the faces with varying mouth sizes.

The model then generated a series of output faces, according to the MFCC coefficients. Finally the corresponding slice of audio used to generate the corresponding output face was combined with the output face image into a Lip-Synced Frame.

A Frame in the system contains a face image with it's associated slice of audio; each frame represented 0.04 seconds of audio. The generated frames were then returned

to the Python interpreter for further processing.

3.4 Frame Merging

Once a group of lip-synced frames had been generated from the audio, the frames were sent to a frame queue. On each step of the video being played, the Frame Player checks the frame queue for new frames. If there is a new frame in the queue, it merges it into the currently selected one using a Dlib Face Swap technique[4]. The video-frames' Dlib facial landmarks were calculated pre-runtime. Once the frame had been merged, or if there was no need to merge frames, the resulting frame was then passed to a pre-trained Pix2Pix GAN network, for inference of the final output image. The output frame is then sent to the screen to view as video. Figure 6 shows an overview of the pipeline.

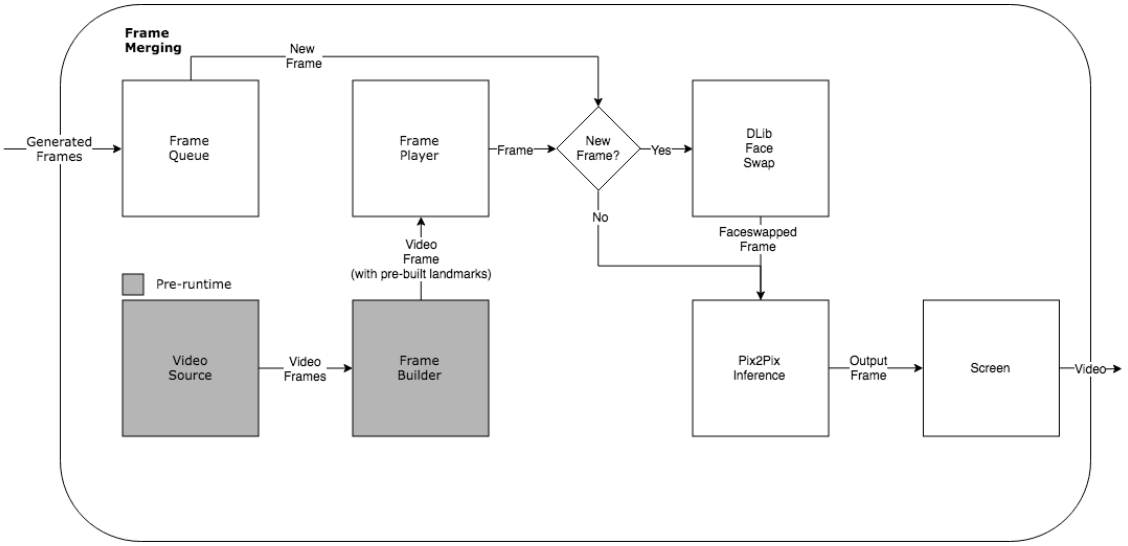

Figure 6: Frame Merging Pipeline

3.4.1 Dlib Face Swap

To merge the generated frames and the video frames the Dlib machine learning toolkit was used [41], along with a 68 point Dlib facial landmark detector.

The facial landmarks for each frame had to be predicted, and a mask of the area

[4]http://matthewearl.github.io/2015/07/28/switching-eds-with-python/

that was going to be swapped had to be built. To build the mask, Dlib facial points 28-36 (the nose) and 49-62 (the mouth) were used, and the edges of the mask were smoothed using a Gaussian kernel of 11x11 pixels. The landmarks and masks were built pre-runtime for the video frames, and during the non-runtime thread for the generated frames (during the Speech-To-Video module in Figure 3). This allowed for a real-time face-swap of the frames.

The transform itself involved 5 steps:

First of all the orthogonal Procrustes problem was solved, in order to find the transformation matrix that best mapped the facial landmarks from the generated frame to the video frame [44].

Secondly, the Procrustes transformation was then applied to the generated frame in order to map it onto the video frame.

Third, the colors of resulting image were corrected using an RGB scaling color correction technique. This works by dividing the generated frame image by a Gaussian blur of itself, before multiplying it by a Gaussian blur of the video frame image. Selecting the appropriate size of the Gaussian kernel is an important factor here, as it dictates how much the facial features from the generated image will show. A value of 0.5 times the inter-pupillary distance was used (the distance from Dlib landmarks 43-48 to 37-42).

