Lip-syncing Videos In The Wild

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

K R Prajwal
2019701002
prajwal.k@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2020

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Lip-syncing Videos In The Wild"** by K R Prajwal, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. C V Jawahar
Date	Adviser: Prof. Vinay P. Namboodiri

To my beloved parents

Acknowledgments

I would like to wholeheartedly thank the people who have made my journey memorable and meaningful. First and foremost, I am deeply indebted to my advisors, Prof. C V Jawahar, Prof. Vinay P. Namboodiri, and Prof. Ponnurangam Kumaraguru, for their constant guidance and support. Throughout my research journey, they have allowed me the freedom to grow and learn by making decisions. When I look back today, I am most thankful to them for inspiring me to venture out of my comfort zone, thus making me a better researcher. I will forever cherish our enjoyable, thought-provoking discussions that have profoundly shaped my research outlook.

I also wish to thank my friends, starting with my partner-in-crime, Rudrabha. I really could not have asked for a better team-mate. Teaming up with someone similar in character, who is diligent and light-hearted with countless entertaining stories to tell along the way has been undoubtedly one of the best things to have happened to me. I thank you for all the memories. This acknowledgment would be incomplete without thanking Harish Rithish (both are the same person) for being my pillar of strength as you have always been for so many years. I cannot be grateful enough to Sindhu Hegde for making my MS courses seem like a cakewalk, for all your curious questions, and, most importantly, for accompanying me to South mess on Sundays. I also wish to thank Kiran, Ashish, Deepak, and Riya for making each day at CVIT stimulating, with countless pranks and laughter. I am very grateful to my friends at the ARL lab, Sangeeth, Tarun, Vishnu, and Lovish, for being there always whenever I needed them.

Finally, it is impossible to ever fully thank the super-heroes behind each and every one of my successful endeavors: my parents and my sister for their selfless affection, support, advice, and making sure I am always happy.

Abstract

The widespread access to the Internet has led to a meteoric rise in audio-visual content consumption. Our content consumption habits have changed from listening to podcasts and radio broadcasts to watching videos on YouTube. We are now increasingly preferring the highly engaging nature of video calls over plain voice calls. Given this considerable shift in desire for audio-visual content, there has also been a surge in video content creation to cater to these consumption needs. In this fabric of video content creation, especially those containing people talking, lies the problem of making these videos accessible across language barriers. If we want to translate a deep learning lecture video in English to Hindi, it is not only that the speech should be translated but also the visual stream, specifically, the lip movements.

Learning to lip-sync arbitrary videos to any desired target speech is a problem with several applications ranging from video translation, to readily creating new content that would otherwise require humongous efforts. However, speaker-independent lip synthesis for any voice, and language is a very challenging task. In this thesis, we tackle the problem of lip-syncing videos in the wild to any given target speech. We propose two new models in this space: one that significantly improves the generation quality and the other significantly improving on lip-sync accuracy.

In the first model, LipGAN, we identify key issues that plague the current approaches for speaker-independent lip synthesis that prevent them from reaching the generation quality of speaker-specific models. Specifically, ours is the first model to generate face images that can be pasted back into the video frame. This feature is crucial for all the real-world applications where the face is just a small part of the entire content being displayed. We show that our improvements in quality lead to multiple real-world applications that have not been demonstrated in any of the previous lip-sync works.

In the second model, Wav2Lip, we investigate why current models are inaccurate while lip-syncing arbitrary talking face videos. We hypothesize that the reason is weak penalization. This finding allows us to create a lip-sync model that can generate lip-synced videos for any identity and voice with remarkable accuracy and quality. We re-think the current evaluation framework for this task and propose multiple new benchmarks, two new metrics, and a Real world lip Sync Evaluation Dataset (ReSyncED). Also, using our model, we show applications on lip-syncing dubbed movies and animating real CGI movie clips to new speech. We also demonstrate a futuristic video call application that is useful for poor network connections. Finally, we present two major applications that our model can impact the most:

social media content creation and personalization and video translation. We hope that our advances in lip synthesis open up new avenues for research in the space of talking face generation from speech.

Contents

List of Figures

List of Tables

Table Page

Chapter 1

Introduction

The wide-spread access to the Internet has altered the way we create and consume content. Every individual with access to the web now has the power to create their own content and spread their ideas across the world almost instantaneously. Over the years, the development of powerful smart devices has fuelled significant changes in the type of content being created and consumed. For instance, textual content, such as blogs and news articles, peaked over ten years ago. Very soon, as the internet speeds rose quickly, music, podcasts, and static images became seamlessly accessible. Today, we are witnessing another such shift: the meteoric rise in streaming video content.

Video-related content is the second most popular search query as of 2019, next only to Facebook. YouTube, the leading video streaming platform, streams over a billion hours of video content a day, equalling the number of Google searches a day! Catering to this demand are the video content creators, who upload over 500 hours of video content every minute. A large chunk of this comprises of "talking face videos", i.e., videos containing a human speaking, such as lecture videos, movies, press conferences, vlogs, etc. One of the major inherent limitations of talking face videos is that they are inaccessible to viewers of different languages, thus significantly limiting their impact. For example, a lecture video in English can be invaluable to a viewer who knows only Hindi if it is translated. While there have been significant efforts in automatically translating speech, there are no works that handle the visual modality in this problem, i.e., altering the speaker's lip movements to match the desired target speech. Even if the speech is manually dubbed, correcting the lip movements in each video frame is a task that demands significant manual effort.

The focus of this thesis is to propose solutions that enable us to automatically lip-sync a talking face video in the wild to match the desired target speech. Besides just translation, such solutions can be used to create new talking faces that can have applications in video conferencing and CGI character animations. With these applications in mind, the thesis presents two solutions, each solving some of the major issues in existing methods. The first work allows us to produce high-quality lip-synced videos while preserving fine-grained visual features such as pose, expressions, and background information. The second work produces significantly more accurate lip-syncing for videos in the wild. Both of these works together aim to produce lip-synced videos of arbitrary identities in the wild with remarkable accuracy and quality. We show that our model can be employed in a variety of real-world applications such as lip-syncing dubbed movies and live-translated press conferences, translating lecture videos, and also low-bandwidth video conferencing.

1.1 Scope of the Thesis

1.1.1 Problem Definition and Challenges

This thesis tackles the problem of speaker-independent speech to lip generation in the wild. Given a video (or an image) of a face and a target speech segment, the goal is to generate a lip-synced video of the face that matches the target speech. There are no explicit constraints in terms of the speaker identity, language, or the voice of the speech. Naturally, learning to synthesize lip movements from speech in such unconstrained settings presents numerous challenges.

- To start with, as our task is only to perform lip-sync, the other visual attributes such as the input face's identity, pose, and expressions are expected to remain unaltered in the generated video. This is quite essential. For instance, if the input is a sequence of faces cropped from video frames, then the preservation of pose, expressions, and identity will allow us to blend the generated faces back into the video frames seamlessly.

- The generated videos need to be temporally coherent. The model needs to generate accurate lip shapes across rapid pose and illumination changes. Producing temporally coherent videos is challenging as it can be seen that past works are incoherent in pose and visual information and produce inaccurate lip motion during silent segments.

- The problem possesses large variations in both modalities. The visual space can vary in terms of identity, pose, background, expressions, and reference lip shapes. The speech can be of any voice, language, content, prosody, accent, and can be even synthetically generated (from a text-to-speech model) or have a fair amount of background noise.

- Finally, precisely evaluating the accuracy of the generated lip-synced videos is still not explored in the current literature. A reliable evaluation framework is crucial to measure the progress in a task. The existing evaluation framework does not reflect the real-world setting and is far from ideal for evaluating lip-sync in unconstrained videos.

1.1.2 Contributions

This thesis tackles each of the above challenges and culminates in a novel, robust model for generating accurate lip-synced videos of arbitrary identities and voices in the wild. Our key contributions are summarized below.

- We propose two novel models aimed at tackling the above challenges. The first model generates high-quality lip-synced videos that also preserves the visual attributes such as pose and expressions. This is the first speaker-independent speech-driven lip synthesis work that allows the generated face images to be seamlessly pasted back into the original video frame.

- The second model investigates the key issues in the first one that hamper it from achieving accurate lip generation for videos in the wild. Consequently, we propose the first model that can accurately lip-sync arbitrary videos in the wild with any target speech.

- We re-examine the current evaluation framework for evaluating lip-sync in talking face videos. Upon finding multiple major shortcomings in the current system, we propose multiple new evaluation benchmarks and metrics for accurate and reliable evaluation of lip-synced videos in the wild.

- We showcase a variety of real-world applications made possible by our final model. We can accurately lip-sync dubbed movies and even live translations in press conferences. We can also animate the lips of CGI characters and lip-sync lecture videos with automatically dubbed synthetic TTS speech. Further, we also show results in two novel applications that are most impacted by our model: “Face to Face Translation” and “Social Media Content Creation”. We attach demo video links containing all these applications.

1.2 Background

We first review the works in speaker-dependent and speaker-independent speech-driven talking face generation. We also survey the various datasets available for our task.

1.2.1 Speaker-specific Talking Face Generation

Generating talking faces driven by audio has been a long-standing problem, first introduced in [8]. Here, frames from a given training video footage of a person are re-ordered to match the phonemic sequence in the new audio. It is further stitched with the background to create the final output. While this was performed in constrained laboratory settings on individual speakers, a recent work [53] using deep learning kick-started efforts for achieving realistic lip-sync in real-world environments. They learn a mapping from the speech of Obama to his lip landmark sequence. The lip landmarks sequence is then synthesized into face frames. This work was personalized for presidential addresses of Obama, and hence also performed head pose correction using re-timing techniques to finally generate high-quality composite videos. Another similar work [44] proposes to lip-sync using random text rather than using Obama’s speech, thus allowing it to generate new sentences that were not actually spoken by Obama. A very recent work [26] involving text and faces, tackled a modification of this problem, where they propose to seamlessly edit a video if some words are added or removed. They require one hour of speaker data for each speaker they want to generate on. A follow-up work [54] of this proposed to learn a speaker-specific network to achieve better visual quality and a speaker-independent network that allows for downstream adaptation to other speakers using a lesser amount of data. However, this deteriorates the lip-sync accuracy significantly. There have also been works [25] specifically designed for generating talking face videos of cartoon characters. We believe that learning specifically on cartoon videos is not ideal, as lip-sync in cartoon videos is already inaccurate. On the other hand, transferring learning from human faces to cartoon faces is more reasonable, and we will show that this can produce good cartoon talking face videos.

1.2.2 Vocabulary-specific Talking Face Generation

There are a number of works that train on datasets that have a more diverse set of speakers but are severely constrained by vocabulary. For instance, a recent work [60] involving generating talking faces using GANs was trained and evaluated on GRID [19] and TIMIT [30] dataset. Both these datasets are speaker-specific and contain $50 - 100$ words per speaker. A follow-up [61] work generalized to a more diverse set of speakers, but the vocabulary was still limited to a thousand words in English from the LRW dataset [17]. The same fact is also true for another work [13] that learns a speech to landmark mapping followed by a temporal generator. We find that training on limited vocabulary and voices significantly reduces the range of speech variations in the real-world that these models can handle. Another work [70] using GANs approaches the problem slightly differently. They use audio-visual speech recognition as a probe task for associating audio-visual representations, and then employ adversarial learning to disentangle the subject-related and speech-related information inside them. They are also limited to 500 English words, but more importantly, they need textual transcripts for training, as they use the speech recognition task for discrimination. We observed that this makes their approach language-dependent. It also becomes hard to reproduce this model for other languages as collecting large video datasets with careful word-level annotated transcripts in various languages is infeasible. A more generic approach would be to devise a fully self-supervised approach that directly learns from phonemes in the speech.

1.2.3 Generating Lip-synced Talking Face Videos In The Wild

This task is the primary focus of this thesis. Given an image (or) a video of an arbitrary person, and an arbitrary driving speech, the task is to generate a lip-synced video that matches the given speech. All the previous works impose different forms of constraints. Some of them are designed to work on only specific speakers, requiring several hours of data for each speaker. This data overhead does not allow them to generate for the multitude of applications in the real world involving arbitrary identities. They only work for the speaker's voice and are inaccurate even if the same voice is synthetically generated, inhibiting them from applications like lip-syncing automatically dubbed content. On the other end, a constraint on the vocabulary can severely limit the model's performance on real-world speech. For instance, the most common and direct application of lip-syncing is to sync translated content. To generalize across languages without training on each language separately, the training speech must be diverse in voices and phonemes. Other similar applications will also be sub-optimal with limited vocabulary training, such as lip-syncing synthetic speech or lip-syncing songs.