Fourth, the mask for the generated image is transformed using the previously calculated Procrustes transformation. It is then combined with the mask of the video frame by taking the element-wise maximum.

Finally, the combined mask is then applied to the corrected color image from the second step, and the image is normalised (using min-max normalisation) to produce the output image.

3.4.2 Pix2Pix GAN Inference

The Pix2pix image translation model [20] has had success in enabling a wide variety of image generation, including that of synthesizing human faces [5]. It implements a conditional GAN framework that can perform image-to-image translation in real-time. In order to train the network it takes an image of the desired input and desired output side-by-side (See Figure 7).

The Pix2Pix model was trained on the first 400 frames from a video recorded at 30 FPS in which the actor recited a book [45] (which was also used to extract the mouth-frames for the Speech2Vid model). The photos were pre-processed by cropping them to 256x256 pixels and extracting all 68 Dlib facial landmarks. The model was trained over 400 epochs in 5 hours 30 minutes, with the default parameters as set in [46].

Figure 7: Example of training image fed to the Pix2Pix network

3.5 Frame Playing

Finally, To display continuous video to the user, a pre-loaded video was run on repeat by the Frame Player. In order to be able to play the video indefinitely, once the Frame Player had reached the last video frame, the order of the frames was

[5] https://github.com/datitran/face2face-demo

reversed and playback continued. This allowed for a continuous smooth transition between frames, and removed the jerkiness that can occur when transitioning from the last to the first frame of a video. An improvement would be to use the trick to select the video frames mentioned before when describing Microsoft's E-Partner system [11].

Chapter 4

Evaluation

4.1 Overview

The system was able to run in real-time, at up to 30 FPS. As the audio generated by the Speech2Vid model had a length of 0.04 seconds, the system was configured to run at 25 FPS. Dialogflow had a mean latency of 27 milliseconds and each second of audio received from Dialogflow took a mean of 2.7 seconds to process (to generate the frames, landmarks e.t.c). An example image at each stage of the pipeline is shown in Figure 8.

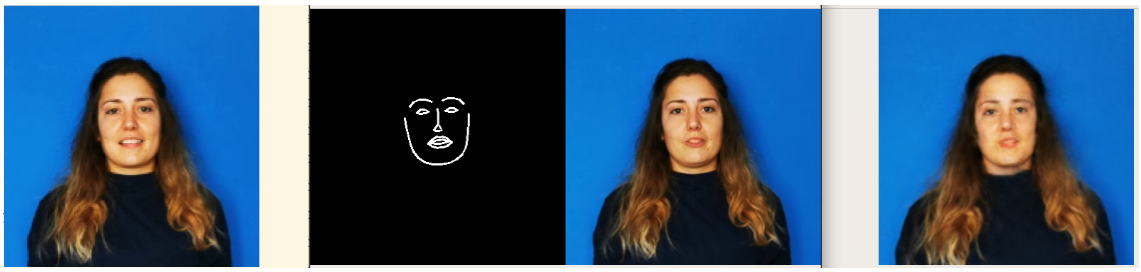

Figure 8: Video Frame (Left) Dlib Landmarks (Left Centre) Face Swap (Right Centre) Pix2Pix Inferred Frame (Right)

It can be seen that during the Dlib face swap, clarity is lost in the actors mouth region, this is because of the low resolution output image of the Speech2Vid model. In the Pix2Pix inferred frame, a further loss of clarity is accrued.

4.2 User Test

In order to evaluate the system, tests were conducted across 10 participants, it was important that the participants had no recognition of the person the system was synthesizing. To conduct the test, the participants were asked to converse with the system for 1 minute. After the conversation, the participants were asked to rate different aspects of the system, speech, dialogue and visuals, using a mean opinion score (MOS) from 1 to 5, 1 being completely artificial, and 5 being completely human like. They were also asked to give an MOS score to the system as a whole. A summary of the scores is shown in Table 1.

Table 1: MOS Score Summary

Component	**Min MOS**	**Max MOS**	**Mean MOS**
Speech	3	5	3.7
Dialogue	1	4	2.5
Visuals	2	4	3.2
Overall	2	4	2.9

The Speech component scored the highest (3.7), however it scored lower than the MOS score as reported by the Wave-net paper of 4.41 [34]. The Dialogue model performed worst overall, interesting to note is that a user had issues with the interaction of the system, attempting to interrupt the dialogue, which although the system is designed to handle, the computer's sound-card did not, meaning that the system missed what the user said, they rated the Dialogue model as 1/5. The visuals achieved a Mean MOS of 3.2, the second highest after the speech. The system overall had a Mean MOS of 2.9.

Chapter 5

Discussion

5.1 System

The goal of constructing a system able to synthesize a human being was partly achieved. The prototype worked, and users were able to interact with it successfully for a minute at least. An MOS score of 2.9 suggests that it was able to synthesize features of a human to some degree. However it has a way to go to be able to indistinguishably pass as a human being.

Interesting to observe from the MOS results, is that components such as Speech received a lower MOS score than when evaluated independently by [34]. Perhaps this suggests that when combining components, human perception of the individual components is affected by the MOS scores of the rest of the components.

Further evaluation proposed is to increase the user sample size for the MOS component test in order to investigate how the MOS of individual components in a system may affect other components MOS, as well as the MOS of the system as a whole. This could help understand which part of the system to improve next. Research such as [15] have evaluated their work (a speech to video frame model) through an online total Turing test (a Turing test that includes audio and video) [5]. This would be an interesting evaluation technique as well. Furthermore, it would be interesting to evaluate how the scores of users who know the person being synthesized compares

to that of a user who does not know the person. One can imagine that a natural next step for a machine in a total Turing test is to not only be be perceived as a human, but perceived as a recognisable human.

Comparing with previous work, Microsoft research's in particular [11], it's interesting to see which parts have advanced the most.

First of all, the TTS system has clearly improved by using Wavenet. Secondly, instead of using an image retrieval system to generate the output image, facial landmarks were predicted using the Speech2Vid model [24]. The predicted landmarks just showed lip movements, and therefore an improvement for the system is to be able to predict more facial expressions. This could be done by implementing [26] for example, which generates facial landmarks from raw waveforms of speech, with varying degrees of emotional intensity. It would be interesting to evaluate whether the landmarks are good enough for lip-sync or whether they could be combined with the lip-synced landmarks from the Speech2Vid model [24]. Furthermore, whether the prediction quality is reduced when predicting from synthesized speech vs actual speech should be investigated. It was not found not to be an issue in the current system, but it was not quantified either. If landmark prediction of emotional facial expressions can be achieved, it would allow for more believable synthesis of humans.

Using the Pix2Pix model for output image inference sacrificed clarity of the image in favour of better image continuity (across the pixel space domain and time domain). Future iterations of the GAN inference should be able to produce clearer images in real-time. However, I think it is important to avoid models that require a long amount of training on a subject dependent dataset, because then the advantage against model-based approaches, of not having to spend time modelling an individual subject, is lost. Investing in subject independent approaches such as [15] or the Speech2Vid model allows for the system to become flexible and re-usable.

Finally, the dialogue model was difficult to compare, as it was constructed with a specific objective of being a tour guide in [11]. However, the dialogue models closest to passing the Turing test remain rule-based, and I think the key part here, is that

the selection of dialogue model depends on the objective of the system. In the case of the system presented here, the Eliza rules were suffice to keep the user engaged for a minute of conversation.

5.2 Ethics

The current age we are living in is being labelled by some as the age of misinformation [47]. Misinformation in terms of news, it has been argued by [4] that false stories dubbed as "Fake News" shared on Facebook[1] influenced the 2016 US presidential election. More recently in 2018, the generation of Fake News has been attributed to the generation of hate speech in Myanmar, leading to what the UN is calling genocide against Rohingya Muslims [48]. Furthermore, technology allowing human image synthesis in videos labelled "Deep Fakes", has been used to create what is known as "Revenge Porn", the distribution of sexually explicit images or video of individuals without their permission [49]. These are not issues to be taken lightly, and given that this research is based on generating a human-like interface, very relevant.