In light of these drawbacks, there has been a recent attempt, "You Said That?" [14] that aims to generate talking face videos of arbitrary identities driven by arbitrary speech. Its core architecture comprises an encoder-decoder model that takes in a short audio segment (0.35 seconds) and a reference frame image and morphs the lip shape. Our first model significantly improves upon this work in several key aspects. One of the critical aspects is that the Speech2Vid model proposed in this paper does not preserve the pose, thus not allowing the generated face to be pasted back into the video frame. We will investigate this further in the next chapter. That being said, this is the first work that is truly "unconstrained," i.e., it is not limited by the range of identities, vocabulary, language, or voices as the model has been trained on a subset of the VoxCeleb [48] dataset that contains thousands of speakers. More details about the various datasets for our task is detailed in the next sub-section.

1.2.4 Datasets for Talking Face Generation

Learning to lip-sync videos is "freely-supervised", i.e., it does not need manually annotated data. It can be learnt from videos of people speaking downloaded from the Internet. However, many videos online can be slightly out of sync. Before we can accurately learn to lip-sync, careful pre-processing is necessary to ensure the data consists of well-synced videos. In recent years, many datasets of talking face videos have been released for research work. The dataset on which a lip-sync network is trained on strongly determines its generalization to real-world videos. Thus, we review the major datasets available for talking face generation. We start with speaker-specific and constrained vocabulary datasets such as GRID [19] and TIMIT [30]. They consist of lab-recorded videos with a vocabulary of less than 100 words. The LRW [17] dataset, on the other hand, consists of real-world videos, but it contains mostly fully-frontal videos, and more importantly, it is limited to a vocabulary of 1000 words, thus biasing any models trained on this dataset towards these words. A final class of datasets that are truly unconstrained in speakers and vocabulary has also been released. VoxCeleb [48] is a large-scale speaker identification dataset with thousands of speakers and 100K+ utterances. AVSpeech [24] is an even more diverse and much larger dataset with 4000 hours spread across several languages all over the world. The feature that sets this apart from VoxCeleb [48] is that it is of a much higher resolution with videos in full-HD compared to low-resolution videos in VoxCeleb. A larger version with more diversity and more number of hours was released in VoxCeleb2 [15]. While all of these datasets are large-scale and diverse for our task, they have not been curated for learning lip sync. As stated before, videos downloaded from the internet can be out-of-sync. Hence, these videos need to be explicitly lip-sync corrected. A fraction of VoxCeleb2 [15] was lip-sync corrected with the help of SyncNet [18]. Videos with lip-sync errors were discarded, and the final cleaned set led to the creation of the LRS dataset [16]. This was exclusively created for lip reading and speech-driven lip synthesis. A larger version of this dataset, namely LRS2 [1], was also released. A later version of the dataset, namely LRS3 [4], was also released publicly, but this consisted of TED videos in indoor stages, with minimal diversity in backgrounds. Both these last two large-scale datasets, namely the LRS2 [1] and LRS3 [4], have been explicitly lip-sync corrected and can be used for our task at hand to achieve accurate lip-sync. They contain a large number of identities with large variations in speech content. Throughout this thesis, we use the LRS2 dataset [1] for training purposes, as it is recorded in a variety of real-world backgrounds, while LRS3 [4] is constrained to indoor stages.

We conclude our survey on talking face generation with the table below comparing existing works with the models proposed in this thesis.

Paper	Speaker Independent	Language Independent	"Freely"-supervised	Smooth blending into video	Accurate for in-the-wild videos
Synth. Obama [53]	×	×	✓	✓	✓
Obamanet [44]	×	×	×	✓	×
Puppetry [54]	×	✓	✓	✓	×
Three-GANs [61]	×	×	✓	×	×
TalkingFace [70]	✓	×	×	×	×
Speech2Vid [14]	✓	✓	✓	×	×
LipGAN (Ours)	✓	✓	✓	✓	×
Wav2Lip (Ours)	✓	✓	✓	✓	✓

Table 1.1 Comparison of recent works on talking face synthesis against our proposed models. LipGAN is the first model to generate lip-synced face images of arbitrary identities/languages that can be blended into the video frames. Our final model (Wav2Lip) improves significantly over LipGAN and is the first model to produce accurate lip-synced talking face videos of arbitrary identities and speeches in the wild.

1.3 Related Works on Learning Multi-modal Correspondences

We conclude our literature survey by also reviewing the works connected to talking face generation or learning correspondences between speech and human gestures (lips, hands, and head pose). To start with talking face generation, a recent work [66] proposes to generate talking face videos given a face landmark image. Instead of conditioning on a speech segment, this work generates a face image corresponding to a given landmark image. The landmark image can be from some other person as well. Another work [28], but this time with speech, learns to generate hand movements corresponding to the speech. Their proposed system does not work for speakers in the wild, but rather for specific speakers, which is reasonable because hand gestures are, to an extent, specific to each person. Finally, another work closely related to ours is SyncNet [18], which proposes to determine the lip-sync error in talking face videos in the wild.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows.

1. In Chapter 2, we outline our first model, the LipGAN. We propose a novel method to morph lip shapes while preserving visual information like pose, background, etc.

2. In Chapter 3, we highlight key issues that hinder current existing approaches, including LipGAN, that make them inaccurate for lip-syncing videos in the wild. Consequently, we design a novel model, Wav2Lip, to accurately lip-sync talking face videos in the wild. We also investigate novel methods to evaluate lip-sync.

3. In the final chapter, we demonstrate various real-world applications enabled by our models and conclude the thesis.

Chapter 2

LipGAN: Lip-syncing while Preserving Other Visual Attributes

Our task is to lip-sync an arbitrary talking face video in the wild to match a target speech. Before we understand the current "open" issues in our task, it would help to establish a general skeletal framework for lip-syncing talking face videos. This has been used in previous works [14, 70, 61] that aim to generalize to multiple speakers and will also be the basic skeleton of our models in this thesis.

2.1 A Skeletal Framework for Lip-Syncing Talking Face Videos

We are given an input video of a person speaking, and our goal is to lip-sync it to match a target speech. To lip-sync an identity in the wild, any speaker-independent model must take the following inputs. Firstly, it needs to input the speech corresponding to which it must sync. This can be in the form of raw audio, melspectrograms, or as mel-frequency cepstral coefficients (MFCC). Secondly, it must input the identity information in some form. It could be one or more face crops extracted from the input video frames, or as just landmark points or graphical face models. Existing speaker-independent works use one or more face crops [14, 70] as the identity prior. All our models in this thesis also use just face crops for identity information. However, the model, is used differently at train and test time, and this point is essential for our further discussion and design of LipGAN. Thus, we provide a basic overview of the training and inference process of a typical speaker-independent lip-sync model in Figure 2.1 and 2.2 and also in the following sub-sections.

2.1.1 An Overview of the Training Procedure

Learning to lip-sync is a "freely" supervised task. That is, no further manual annotation is needed, and the supervision can be obtained from readily available videos on the web. Given a video, we can easily create an out-of-sync face-speech pair by sampling each of them from different time-steps. The ground-truth for training would be the face from the frame corresponding to the sampled speech segment. The network is trained to minimize the L1 reconstruction loss between the generated face and the target. An illustration of this training process is depicted in Figure 2.1.

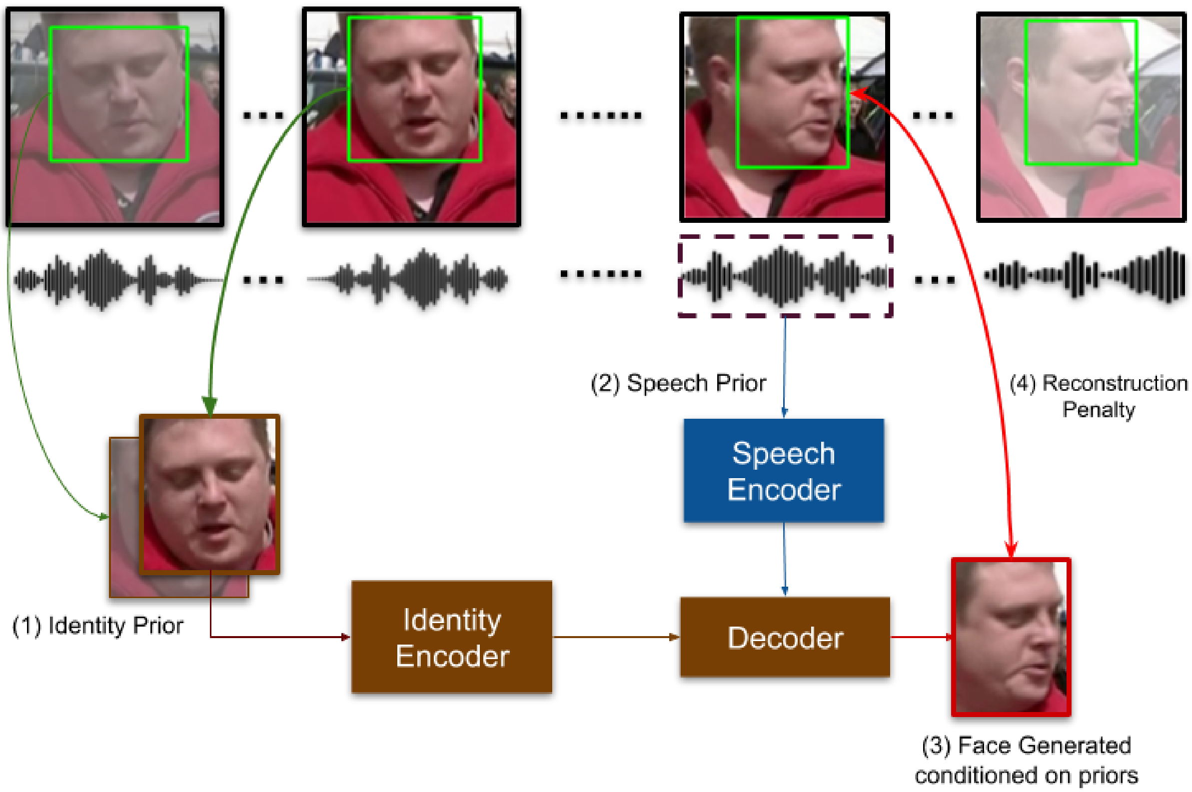

Figure 2.1 Overview of a basic training framework of a model for lip-syncing videos in the wild. An identity prior is provided in the form of a face image or landmarks or a graphical face model. The model is trained with a reconstruction penalty to generate a face image lip-synced with a speech prior that is sampled from a random time-step. Notice that the reference identity image and the desired target can have completely different poses, expressions, background, and lighting.

2.1.2 Lip-syncing videos during Test time

Lip-syncing real-world videos during test time using the above model is straightforward. The model needs to morph the lip shapes in each frame of the video. The speech segments are taken from a new target speech, and the face crops in which lips are to be morphed are fed into the lip-sync model. An essential aspect of lip-syncing such videos is that the generated face needs to be smoothly pasted back into the frame so that it appears as though only the lip and mouth movements are morphed. That is, the lip sync model must not alter the pose, expressions, background information, etc.

2.2 Challenges in Speaker-Independent Speech-driven Lip-Syncing

Now that we have an overview of the skeleton of a speaker-independent model and how it functions during training and inference, we can compare this class of models to existing speaker-specific works. Compared to the speaker-specific works [53, 44, 26, 54, 60], generating talking face videos of arbitrary

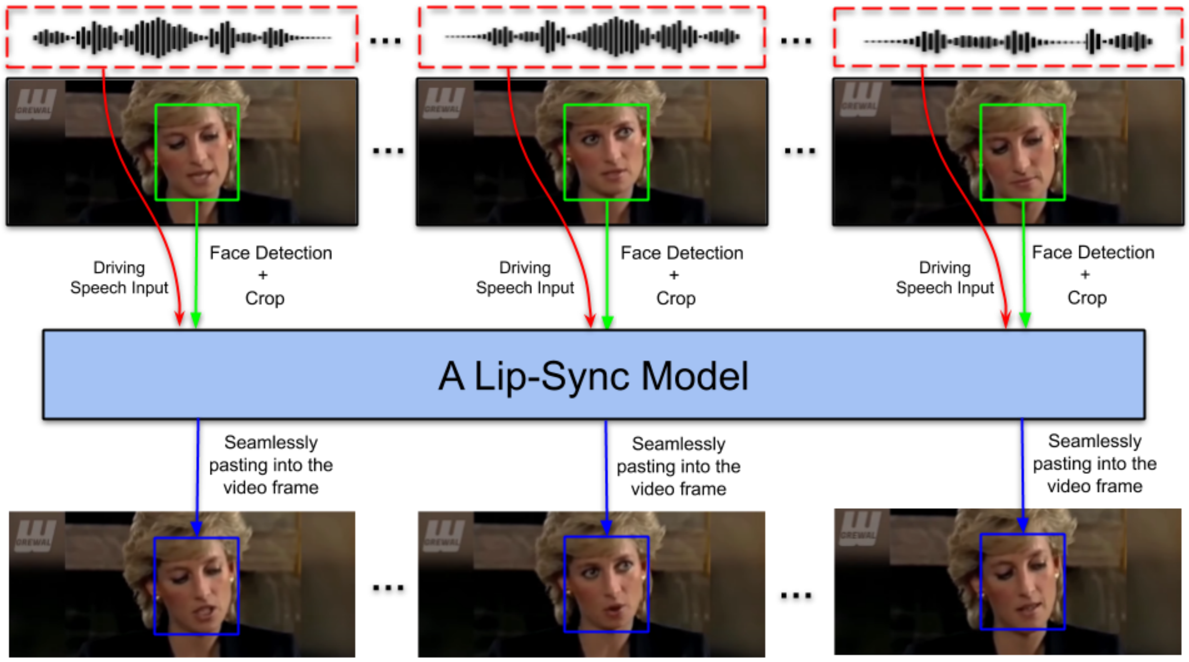

Figure 2.2 Overview of the testing framework of a model to lip-sync videos in the wild. Notice that the model is expected to preserve other visual attributes for seamless pasting into the video frame. Video frames are taken from: `https://www.youtube.com/watch?v=7jR52mrlvkg`

identities presents a different set of challenges or "open" issues. After studying these issues, we can then attempt to resolve them to achieve results as good as single-speaker models and have the massive advantage of working for arbitrary identities and speeches.

2.2.1 Inferior Visual Quality in Speaker-Independent Models

Existing works that are speaker-specific [53, 44, 26, 54, 60] can generate high-quality lip-synced videos with minimal visual artefacts. This is in sharp contrast to works that are speaker-independent [14, 70], where a significant amount of the visual attributes are lost. This is quite evident as one of the works use a de-blurring model [14] and they either show results only on static faces [70] or on videos only after rigorous post-processing [14]. This limits the applications of speaker-independent models to at most lip-syncing static face images. While this could be useful, for instance, as talking avatars in-game video streams, etc., the wide variety of other real-world applications needs good quality outputs on real-world videos that can have many poses, expressions, backgrounds, and identities. The stark difference in generation quality of current speaker-specific works and speaker-independent models is clearly illustrated in Figure 2.3.

We can see that existing speaker-independent approaches generate talking faces of inadequate quality to be used in real-world videos. For instance, the generated faces of the speaker-independent models are blurry, with a significant loss in texture and fine-grained identity attributes. The background in the

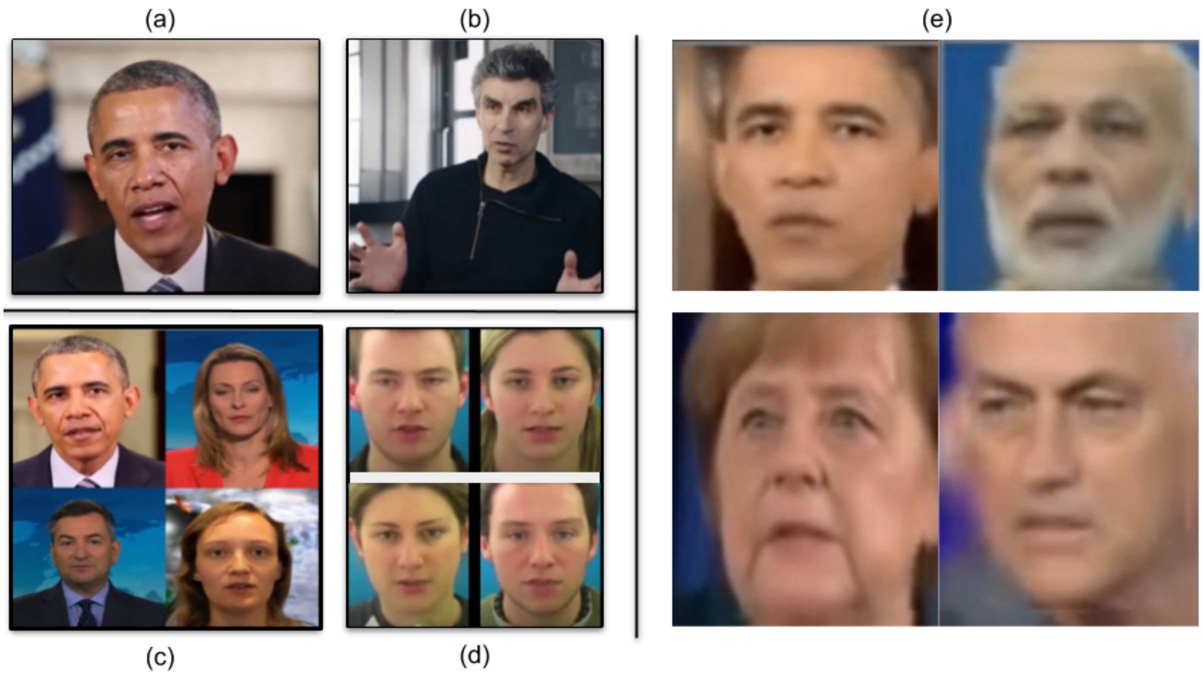

Figure 2.3 Speaker-specific models (a – d): (a) Synth. Obama [53], (b) Text-based Editing [26], (c) Neural Puppetry [54], and (d) Three-GANs [61]. Current best Speaker-Independent Approach: (e) Speech2Vid [14]. In contrast to the quality achieved by existing speaker-specific approaches, even the best speaker-independent method generates sub-par quality faces, with a significant amount of visual attributes lost. Most importantly, the speaker-independent model's generated face cannot be pasted back into the video frame, as the pose, background, expressions are not preserved.

face crop is also blurry and altered. When the face crop is pasted back into the frame, the crop will look completely artificial and out of place. Existing models are also not trained to preserve other vital visual attributes such as pose and expressions. In essence, these methods focus on generating face crop with lip-sync but do not consider how these face crops fit into the larger video frame. Consequently, we found that they arbitrarily change the poses and expressions (or lose them entirely) in the generated faces, making them unusable for lip-syncing videos.

2.2.2 Inter-person variations

Apart from being unable to match the quality of speaker-specific models, speaker-independent models face further challenges inherent to the problem. For instance, speaker-specific models have an advantage of "memorizing" the fine-grained visual attributes such as lip shapes, skin color, backgrounds, pose variations, and how a person utters phonemes, etc. For instance, the most popular candidate for speaker-specific models, President Obama, usually speaks in front of very similar backgrounds with a nearly frontal head pose and neutral expressions. The model can also learn his own speaking style, that is how he utters each phoneme. On the other hand, speaker-independent models have to learn phoneme

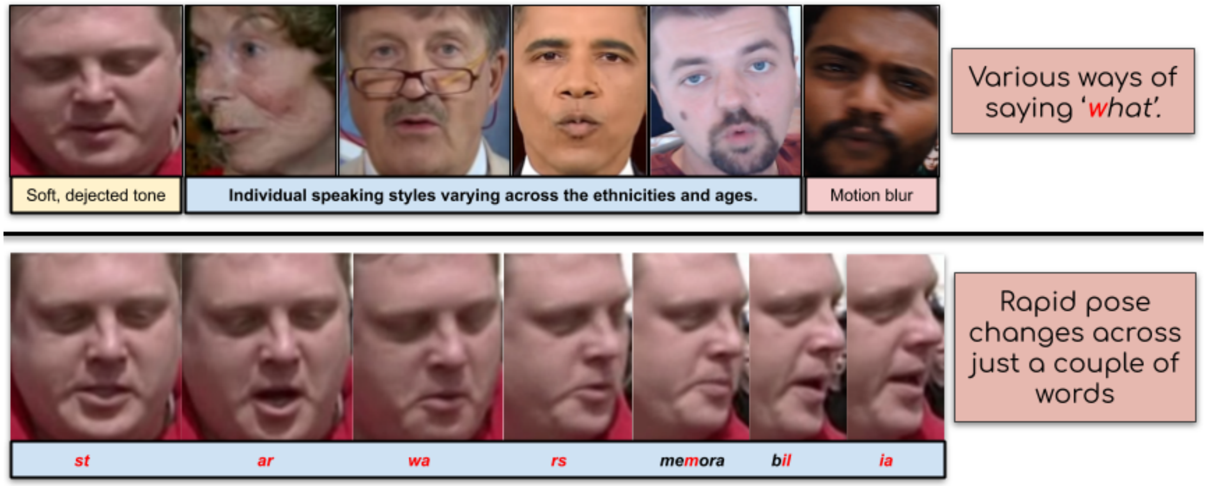

Figure 2.4 Some of the challenges faced by a speaker-independent speech-driven lip sync model. Each person has their own speaking style and personalized visemes, i.e., how they utter sounds. Notice the various ways people say the first part of the word "what". It can vary across people of different ages, races, and the underlying emotions. Another challenge of rapid changes in visual attributes such as pose and expressions within short intervals is shown in the second row. A speaker-independent lip-sync model is expected to retain these changes across frames while only morphing the lip shapes.

to viseme mappings that generalize across identities, expressions, poses, and backgrounds. Figure 2.4 depicts some of these variations. In the first row, notice how uttering the starting of the word "what" is very different across identities. The utterance can be influenced by emotion and rapid motion. Some of them can have very distinctive lip shape for the utterance (Obama, for instance), and some of them can be not so distinctive (left of Obama). In the second row, we can see how a person in a real-world video can change his (her) head pose rapidly across even a very short duration, going from a full-frontal view to a profile-view, which also changes the scale of the face crop. The model has to handle and preserve all these differences while only altering the lip shape.

2.2.3 Inadequate lip-sync penalization

We identify two major reasons for inferior lip-sync in existing models.

2.2.3.1 Lip sync is a fine-grained attribute

Given these large amounts of variations, we can see why the reconstruction loss in the current training framework is inadequate for penalizing fine-grained lip-sync errors. Consider the variation in the second row of Figure 2.4. The current training framework outlined in Figure 2.1, can sample a random identity frame from a profile view as shown in the second row of Figure 2.4 and expect the network to generate a target face in the front pose. Altering this pose is a challenging task and would thus constitute a large amount of the reconstruction loss as it is spread across the whole image. On the other hand, the lip region is less than 4% of the pixels and only requires a smaller fine-grained change.