5.3 Data Rights

It is clear that we are living in a time where data-privacy and the ethics of data use is at the top of societies' agenda. Evidence of this include the controversy of the 2018 net neutrality changes in the US and India or the general data protection act passed in Europe. The system that this paper discusses not only depends on copyright data, such as images of people, but produces data whose copyright is uncertain. This is because under current law, in the US, Germany and England for example, it is not possible for the AI to be considered the author of the work [50]. Finding who the author of the work varies depending on the legal system, and can often remain ambiguous. For example, in England the person who made the arrangements necessary for the work to be created becomes the author. However, in the common case of a developer using an AI model which implements an algorithm designed by

[1] https://www.facebook.com/

someone else, the author is not clear. The rapid progress of algorithms such as the Pix2Pix model which allow easy content creation by an AI algorithm, mean that the legal systems have yet to adapt to find an efficient way of determining who is accountable for what. The problem with this is that it allows moral complacency and lack of responsibility when using the algorithms. For example, if the justice system is not able to conclude whether you are to be held accountable for creating a Deep Fake Revenge-Porn video, then it may detract from the terribly unethical context of it. However, the legal system being too slow to adapt to emerging technology, should not be used as an excuse to morally justify an action.

So what should you do if you are not sure and there is no legal precedent to guide you? The case of Revenge Porn is extreme, but what about a more subtle case, such as that of using an AI algorithm to generate a photo from another? I think in these cases, understanding the data involved is important, and ensuring that you communicate with the stakeholders of the data to ensure that they approve of any proposed changes. Also, it should be encouraged to only use pre-trained AI models that have been trained on data which the person doing the training had permission for.

5.4 Breakdown of human assumption

In chapter 1.2 Motivations, it was explained that the assumption of interacting with another human when interacting in a human-like way is now being broken. The implications of this assumption being broken lead to more ethical issues.

It is hard to envision many use cases where deceit is not involved when synthesising aspects of other people. For example with Google Duplex [9], the system is used to call a restaurant pretending to be human, and book a table for its respective owner. However, the person answering the call has no idea that they are talking to a machine. Is this ethically correct? Although Google seems to think so, I can't help but feel that deceit lies at the heart of this interaction.

On the one hand, the Duplex system is being controlled by a human through imper-

ative voice commands (e.g. "Book me a table for X at restaurant Y"). Therefore one could argue that the interaction process, the cooperation involved to book a table, is no different from booking a table through the restaurant's website. Either way the restaurant does not know the actual human on the other end of the interaction. However, I think the issue lies a bit deeper than this.

At the heart of human-like interfaces lies human cooperation [51]. This cooperation depends heavily on interactions with between humans, and over tens of thousands of years humans have built an extremely complex interaction model which allows them to make decisions on whether to cooperate or not. The issue with synthesizing human-like interfaces through the power of computing, means that our interaction models are at risk of becoming inadequate and even being exploited. To explain this further, let's go back to the restaurant booking example.

When the restaurant person picks up the phone and answers a Google Duplex call, they prepare their interaction and cooperation model based on if they were interacting with a human. One assumption of this model for example, is that the human they are interacting with cannot call them up 100 times with different voices and book every table, whilst booking every other table of every other restaurant in town at the same time. The Google Duplex system can take advantage of this assumption for example to do just that, wrecking havoc on the restaurant business. This is just one trivial assumption picked at random, but I hope it starts to make more obvious how the human interaction model can be exploited.

The first step to ensuring people aren't taking advantage of, is for them to understand that they can be taken advantage of. That is why it is important for these systems to make it clear that they are not human, and for people to learn of their capabilities. Unfortunately, this will probably be learned the hard way, and has been an issue since the start of the internet, for example, email phishing scams. However, ultimately, at least for now, there is a human commanding the machine.

5.5 Conclusion

This work has investigated the latest in human-synthesis technology and developed a working photo-realistic voice-bot. The evaluation of the prototype yielded an MOS score of 2.9 across 10 participants, and although the work was not successful in making the prototype indistinguishably human-like, the work also explored whether it is ethically correct to do so, arguing that identifying the data stakeholders is important as well as educating others about the capabilities of the technology.

To improve the prototype, future work could begin by improving the speech to video model, so that it predicts a wider variety of facial landmarks, allowing for more complex facial expressions, this could be done by integrating [27] or [26]. The quality of the image synthesis, which is currently a bit blurry, could be improved by further training of the Pix2Pix model, however it would be more valuable to find a facial-landmark to image model, that does not require retraining on individual subjects. This would be an interesting research topic and would also benefit the system. To improve the text-to-speech module, although the quality of the Wavenet model is high, it could be improved by adapting it to be able to clone the subject's voice, as proposed in [35].