2.2.3.2 Inadequate visual prior

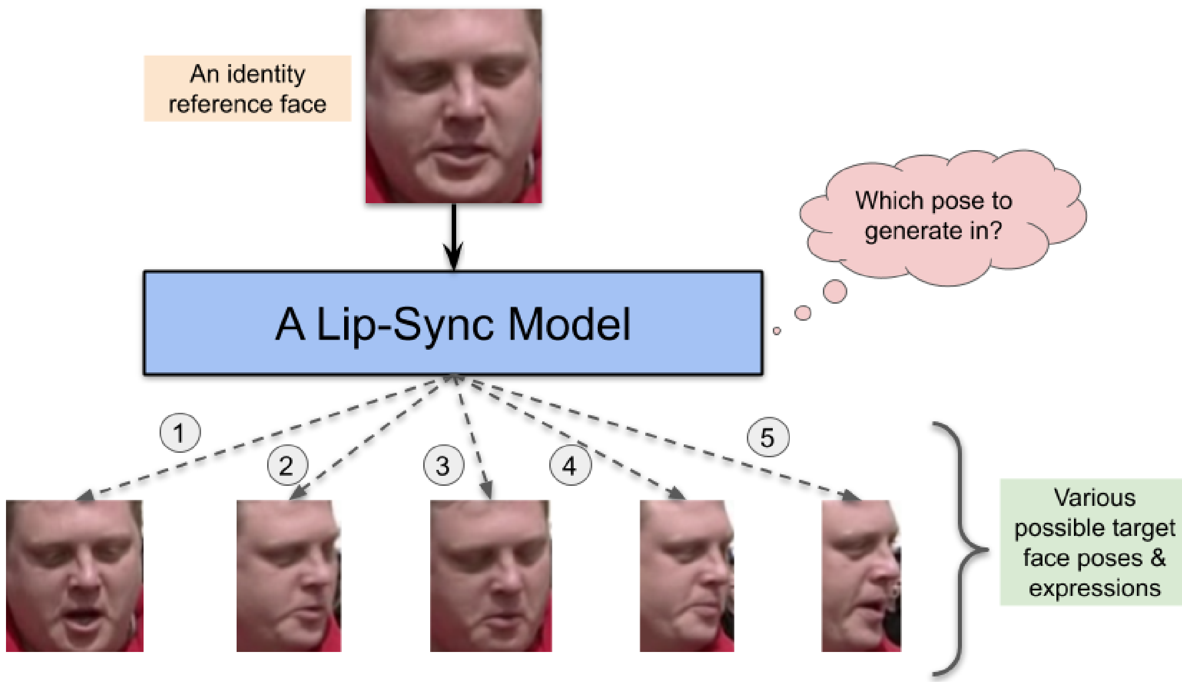

Figure 2.5 We find that the existing speaker-independent lip-sync models suffer from the problem of inadequate visual priors. The model is expected to morph various visual attributes such as pose, expressions of an arbitrarily sampled identity face to match a very different ground-truth face image. As the only signal this model can extract from the audio is the lip movements, it does not know which ground-truth pose and expression it should generate.

As already described in the training framework (Figure 2.1), a random identity face is sampled as a reference from an arbitrary time-step, and the frame aligned with the speech is chosen as the ground-truth face. As discussed before, the goal of our problem is only to morph the lip movements. This is partially aligned with the current training framework, where the only clear signal that can be extracted from the short audio segment is regarding the lip movement. Thus the model is well set to morph just the lip movements. However, the sampled identity and ground truth pairs vary not just in lip shape but also in terms of several other visual attributes like pose, expressions, background, etc. The model has no prior information on how to change these attributes in the identity image to match the ground-truth image. The changes that need to be made are quite large, as shown in Figure 2.5, where we can see that if the ground-truth image is, for example, the fifth image, a large change is necessary to minimize the reconstruction loss.

2.3 Designing the LipGAN Architecture

Now that we have studied the various issues highlighted in the previous section, we can now design the LipGAN model (Figure 2.6) that can generate high quality lip-synced talking face videos in the wild. To do this, we make several design choices. We outline each of them below and how they handle one or more of the issues mentioned above. Note that we roughly follow the training framework outlined in Figure 2.1, so we have a face identity encoder, a speech encoder, and a face decoder. Along with this, we also have an additional discriminator, which we will get into later.

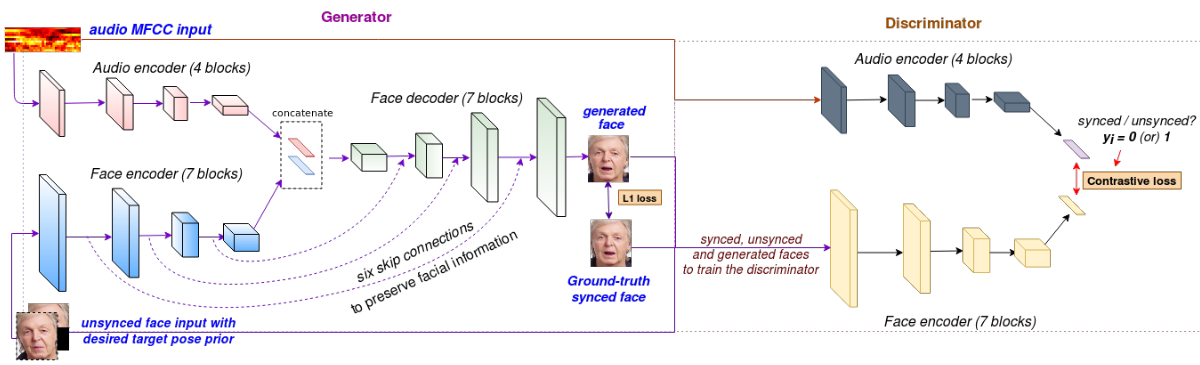

Figure 2.6 We follow the training framework already outlined before. Additionally, we make two key improvements. We use a pose prior to condition the model to preserve the target pose and expressions. We also use extensive skip connections throughout the generator model. We train our LipGAN network in an intuitive GAN setup. The generator generates face images conditioned on the audio input. The discriminator checks whether the generated frame and the input audio are in sync. Note that while training the discriminator, we also feed extra ground-truth synced / unsynced samples to ensure that the discriminator learns to specifically check for superior lip-sync and not just the image quality.

2.3.1 Extensive Skip Connections

We first address the problem of inferior visual quality in the current speaker-independent models. Existing models such as Speech2Vid [14] struggle to preserve the fine-grained visual details in the generated faces. They resort to using an additional de-blurring network for this task. Instead, we take inspiration from image segmentation works [51, 22] and use extensive skip connections between our face encoder and decoder. While these are “long” skip connections [22], we also employ short residual identity skip connections [31] to improve learning. We find that these short and long skip connections preserve the visual features significantly better.

2.3.2 Target Pose Prior

The next issue that we would like to handle is linked to inadequate visual priors. In order to deal with the large variations between the target and reference images, existing speaker-independent works [14, 70] propose to perform face registration and alignment. That is, they alter the pose of all the face frames to match a standard frontal face. This process is shown in Figure 2.7. This is definitely helpful in

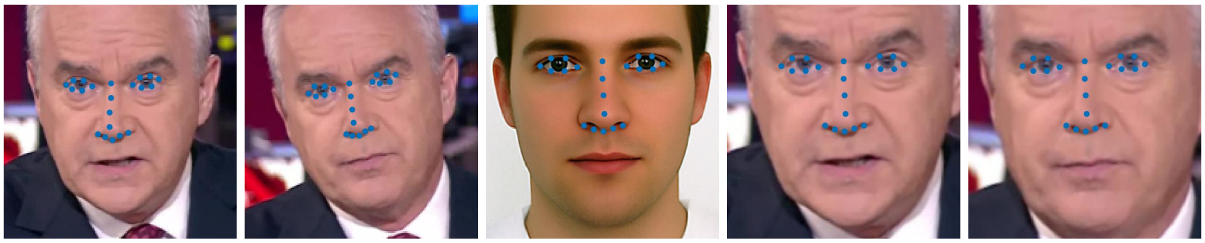

Figure 2.7 Illustration from [14] showing that facial registration is performed to handle large pose variations between sampled identity and target face pairs. While this eases the learning of audio-visual correspondence, it introduces new issues such as being unable to paste the generated face back into the original video frame. Notice how the pose and the visual content is drastically changed with respect to the original crop.

learning to lip-sync from speech, as the lips are roughly in the same location, and more importantly, it aligns the lip positions of the ground truth and the identity frame. However, this introduces several limitations. To start with, facial alignment is computationally expensive; in fact, more expensive than the generation itself. But more importantly, it destroys the visual information present in the original crop. In Figure 2.7, we can see that the background, face scale, and pose are entirely altered as a result of facial registration. But, without registration, the pose, expression, etc. change drastically across the input identity and the target face, making it extremely infeasible for the model to learn. *Thus, the question is, how can we better assist the model in preserving all the other visual attributes as much as possible while only morphing the mouth regions?*

As we can see in Figure 2.5 and have made earlier, a key observation is that the current model does not have adequate priors on which target face to generate. Thus, if we can provide as input a piece of prior information about the ground-truth frame we expect from the model, then the model can learn to generate faces that closely match the desired target face. However, one of the essential aspects is that we cannot leak the target mouth region to the model, as that is something the model must learn from the input speech.

With these points in mind, we propose a novel "in-painting based" approach. We provide the top half of the target face, which contains pose, expressions, etc., as a prior. We would like to in-paint the bottom half of the face, mainly consisting of the lip, mouth, and jaw region, all of which are strongly correlated with the speech. The model can learn to generate the bottom half that matches the top half. This is not ideal, as the model can still face difficulties in handling substantial pose changes. But, it is significantly better than having no prior, as at least, in this case, the model knows which target face to generate.

2.3.2.1 Pose prior during training

There is an essential (and beneficial) difference in how we use our pose prior during training and inference times. During training, we follow the training framework already illustrated before. We sample a target face and the corresponding speech, and also a random identity image. Instead of feeding just the random identity image, we blacken the lower half of the ground-truth image to mask the mouth and concatenate it channel-wise with the identity image. Thus our identity input is $(height, width, 6)$.

2.3.2.2 Pose prior during testing

A beneficial aspect of our approach is that during testing, the model does not need to alter the pose at all. The same face crop is given as the identity image and the pose prior. That is because, in the end, we do not want our model to alter any other visual attributes other than the mouth region. During training, our model has been trained to generate in a specific pose specified by the pose prior. At the time of inference, it has to do the most straightforward task of not changing anything at all. This feature leads to our model being sub-par during training but produces high-quality outputs with minimal to no artifacts during test time.

2.3.3 Penalizing incorrect lip generation

As discussed in the issues section, plain reconstruction loss is not a "specific" penalization of incorrect lip shapes. Given that the lip shape is a small region in the entire face (less than 4%), an additional loss is needed to penalize inaccurate lip shape generation. Multiple methods come to mind. We explore a few of them below.

2.3.3.1 Optimizing for the SSIM metric

Structural Similarity Index (SSIM) [63] has been widely used to measure and evaluate how well the models generate image structures and textures. Compared to other image quality assessment metrics such as PSNR, SSIM co-relates better with attributes such as lip shape. To improve lip generation, we can optimize the SSIM metric on just the lip area or the face image. However, as this metric was not developed for lip-sync evaluation, this method did not yield significant improvements in lip generation.

2.3.3.2 Optimizing for the lip landmark distance

Compared to SSIM, we could also use some features specific to lips such as lip landmarks. A recent work [10] showed that we can perform accurate landmark detection on faces in-the-wild. However, using lip landmarks for discrimination is still not advisable as it poses two problems. Firstly, lip landmarks are far less accurate than the remaining face landmarks. Secondly, detecting lip landmarks on noisy generated faces would be extremely inaccurate. Thus, it would not be an accurate penalization method.

2.3.3.3 Using a lip-reading discriminator

This is a more practical method proposed in [70]. They train on the LRW dataset, and they use a discriminator trained to perform visual speech recognition a.k.a. lip-reading. Since a lip-reading model focuses primarily on lip regions, it can be expected to penalize inaccurate lip-sync. The model expects that the lips in the generated face images must lead to correct word-level lip-reading. While this is an appealing approach, it presents significant limitations. Firstly, this training process is not "freely-supervised" like our problem. It requires accurate textual transcripts. Moreover, these texts must be well-aligned with the speech and video, at least at the word-level. The LRW dataset [17] consists of 1000 words in English with such annotations. Straightaway, we see the limitation in the variety and size of the dataset. The model becomes extremely biased towards these words and cannot generalize to unseen speeches and languages. Thus, this is an infeasible approach.

2.3.3.4 Using a real (or) fake discriminator

Generative Adversarial Networks (GANs) have shown tremendous improvements in multiple image generation tasks ranging from super-resolution [62, 65, 37], style-transfer [40, 37, 45], photo-realistic face generation [39, 33], pose generation [46], face editing [50, 9], face aging [6] and de-aging [67]. Thus, it is only reasonable to employ them to improve the visual quality of the generated faces. It can also penalize unrealistic lip shapes. However, by just training a real or fake image discriminator, it cannot penalize inaccurate lip shapes that do not match with the input audio. This method can only penalize unrealistic lip shapes and not inaccurate lip shapes.

2.3.3.5 Learning to synthesize by testing for synchronization

We formulate our lip synthesis problem as "learning to synthesize by testing for synchronization". Concretely, our setup contains two networks, a generator G that generates faces by conditioning on audio inputs and a pose prior and a discriminator D that tests whether the generated face and the input audio are in sync. By training these networks together in an adversarial fashion [29], the generator G learns to create photo-realistic faces that are accurately in sync with the given input audio. The setup is illustrated in Figure 2.6. Since the discriminator also has access to the input audio, it can specifically check for inaccurate lip generation rather than just visual quality. We detail in the training section of the discriminator on how we force the discriminator to specifically check for lip-sync rather than just the visual quality.

2.4 Details of the LipGAN architecture

Now that we have established the design principles of the LipGAN architecture, we now delve into each module in detail. Note that our model contains two major blocks: the generator and the discriminator. We start with the generator block.

2.4.1 Generator network

The generator network is a modification of [14] and contains three branches: (i) Face encoder, (ii) Audio encoder and a (iii) Face Decoder. This is similar to the training framework established before but with additional design decisions made above.

2.4.1.1 The Face Encoder.

We design our face encoder a bit differently from [14]. The first difference is at the input level itself, i.e., an additional pose prior to handle the pose mismatch between the generated and target face. Along with the random identity face image I, we also provide the desired pose information of the ground-truth as input to the face encoder. We mask the lower half of the ground truth face image and concatenate it channel-wise with I. The masked ground truth image provides the network with information about the target pose while ensuring that the network never gets any information about the ground truth lip shape. Thus our final input to the face encoder is a HxHx6 image. The second difference is that we use residual layers instead of vanilla CNN layers with gradual intermediate down-sampling layers. The exact details of each layer are shown below in Table 2.4.1.1. The encoder is responsible for altering the pose of the identity image to match the target pose. It also embeds the identity features at different spatial resolutions to be used by the decoder. The face encoder embeds the given input face image into a face embedding of size h.

Attribute	block1	block2	block3	block4	block5	block6	block7
No. of Conv. Layers	1C	1C; 2R	1C; 3R	1C; 2R	1C; 2R	1C	2C
Kernel Size	7	5; 3; 3	3	3	3	3	3; 1
Stride	1	2; 1; 1	2; 1; 1	2; 1; 1	2; 1; 1	2	3; 1
No. of filters	32	64	128	256	512	512	512
Input Resolution	96	96	48	24	12	6	3
Output Resolution	96	48	24	12	6	3	1

Table 2.1 Full details of our modular face encoder. "1C" means one vanilla convolution block. Each vanilla convolutional block consists of a convolution layer followed by batch-normalization [34] and a ReLU [49] activation."2R" means two residual convolution blocks. Each residual block consists of two vanilla convolution blocks with an identity skip connection across them. Note that we use strided convolution for gradual down-sampling to preserve the spatial information as much as possible. All resolutions have the same height and width, indicated by a single number.

2.4.1.2 Audio Encoder.

The audio encoder is also adapted from [14] and consists of a series of residual convolution layers. It takes a Mel-frequency cepstral coefficient (MFCC) heatmap of size MxTx1 and creates an audio embedding of size h. The audio embedding is concatenated with the face embedding to produce a joint audio-visual embedding of size 2xh. The complete layer-wise details are provided in Table 2.4.1.2

Attribute	block1	block2	block3	block4	block5
No. of Conv. Layers	1C; 3R	1C; 2R	1C; 2R	1C; 1R	2C
Kernel Size	3	3	3	3	4×5; 1
Stride	1	1×2; 1; 1	2; 1; 1	2; 1	4×5; 1
No. of filters	128	256	512	512	512
Input Shape	12×35	13×35 (1ZP)	13×18	7×9	4×5
Output Shape	12×35	13×18	7×9	4×5	1×1

Table 2.2 Full details of our modular audio encoder. "ZP" stands for zero-padding. Other notations are same as the Table 2.4.1.1.

2.4.1.3 Face Decoder.

This branch produces a lip-synchronized face from the joint audio-visual embedding by inpainting the masked region of the input image with an appropriate mouth shape. It contains a series of residual blocks with a few intermediate deconvolutional layers that upsample the feature maps. The output layer of the face decoder is a sigmoid activated 1x1 convolutional layer with 3 filters, resulting in a face image of HxHx3. While [14] employ only 2 skip connections between the face encoder and the decoder, we employ 6 skip connections, one after every upsampling operation to ensure that the fine-grained input facial features are preserved by the decoder while generating the face. As we feed the desired pose as input during training, the model generates a morphed mouth shape that matches the given pose. Indeed, in our results, it can be seen that we preserve the face pose and expression much better than [14] and only alter the mouth regions to match the input speech. This allows us to seamlessly paste the generated face crop into the given video without any artifacts, which was not possible with [14] due to the random, uncontrollable pose variations. The complete details of the face decoder branch are given in Table 2.4.1.3.

2.4.2 Discriminator network

While using only an L2 reconstruction loss for the generator can generate satisfactory talking faces [14], employing strong additional supervision can help the generator learn robust, accurate phoneme-viseme mappings and also make the facial movements more natural. [70] employed audio-visual speech recognition as a probe task to associate the acoustic and visual information. However, this makes the setup

Attribute	block1	block2	block3	block4	block5	block6	Output
No. of Conv. Layers	1T	1T; 2R	1T; 2R	1T; 2R	1T; 2R	1T; 2R	1S
No. of filters	512	512	256	128	64	32	3
Input Resolution	1	3	6	12	24	48	96
Output Resolution	3	6	12	24	48	96	96
Encoder Connection	block6,7	block5	block4	block3	block2	block1	-

Table 2.3 Full details of our modular face decoder. "1T" means one transpose convolution block. It is the same as the vanilla convolutional block detailed in the previous tables but uses a transpose convolution instead. The final layer "1S" is a sigmoid activated convolutional layer. Other notations can be inferred from the previous tables. All kernel sizes are 3×3, and all transposed convolutions are performed with a stride of 2, except the first one, which is done with a stride of 3. The feature maps from the encoder blocks (last row) are concatenated channel-wise after each upsampling operation. Note that our complete generator is fully convolutional.

language-specific and offers only indirect supervision. We argue that directly testing whether the generated face synchronizes with the audio provides a stronger supervisory signal to the generator network. Accordingly, we create a network that encodes an input face and audio into fixed representations and computes the normalized L2 distance d between them. The face encoder and audio encoder are similar to those used in the generator network, but with much lesser parameters. We use a series of vanilla convolutional layers with rapid down-sampling. We found that using Instance Normalization [58] and LeakyReLU [47] in the discriminator was more stable than using Batch normalization [34] with ReLU activations. Complete architectural details are given in Table 2.4.2.

Attribute	conv1	conv2	conv3	conv4	conv5	conv6
Kernel Size	7	5	3	3	3	3
No. of filters	64	128	256	512	512	512
Input Resolution	96	48	24	12	6	3
Output Resolution	48	24	12	6	3	1

Table 2.4 Full details of our face encoder in the discriminator. Each convolutional block consists of a convolution layer followed by instance-normalization and a LeakyReLU [49] activation. All layers have a stride of two except the last layer that has a stride of three.

2.4.3 Joint training of the GAN framework

Our training process is as follows. We randomly select a T millisecond window from an input video sample and extract its corresponding audio segment A, re-sampled at a frequency F Hz. We choose the middle video frame S in this window as the desired ground-truth. We mask the mouth region (assumed to be the lower-half of the image) of a person in the ground truth frame to get S_m. We also sample

Attribute	block1	block2	block3	block4
No. of Conv. Layers	2C	2C	2C	1C
Kernel Size	3	3	3	4×5
Stride	1	1×2; 1	2	4×5;
No. of filters	64; 128	256	256; 512	512
Input Shape	12×35	13×35 (1ZP)	13×18	4×5
Output Shape	12×35	13×18	4×5	1×1

Table 2.5 Full details of our discriminator's audio encoder. "ZP" stands for zero-padding. Other notations are same as the Table 2.4.1.2.

a negative frame S', i.e., a frame outside this window which is expected to not be in sync with the chosen audio segment A. At each training batch to the generator, the unsynced face S' concatenated channel wise with the masked ground truth face S_m and the target audio segment A is provided as the input. The generator is expected to generate the synced face $G([S'; S_m], A) \approx S$. Each training batch to the discriminator contains three types of samples: (i) Synthetic samples from the generator $(G(S', A), A); y_i = 1, (ii)$ Actual frames synced with audio $(S, A); y_i = 0$ and (iii) Actual frames out of sync with audio $(S', A); y_i = 1$. The third sample type is particularly important to force the discriminator to take into account the lip synchronization factor while classifying a given input pair as real / synthetic. Without the third type of sample, the discriminator would simply be able to ignore the audio input and make its decision solely on the quality of the image. The discriminator learns to detect synchronization by minimizing the following contrastive loss:

$$L_c(d_i, y_i) = \frac{1}{2N} \sum_{i=1}^{N} (y_i \cdot {d_i}^2 + (1 - y_i) \cdot max(0, m - d_i)^2) \tag{2.1}$$

where m is the margin, which we set to 2. The generator learns to reconstruct the face image by minimizing the L1 reconstruction loss:

$$L_{Re}(G) = \frac{1}{N} \sum_{i=1}^{N} ||S - G(S', A)||_1 \tag{2.2}$$

We train the generator G and discriminator D using the following GAN objective function:

$$L_{\text{real}} = \mathbb{E}_{z,A}[L_c(D(z, A), y)] \tag{2.3}$$

$$L_{\text{fake}} = \mathbb{E}_{S',A}[L_c(D(G([S'; S_m], A), A), y = 1)] \tag{2.4}$$

$$L_a(G, D) = L_{\text{real}} + L_{\text{fake}} \tag{2.5}$$

where $z \in \{S, S'\}$. Here, G tries to minimize L_a and L_{Re} and D tries to maximize L_a. Thus, the final objective function is:

$$G^* = \arg \min_G \max_D L_a(G, D) + L_{Re} \tag{2.6}$$

2.4.4 Dataset and Implementation Details

We use the train split of the LRS 2 dataset [2], which contains over 29 hours of talking face videos. We train on a single NVIDIA TITAN X GPU with a batch size of 64. We detect faces in our input frame using `dlib` [41] and resize the face crops to 96x96x3. We use the Adam [42] optimizer with an initial learning rate of $1e^{-3}$ and train for about 20 epochs. The learning rate was reduced by five when the loss plateaued. We have created two implementations, one using MATLAB audio features, which is identical to [14]. Since the rest of the pipeline is in Python, we also create a fully-pythonic version using audio features using the librosa library[1]. We detail both these implementations further below. Both these implementations give very similar results.

2.4.4.1 MATLAB MFCC features for audio

We extract 13 MFCC features from each audio segment $(T = 350, F = 100)$ and discard the first feature similar to [14]. This yields an MFCC feature of shape 12×35. This is readily compatible with our model detailed above.

2.4.4.2 Python Librosa Melspectrogram features for audio

We also devise a Python implementation as it is cumbersome for applications to handle two very different programming languages. We use the librosa library to extract the audio features. First, we sample the audio at $F = 16000$ Hz. We extract linear spectrograms by performing a short-time-fourier transform on the raw audio. We use a window size of 50ms (800) and a hop size of 12.5ms (200). We further compute mel-spectrograms with 80 mel features. Thus for each second of audio, we get a mel-spectrogram window of 80×80. The mel-spectrograms are normalized between $[-4, 4]$. Similar to how we take an audio window with a frame in the center, here we take a mel window of 27 time-steps (corresponding to approximately 350ms). We modify our audio network to match the new input shape, but we keep the remaining portions intact.

Both the MATLAB and Python code are released publicly at: github.com/Rudrabha/LipGAN.