Finally, to improve the dialogue model, it really depends on the use-case of the system. For example, if it has a role of needing to have long, autonomous conversations, then a more complex model than the Eliza based rules should be found. However, the system could be used without an autonomous dialogue model. For example, the system could be configured in a Wizard-Of-Oz setup, where a user could control the synthesis through a text-based user interface. A use-case for this could be to help those who have trouble communicating with other people themselves, but who want to retain their sense of identity.

List of Figures

List of Tables

Bibliography

[1] Asimov, I. *The caves of steel: by Isaac Asimov* (T.V. Boardman, 1954).

[2] *Her* (Annapurna Pictures, 2013).

[3] Matias, Y. & Leviathan, Y. Google duplex: An ai system for accomplishing real-world tasks over the phone (2018). URL `https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html`.

[4] Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* **31**, 211–36 (2017). URL `http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211`.

[5] Powers, D. M. W. The total turing test and the loebner prize. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, 279–280 (Association for Computational Linguistics, Stroudsburg, PA, USA, 1998). URL `http://dl.acm.org/citation.cfm?id=1603899.1603947`.

[6] Wang, L., Han, W., Soong, F. & Huo, Q. Text-driven 3d photo-realistic talking head (International Speech Communication Association, 2011). URL `https://www.microsoft.com/en-us/research/publication/text-driven-3d-photo-realistic-talking-head/`.

[7] Pandzic, I. S., Ostermann, J. & Millen, D. User evaluation: Synthetic talking faces for interactive services. *The Visual Computer* **15**, 330–340 (1999). URL `https://doi.org/10.1007/s003710050182`.

[8] Muhammad, A., Ahmad, W., Tooba, M. & Anwar, S. Assistive technology for disabled persons. In *International Conference on Recent Advances in Computer Systems* (Atlantis Press, 2015/11). URL `https://doi.org/10.2991/racs-15.2016.12`.

[9] O'Leary, D. Google's duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management* **26**, 46–53 (2019).

[10] Cosatto, E. & Graf, H. Photo-realistic talking-heads from image samples. *Multimedia, IEEE Transactions on* **2**, 152 – 163 (2000).

[11] Zhang, B., Liu, Z. & Guo, B. *Photo-Realistic Conversation Agent*, 219–240 (Springer US, Boston, MA, 2004). URL `https://doi.org/10.1007/1-4020-7775-0_12`.

[12] Fan, B., Xie, L., Yang, S., Wang, L. & Soong, F. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications* (2015).

[13] Hartholt, A. *et al.* All Together Now: Introducing the Virtual Human Toolkit. In *13th International Conference on Intelligent Virtual Agents* (Edinburgh, UK, 2013). URL `http://ict.usc.edu/pubs/All%20Together%20Now.pdf`.

[14] Wiles, O., Koepke, A. & Zisserman, A. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision* (2018).

[15] Vougioukas, K., Petridis, S. & Pantic, M. End-to-end speech-driven facial animation with temporal gans. In *BMVC* (2018).

[16] Anderson, R., Stenger, B., Wan, V. & Cipolla, R. Expressive visual text-to-speech using active appearance models (2013).

[17] Liu, K. & Ostermann, J. Realistic facial expression synthesis for an image-based talking head. In *2011 IEEE International Conference on Multimedia and Expo*, 1–6 (2011).

[18] Thies, J., ZollhÃűfer, M., Stamminger, M., Theobalt, C. & NieÃ§ner, M. Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2387–2395 (2016).

[19] Averbuch-Elor, H., Cohen-Or, D., Kopf, J. & Cohen, M. F. Bringing portraits to life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)* **36**, 196 (2017).

[20] Isola, P., Zhu, J., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. *CoRR* **abs/1611.07004** (2016). URL `http://arxiv.org/abs/1611.07004`. `1611.07004`.

[21] Kumar, R., Sotelo, J., Kumar, K., de Brébisson, A. & Bengio, Y. Obamanet: Photo-realistic lip-sync from text. *CoRR* **abs/1801.01442** (2018). URL `http://arxiv.org/abs/1801.01442`. `1801.01442`.

[22] Suwajanakorn, S., M. Seitz, S. & Kemelmacher, I. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics* **36**, 1–13 (2017).

[23] Fried, O. *et al.* Text-based editing of talking-head video. *ACM Trans. Graph.* **38**, 68:1–68:14 (2019).