2.5 Results and Evaluation

We evaluate our novel LipGAN architecture quantitatively and also with subjective human evaluation. All our evaluation and ablation studies are performed using the MATLAB version of the code, as this allows for a fair comparison with previous works that use the same code-base for speech processing.

During inference, the model generates the talking face video of the target speaker frame-by-frame. In each frame, we detect the face crop, resize it to 96×96 and use it as an input to our model. This visual input is the current face crop concatenated with the same cropped face with the lower-half masked.

[1] https://librosa.org/librosa/

That is, during inference, we expect the model to morph the input shape and preserve other aspects like pose and expression. Along with each of the visual inputs, we feed a $T = 350$ms audio segment. The generated faces are resized back to their original sizes and can be just placed into the video frame without any further post-processing. In Figure 2.8, we compare the talking faces on static face images generated by the 3 models: Speech2Vid [14], TalkingFace [70] and ours on audio segments actually spoken by PM Narendra Modi and Elon Musk.

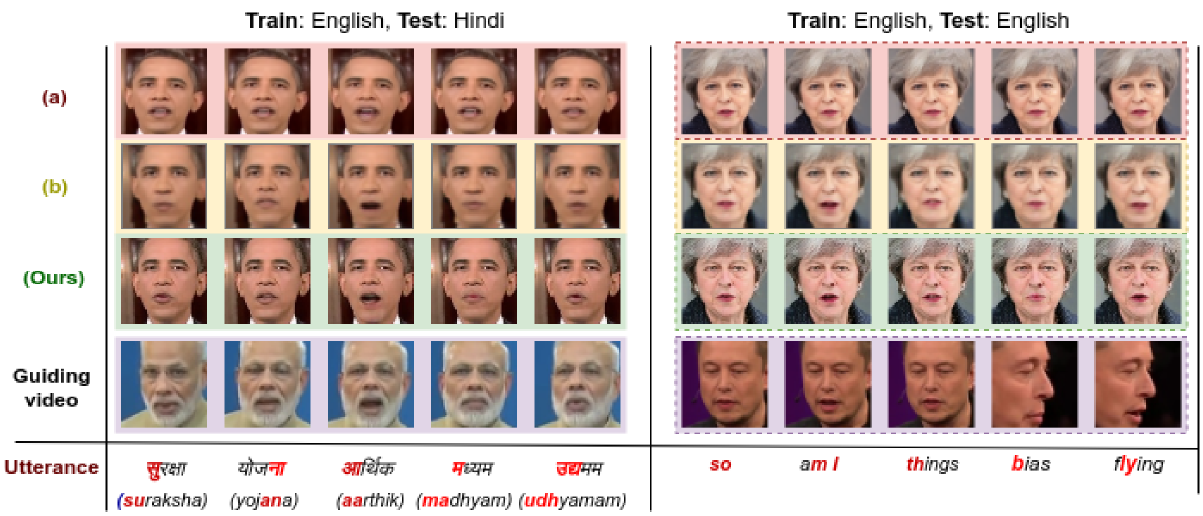

Figure 2.8 Visual comparison of faces generated by different models when they try to speak specific segments of the words shown in the last row. The audio segments corresponding to these word segments are extracted from the guiding video and fed into each of the models compared above. From top to bottom row: (a)TalkingFace [70] (b)Speech2Vid [14] and Our LipGAN model. While (a) achieves poor lip-sync across languages, and (b) generates unnatural lip movements, our LipGAN model produces consistent accurate, natural talking faces across languages.

2.5.1 Quantitative evaluation

To evaluate our lip synthesis quantitatively, we use the LRW test set [17]. We follow the same inference method mentioned above, but with one change. Instead of feeding the current frame as input as mentioned above, we feed a random input frame of the speaker, concatenated with the masked current frame for the pose prior. This is done to ensure we do not leak any lip information to the model while computing the quantitative metrics. In Table 2.6, we report the scores obtained using standard metrics: PSNR, SSIM [63] and Landmark distance. As shown in Table 2.6, our model significantly outperforms existing works across all quantitative metrics. These results highlight the superior quality of our generated faces (judged by PSNR) and also a highly accurate lip synthesis (LMD, SSIM). The noted increase in SSIM and the decrease in LMD can be attributed to the direct lip-synchronization supervision provided by the discriminator, which is absent in prior works.

Algorithm	PSNR	SSIM	LMD
Speech2Vid [14]	28.06	0.460	2.22
TalkingFace [70]	26.80	0.884	-
LipGAN (Ours)	**33.4**	**0.960**	**0.60**

Table 2.6 Our proposed LipGAN model achieves significant improvements over existing competitive approaches across all standard quantitative metrics.

2.5.2 Importance of the lip sync discriminator.

To illustrate the effect of employing a discriminator in the LipGAN network that tests whether the generator faces are in sync, we conduct the following experiment. We train the talking face generator network separately on the same train split of LRS2 without changing any other hyper-parameters. We feed the unseen test images shown in Figure 2.9 along with unseen audio segments as input to our LipGAN network and the plain generator network that was trained without the discriminator. We plot the activations of the penultimate layer of the generator in both these cases. From the heatmaps in Figure 2.9, it is evident that our LipGAN network learns to attend actively on the lip and mouth regions compared to the one that is not trained with a lip-sync discriminator. These findings also concur with the significant increase in the quantitative metrics and the natural movement of the lip regions in the generated faces.

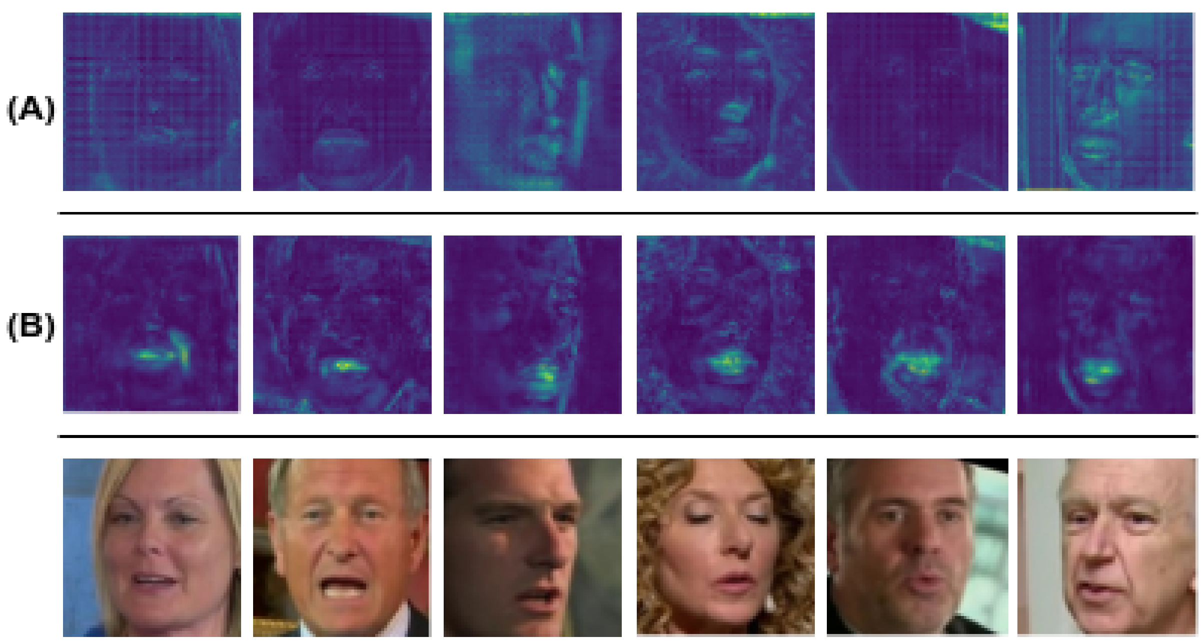

Figure 2.9 Activation heatmaps from the penultimate layer of two generator networks, one trained without a lip-sync discriminator (A) and the LipGAN network (ours) with a discriminator (B). Our network with the discriminator is highly attentive towards lip and mouth regions.

2.5.3 Human evaluation.

Talking face generation is primarily done for direct human consumption. Hence, alongside the quantitative metrics, we also subject it to human evaluation. We choose 10 audio samples, with an equal number of English and Hindi speech videos. We generate talking faces for each audio sample using three different models for 5 popular identities to yield a total of 150 samples. We compare the faces generated by three different models: (i)Speech2Vid [14], (ii)TalkingFace [70] and (iii) Our LipGAN model. We conduct a user study with the help of 20 participants who are asked to rate each of the videos on a scale of 1 to 5 based on the extent of lip synchronization and realistic nature. As shown in Table 2.7, our model obtains significantly higher scores compared to existing works.

Approach	Lip-sync rate	Realistic rate
TalkingFace [70]	2.41	2.42
Speech2Vid [14]	2.95	3.10
Ours	**3.68**	**3.73**

Table 2.7 LipGAN achieves significantly higher scores for both realistic rate and the extent of lip synchronization

2.6 Various Real-World Examples

In this section, we show several real-world examples and some hand-picked real-world applications made possible by our LipGAN model. Each of them depicts a different aspect of our model. We also attach a time-stamp of the publicly available video for each example shown below.

2.6.1 Animating CGI characters with speech

Animating computer-generated faces requires a skilled person to carefully observe the speech audio and manually overlay visemes on the animated faces. Such practices can be expensive and time-consuming. Also, as it is tedious to correct in each video frame and to do it very accurately, several lip movements, even in professional movies, can look unnatural for CGI characters. Even though our LipGAN model was trained only on human faces, we can generate lip movements for animated characters given a speech segment. We show sample frames in Figure 2.10, with the corresponding red letters highlighting the phoneme being uttered in that frame.

2.6.2 Multi-language support

Ours is the first work to showcase multi-language support. Since there is a significant phoneme overlap across languages, training on a large amount of speech from only a single language like English

Figure 2.10 Here, we make the Tintin animated character sing a Beatles song. Watch the video clip here: https://youtu.be/aHG6Oei8jF0?t=17. We can see that our model can preserve the facial features to allow for seamless pasting of the generated face crop onto the video frame.

allows us to generalize to several other languages. This is extremely beneficial for applications like correcting lip-sync in dubbed movies, without having to train a separate lip-sync model on each language. Some generations from three languages are shown in Figure 2.11.

2.6.3 Lip-syncing dubbed movie scenes

Movies are generally dubbed by dubbing artists manually. The dubbed audio is then overlaid to the original video. This causes the actors' lips to be out of sync with the audio, affecting the viewer experience. We demonstrate that we can synthesize and synchronize lips in manually dubbed videos, thus automatically correcting any dubbing errors. LipGAN is the first model to show results in such unconstrained real-world videos. Lip-sync corrected frames from the Hindi-dubbed Harry Potter movie, which was originally in English, is shown in Figure 2.12.

2.6.4 Lip-syncing for TTS generated speech

Even though our model was trained only on human speech, we can generate lip movements to match the synthetic voices generated by a TTS model. This is extremely useful for applications like automatic dubbing and video content creation. We are the first to show results on TTS generated samples as well. Generated samples for various languages and voices are shown in Figure 2.13. Note that due to the domain shift, we observe that the lip-sync accuracy drops with TTS speech.

Figure 2.11 Our LipGAN model can lip-sync speech in different languages even though it was trained only on English. Watch the video clip here: https://youtu.be/aHG6Oei8jF0?t=73. Note how the visual quality is competitive in several cases to that of the speaker-specific works.

Figure 2.13 Our LipGAN model can lip-sync faces to match a synthetic speech segment from a TTS model. Watch the video clip here: https://youtu.be/aHG6Oei8jF0?t=250. All the words are uttered by a TTS model in different voices.