[24] Jamaludin, A., Chung, J. S. & Zisserman, A. You said that? : Synthesising talking faces from audio. *International Journal of Computer Vision* (2019).

[25] Chung, J. S. & Zisserman, A. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV* (2016).

[26] X. Pham, H., Wang, Y. & Pavlovic, V. End-to-end learning for 3d facial animation from raw waveforms of speech (2017).

[27] Karras, T., Aila, T., Laine, S., Herva, A. & Lehtinen, J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.* **36**, 94:1–94:12 (2017). URL `http://doi.acm.org/10.1145/3072959.3073658`.

[28] Gold, B. *Speech and audio signal processing* (Wiley-Blackwell Publishing, 2011).

[29] Mattheyses, W. & Verhelst, W. Audiovisual speech synthesis. *Speech Commun.* **66**, 182–217 (2015). URL http://dx.doi.org/10.1016/j.specom.2014.11.001.

[30] Prenger, R., Valle, R. & Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621 (2019).

[31] Sotelo, J. *et al.* Char2wav: End-to-end speech synthesis. In *ICLR (Workshop)* (OpenReview.net, 2017).

[32] Li, N. *et al.* Close to human quality TTS with transformer. *CoRR* **abs/1809.08895** (2018). URL http://arxiv.org/abs/1809.08895. 1809.08895.

[33] Shen, J. *et al.* Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR* **abs/1712.05884** (2017). URL http://arxiv.org/abs/1712.05884. 1712.05884.

[34] van den Oord, A. *et al.* Wavenet: A generative model for raw audio. *CoRR* **abs/1609.03499** (2016). URL http://arxiv.org/abs/1609.03499. 1609.03499.

[35] Ar, S. O., Chen, J., Peng, K., Ping, W. & Zhou, Y. Neural voice cloning with a few samples. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, 10040–10050 (Curran Associates Inc., USA, 2018). URL http://dl.acm.org/citation.cfm?id=3327546.3327667.

[36] Jin, Z., Mysore, G. J., Diverdi, S., Lu, J. & Finkelstein, A. Voco: Text-based insertion and replacement in audio narration. *ACM Trans. Graph.* **36**, 96:1–96:13 (2017). URL http://doi.acm.org/10.1145/3072959.3073702.

[37] Levesque, H. J. *Common sense, the Turing test, and the quest for real AI* (MIT Press, 2017).

[38] Venkatesh, A. *et al.* On evaluating and comparing conversational agents. *CoRR* **abs/1801.03625** (2018). URL `http://arxiv.org/abs/1801.03625`. `1801.03625`.

[39] Li, J., Galley, M., Brockett, C., Gao, J. & Dolan, B. A persona-based neural conversation model. *CoRR* **abs/1603.06155** (2016). URL `http://arxiv.org/abs/1603.06155`. `1603.06155`.

[40] Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966). URL `http://doi.acm.org/10.1145/365153.365168`.

[41] E. King, D. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* **10**, 1755–1758 (2009).

[42] Afouras, T., Chung, J. S., Senior, A., Vinyals, O. & Zisserman, A. Deep audio-visual speech recognition. In *arXiv:1809.02108* (2018).

[43] Vedaldi, A. & Lenc, K. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia* (2015).

[44] Berge, J. J.c. gower and g.b. dijksterhuis.procrustes problems. new york: Oxford university press. *Psychometrika* **70** (2005).

[45] Seuss, D. *Green eggs and ham* (HarperCollins Childrens Books, 2016).

[46] Hesse, C. Project title. `https://github.com/affinelayer/pix2pix-tensorflow` (2013).

[47] OConnor, C. & Weatherall, J. O. *The misinformation age: how false beliefs spread* (Yale University Press, 2019).

[48] Myanmar military leaders must face genocide charges âĂŞ un report. *UN News* (2018). URL `https://news.un.org/en/story/2018/08/1017802`.

[49] Anne Franks, M. Criminalizing revenge porn: A quick guide. *SSRN Electronic Journal* (2013).

[50] Stephens, K. & Bond, T. Artificial intelligence navigating the ip challenges. *PLC Magazine* (2018).

[51] Rand, D. G. & Nowak, M. A. Human cooperation. *Trends in Cognitive Sciences* **17**, 413 – 425 (2013). URL `http://www.sciencedirect.com/science/article/pii/S1364661313001216`.