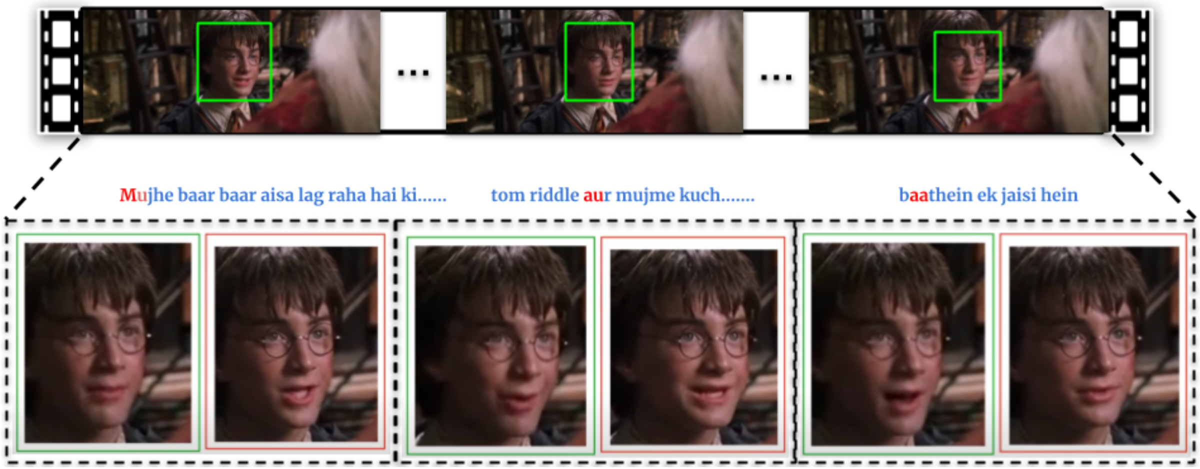

Figure 2.12 Our LipGAN model can lip-sync dubbed movie scenes to match the dubbed audio. Watch the video clip here: https://youtu.be/aHG6Oei8jF0?t=196. Notice how the face box is seamlessly pasted into the video frame. Also, our model is able to correct lip-sync in movies scenes even in the presence of background music.

2.6.5 Further Applications

Our method also enables automatic lip-sync of translated educational content and television news interviews. In the future, this can be an essential module in conjunction with text-to-speech and machine translation models. Such applications can be widely impactful in a country like India, where hundreds of languages are spoken across a large population. Thus lip-sync can have future applications to make video content accessible across people who speak different languages.

2.7 Discussion and Summary

In this work, we proposed a novel model that can lip-sync videos in the wild with far superior visual quality and accuracy than previous works. Ours is the first work that can lip-sync any identity in arbitrary poses without any facial registration. We propose a novel in-painting approach to achieve this. This allows us to blend the generated faces back onto the video frame seamlessly. To improve the accuracy of the generated lips, we train the generator with a discriminator that checks specifically for lip synchronization. These improvements allow us to demonstrate speech-driven lip-sync applications in several real-world scenarios. All resources pertaining to this chapter, such as demo video containing the results, code, and models can be found on the project page: `cvit.iiit.ac.in/research/projects/cvit-projects/facetoface-translation`

We believe that LipGAN shifts the focus for future directions from improving the visual quality to improving the lip-sync accuracy in unconstrained real-world videos. Being able to accurately lip-sync unconstrained videos in the wild still remains a challenge in this field.

Chapter 3

Accurately Lip-syncing Videos In The Wild

3.1 Introduction

In the previous chapter, we witnessed the numerous applications made possible by our lip-sync model. We observed that almost all of these essential applications need a lip-sync model to accurately morph the lip movements on real-world videos and not just on static face images. In this chapter, we re-visit the progress in speaker-independent lip-syncing over the recent years and how far we are from achieving highly accurate, reliable lip-syncing on real-world videos.

3.1.1 Landscape of Talking Face Videos

The current space of generating lip-synced talking face videos is depicted in Figure 3.1. Initial works [44, 53] using deep learning in this space learned a mapping from speech representations to lip landmarks using several hours of a single speaker. More recent works [26, 54] in this line directly generate images from speech representations and show exceptional generation quality for specific speakers which they have been trained upon. Numerous practical applications, however, require models that can readily work for generic identities and speech inputs. This led to the creation of speaker-independent speech to lip generation models [14] that are trained on thousands of identities and voices. In the previous chapter, we proposed LipGAN, which significantly improved upon this class of works in terms of the visual quality and natural lip movements. Both Speech2Vid [14] and LipGAN can generate accurate lip motion on a single, static image of any identity in any voice, including that of a synthetic speech generated by a text-to-speech system. However, to be used to generate natural video content, like translating a lecture/TV series, for example, these models need to be able to morph the broad diversity of lip shapes present in these dynamic, unconstrained videos as well, and not just on static images.

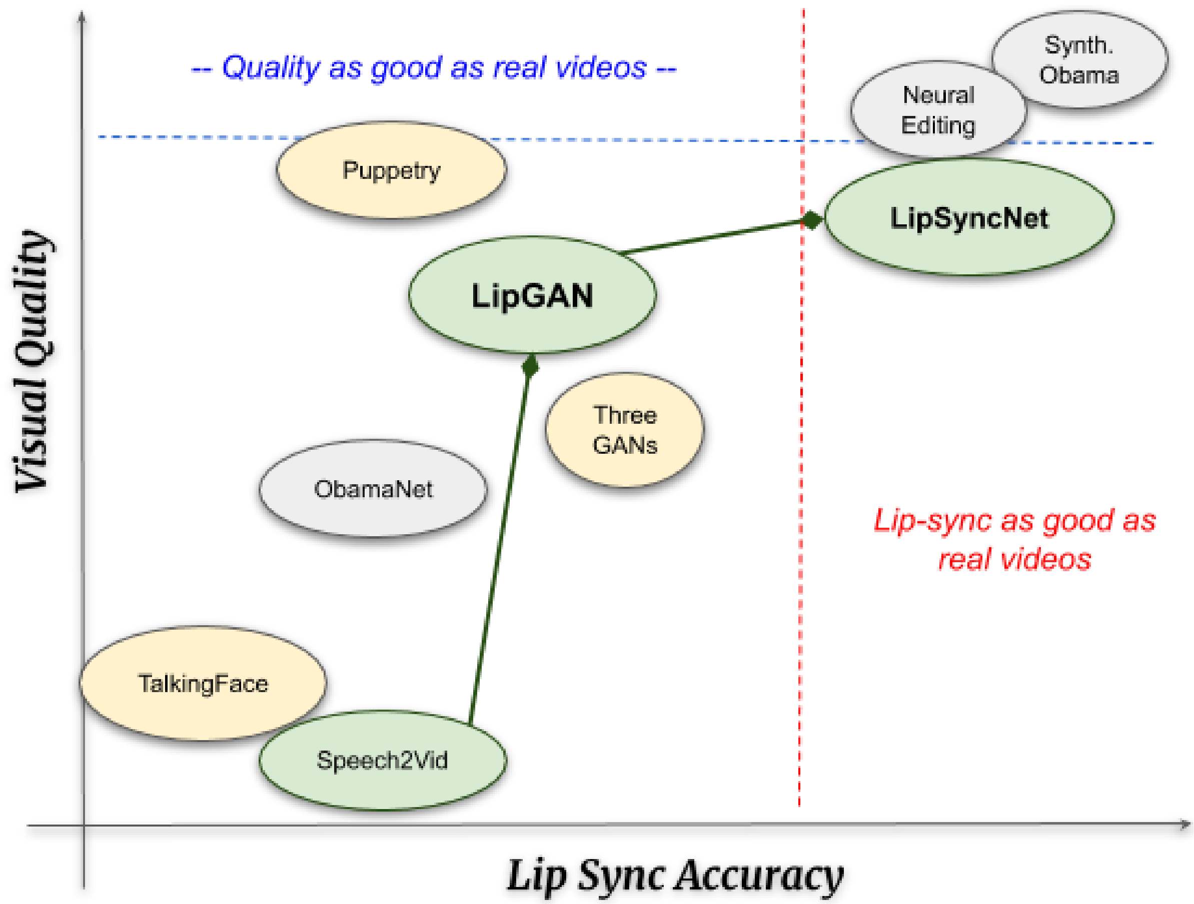

Figure 3.1 The current landscape of talking face videos. We consider several works in this space and illustrate here the key works that achieve improvements in two specific aspects: visual quality and lip-sync accuracy (focusing on the latter in this chapter). The works are color-coded based on their nature. Grey colored works denote speaker-specific models that need a large amount of data of a person. These are (i) Synth. Obama [53], (ii) Neural Editing [54], and (iii) ObamaNet [44]. Yellow color bubbles denote works that are multi-speaker but are constrained by vocabulary (TalkingFace [70] and Three-GANs [61]) or need additional data for adaptation to a new speaker (Neural Voice Puppetry [54]. In green, we represent works [14] that are speaker-independent and are expected to lip-sync videos in the wild. Our LipGAN model improves the visual quality by a significant margin. On the other hand, we now propose a novel model that is significantly more accurate than all the previous works on unconstrained real-world videos. We see that this model almost bridges the gap between speaker-specific and speaker-independent models in terms of both quality and lip-sync accuracy.

Our work builds upon this latter class of speaker-independent works that aspire to lip-sync talking face videos of any identity and voice. We find that these models that work well for static images are unable to accurately morph the large variety of lip shapes in unconstrained video content, leading to significant portions of the generated video being out-of-sync with the new target audio. A viewer can recognize an out-of-sync video segment as small as just $\approx 0.05 - 0.1$ seconds [18] in duration. Thus, convincingly lip-syncing a real-world video to an entirely new speech is quite challenging, given the tiny degree of allowed error. Further, the fact that we are aiming for a speaker-independent approach without any additional speaker-specific data overhead makes our task even more difficult. Real-world videos contain rapid pose, scale, and illumination changes and the generated face result must also seamlessly blend into the original target video.

3.1.2 Contributions

We start by inspecting the existing speaker-independent approaches for speech to lip generation. We find that these models do not adequately penalize wrong lip shapes, either as a result of using only reconstruction losses or weak lip-sync discriminators. We adapt a powerful lip-sync discriminator that can enforce the generator to consistently produce accurate, realistic lip motion. Next, we re-examine the current evaluation protocols and devise new, rigorous evaluation benchmarks derived from three standard test sets. We also propose reliable evaluation metrics using SyncNet [18] to precisely evaluate lip sync in unconstrained videos. We also collect and release ReSyncED, a challenging set of real-world videos that can benchmark how the models will perform in practice. We conduct extensive quantitative and subjective human evaluations and outperform previous methods by a large margin across all benchmarks. Our key contributions/claims are as follows:

- We propose a novel lip-synchronization network, Wav2Lip, that is significantly more accurate than previous works for lip-syncing arbitrary talking face videos in the wild with arbitrary speech.

- We propose a new evaluation framework, consisting of new benchmarks and metrics, to enable a fair judgment of lip synchronization in unconstrained videos.

- We collect and release ReSyncED, a Real-world lip-Sync Evaluation Dataset to benchmark the performance of the lip-sync models on completely unseen videos in the wild.

- Wav2Lip is the first speaker-independent model to generate videos with lip-sync accuracy that matches the real synced videos. Human evaluations indicate that the generated videos of Wav2Lip are preferred over existing methods and unsynced versions more than 90% of the time.

3.2 Delving into the Existing Speaker-Independent Models

Our focus is to lip-sync arbitrary identities and voices in the wild. We will now highlight the key aspects missing in the existing models that aspire to do the same task. Before that, we present a brief recap of these existing models.

3.2.1 Brief Overview of the Existing Models

Despite the rise in the number of works on speech-driven face generation, there are surprisingly very few works that have been designed to lip-sync videos of arbitrary identities, voices, and languages. They are not trained on a small set of identities or a narrow vocabulary. This allows them to, at test time, lip-sync random identities for any speech. To the best of our knowledge, only two such prominent works exist in the current literature. The first is the Speech2Vid model [14]. Note that [14] is an extended version of [14]. We will continue referring to [14] for consistency with the previous chapters. The second model after the Speech2Vid model is our LipGAN model detailed in the previous chapter. Both these works formulate the task of learning to lip-sync in the wild as follows: *Given a short speech segment S and a random reference face image R, the task of the network is to generate a lip-synced version L_g of the input face that matches the audio*. Additionally, the LipGAN model also inputs the target face with bottom-half masked to act as a pose prior. We found that this was crucial, as it allowed the generated face crops to be seamlessly pasted back into the original video without any further post-processing. LipGAN also trains a discriminator in conjunction with the generator to discriminate sync or out-of-sync audio-video pairs. Both these models, however, suffer from a significant limitation: they work very well on static images of arbitrary identities but produce inaccurate lip generation when they try to lip-sync unconstrained videos in the wild.

3.2.2 Why are the existing models inaccurate on real-world videos?

Our core architecture can be summed up as *"Generating accurate lip-sync by learning from a well-trained lip-sync expert"*. To understand this design choice, we first identify two key reasons why existing architectures produce inaccurate lip-sync for videos in the wild. We argue that the loss functions, namely the L1 reconstruction loss used in both the existing works, LipGAN and Speech2Vid [14] and the discriminator loss in LipGAN are inadequate to penalize inaccurate lip-sync generation.

3.2.3 Pixel-level Reconstruction loss is a Weak Judge of Lip-sync

The face reconstruction loss is computed for the whole image, to ensure correct pose generation, preservation of identity, and even background around the face. The lip region corresponds to less than 4% of the total reconstruction loss (based on the spatial extent), so a lot of surrounding image reconstruction is first optimized before the network starts to perform fine-grained lip shape correction. This is further supported by the fact that the network begins morphing lips only at around half-way ($\approx 11^{th}$

epoch) through its training process ($\approx$ 20 epochs as stated in the previous chapter). Thus, it is crucial to have an additional discriminator to judge lip-sync, as also done in LipGAN. But, how powerful is the discriminator employed in LipGAN?

3.2.3.1 A Weak Lip-sync Discriminator

We find that the LipGAN's lip-sync discriminator is only about 56% accurate while detecting off-sync audio-lip pairs on the LRS2 test set. For comparison, the expert discriminator that we will use in this work is 91% accurate on the same test set. We hypothesize two major reasons for this difference. Firstly, LipGAN's discriminator uses a single frame to check for lip-sync. In an ablation study later in the chapter, we show that lip-sync is very hard to discriminate from just a single frame. But the difficulty in lip-sync discrimination can also be illustrated visually, as we do in Figure 3.2. We can observe that if we feed a single frame to the discriminator and the surrounding speech, the model cannot judge it accurately. This is because an utterance occurs over a series of frames. Consequently, several utterances can share common intermediate lip shapes, but their overall sequential lip motion would be different. Thus, using a single-frame to discriminate as done in previous works is far from ideal. Secondly, the

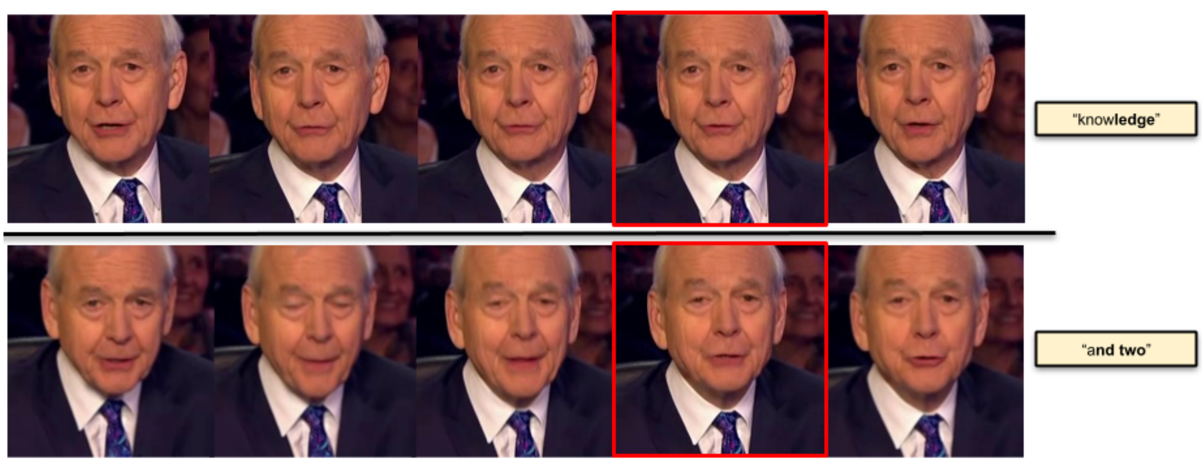

Figure 3.2 A single-frame is hard to discriminate lip-sync. We can see that in a single frame (marked red), lip shapes can look identical, and either one can be in place of the other, if not for the surrounding context. Thus, given a surrounding context of 5 frames, for example, the sequence of lip movements is more distinctive to discriminate.

generated images during training contain many artifacts due to large scale and pose variations between the reference and target images. In Figure 3.3, we show different examples of artifacts in the generated faces during training. We hypothesize that training on these noisy generated images, as done in LipGAN, significantly hinders the network from learning the fine-grained audio-visual association. The artifacts in these generated faces allow the discriminator to use them as shortcuts to discriminate the generated images. Indeed, we show in an ablation study later in the chapter that even fine-tuning our own proposed expert discriminator on the generated frames in a GAN setup also leads to a large drop in off-sync

Figure 3.3 Fine-tuning the discriminator on the noisy generated faces deteriorates its ability to penalize lip-sync. The artifacts in these generated faces allow the discriminator to use them as shortcuts to discriminate the generated images. This clearly is not the case with a discriminator trained only on real faces as it can only judge differences in lip-sync.

detection accuracy. We argue and show that the "real", accurate concept of lip-sync captured from the actual video frames can be used to accurately discriminate and enforce lip-sync in the generated images.

3.3 A Lip-sync Expert Is All You Need

Based on the above two findings, we propose to use a pre-trained expert lip-sync discriminator that is accurate in detecting sync in real videos. Also, it should not be fine-tuned further on the generated frames like it is done in LipGAN. One such network that has been used to correct lip-sync errors for creating large lip-sync datasets [1, 4] is the SyncNet [18] model. We propose to adapt and train a modified version of SyncNet [18] for our task.

3.3.1 Overview of SyncNet

SyncNet inputs a window V of T_v consecutive face frames (lower half only) and a speech segment S of size $T_a \times D$, where T_v and T_a are the video and audio time-steps respectively. It is trained to discriminate sync between audio and video by randomly sampling the audio window $T_a \times D$ to be either aligned with the video (sync) or from a different time-step (out of sync). It contains a face encoder F and an audio encoder A, both comprising of a stack of 2D-convolutions. L2 distance is computed between the embeddings generated from these encoders, and the model is trained with a max-margin loss to minimize (or maximize) the distance between synced (or unsynced) pairs.

3.3.2 Our expert lip-sync discriminator

We make the following changes (Figure 3.4) to the previously proposed SyncNet model to improve its convergence speed and accuracy. Firstly, instead of feeding gray-scale images concatenated channel-wise as in the original model, we feed color images. Second, our model is significantly more deeper, with residual skip connections [31]. Third, we use a different loss function: cosine-similarity with binary cross-entropy loss. That is, we compute a dot product between the ReLU-activated video and

Chapter 5

Conclusions

In this thesis, we extensively investigated the problem of speech-driven lip synthesis in the wild. We identified major issues in existing works and designed two novel approaches that significantly improve the quality and accuracy of lip-synced talking face videos.

In Chapter 2, we started by introducing a basic, skeletal framework for lip-syncing an arbitrary identity in the wild. We then delved into this framework and investigated why existing speaker-independent methods do not produce results comparable to the models that are trained for a specific speaker. Even more importantly, we found that they completely lose visual information such as pose, expressions, background, etc. The reason for this was because of inadequate visual priors. We solved this with a novel in-painting approach that allowed us to control the pose, expression, background in the generated frame. Thus, our new model, LipGAN, generated high-quality lip-sync videos compared to previous works.

In Chapter 3, we investigated why existing approaches, including LipGAN, are consistently inaccurate while lip-syncing videos in the wild. We found that both L1 reconstruction loss and the discriminator penalization in LipGAN are inadequate to penalize the inaccurately generated lip shapes. Thus, in Wav2Lip, we employed a powerful lip-sync discriminator that is far more accurate. Our final model, Wav2Lip, was able to lip-sync videos in the wild with remarkable accuracy and quality.

Finally, in 4, we demonstrated our model for several attractive real-world applications. Our work opens up new avenues for learning and synthesizing face and speech attributes from one another. Future works can aim to tackle problems such as inferring emotion and expressions from speech. Learning head gestures, hand gestures from the speech is also worth exploring. Another interesting problem would be to handle the length differences between the video and speech before and after translations. Adapting the lip synthesis models for artificial data like CGI faces and TTS generated voices can lead to even more applications. We hope that this work encourages future research works in similar promising directions.

Related Publications

1. **Towards Automatic Face-to-Face Translation**, Prajwal K R*, Rudrabha Mukhopadhyay*, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C V Jawahar, *ACM International Conference on Multimedia 2019, Nice, France.*

2. **A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild**, Prajwal K R*, Rudrabha Mukhopadhyay*, Vinay Namboodiri, C V Jawahar, *ACM International Conference on Multimedia 2020, Seattle, USA.*

Other publications during MS that are not part of this thesis:

1. **Towards Increased Accessibility of Meme Images with the Help of Rich Face Emotion Captions**, K R Prajwal, C V Jawahar, Ponnurangam Kumaraguru, *ACM International Conference on Multimedia 2019, Nice, France.*

2. **Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis**, Prajwal K R*, Rudrabha Mukhopadhyay*, Vinay Namboodiri, C V Jawahar, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, Seattle, USA.*

3. **IndicSpeech: Text-to-Speech Corpus for Indian Languages**, Nimisha Shrivastava, Rudrabha Mukhopadhyay*, Prajwal K R*, C V Jawahar, *The 12th Language Resources and Evaluation Conference (LREC), 2020, Marseille, France.*

4. **Visual Speech Enhancement Without A Real Visual Stream**, Sindhu Hegde, Prajwal K R*, Rudrabha Mukhopadhyay*, Vinay Namboodiri, C V Jawahar, *(under review)*.

(*equal contribution)

Bibliography

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

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018.

[4] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018.

[5] D. Aneja and W. Li. Real-time lip sync for live 2d animation. *arXiv preprint arXiv:1910.08685*, 2019.

[6] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[8] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

[9] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

[10] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[11] M. Busta, L. Neumann, and J. Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.

[12] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[13] L. Chen, H. Zheng, R. K. Maddox, Z. Duan, and C. Xu. Sound to visual: Hierarchical cross-modal talking face video generation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019.

[14] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

[15] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[16] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[18] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[20] B. Dancygier and L. Vandelanotte. Internet memes as multimodal constructions. *Cognitive Linguistics*, 28(3):565–598, 2017.

[21] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge dataset, 2020.

[22] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

[23] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[24] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[25] O. Fried and M. Agrawala. Puppet dubbing. *arXiv preprint arXiv:1902.04285*, 2019.

[26] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[27] S. Furukawa, T. Fukusato, S. Yamaguchi, and S. Morishima. Voice animator: Automatic lip-synching in limited animation by audio. In *International Conference on Advances in Computer Entertainment*, pages 153–171. Springer, 2017.

[28] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[30] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee. Deep fake image detection based on pairwise learning. *Applied Sciences*, 10:370, 2020.

[33] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448, 2017.

[34] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[35] A. Jamaludin, J. S. Chung, and A. Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11-12):1767–1779, 2019.

[36] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490, 2018.

[37] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[38] S. Kankanahalli. End-to-end optimized speech coding with deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2521–2525. IEEE, 2018.

[39] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[40] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

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

[42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[43] V. Kuleshov, S. Z. Enam, and S. Ermon. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017.

[44] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[45] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[46] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[47] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[48] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[49] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[50] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[51] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[52] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[53] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[54] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. *arXiv preprint arXiv:1912.05566*, 2019.

[55] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[56] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection, 2020.

[57] E. Tursman, M. George, S. Kamara, and J. Tompkin. Towards untrusted social video verification to combat deepfakes via face geometry consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[58] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[60] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018.

[61] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019.

[62] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[64] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[65] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[66] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468, 2019.

[67] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[68] K. Zhen, M. S. Lee, J. Sung, S. Beack, and M. Kim. Efficient and scalable neural residual waveform coding with collaborative quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 361–365. IEEE, 2020.

[69] K. Zhen, J. Sung, M. S. Lee, S. Beack, and M. Kim. Cascaded cross-module residual learning towards lightweight end-to-end speech coding. *arXiv preprint arXiv:1906.07769*, 2019.

[70] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.