A Deep Learning Approach to Learn Lip Sync from Audio

BY

Mazharul Islam Bhuiyan

Id: 171-15-1425

Milon Mahato

Id: 171-15-1472

Md. Nazmul Hassan

Id: 171-15-1487

Md. Habibur Rahman

Id: 171-15-1471

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Md. Reduanul Haque

Senior Lecturer

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised By

Mr. Md. Mahfujur Rahman

Lecturer

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

14th January 2021

DECLARATION

We hereby declare that, this Paper has been done by us under the supervision of **Mr. Md. Reduanul Haque, Senior Lecturer, Department of Computer Science & Engineering** & co-supervision of **Mr. Md. Mahfujur, Lecturer, Department of Computer Science & Engineering** Daffodil International University. The work embodied in this paper has not been submitted to any other University or Institute for the award of any degree or diploma.

Supervised by:

Mr. Md. Reduanul Haque
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Md. Mahfujur Rahman
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Mazharul Islam Bhuiyan
ID: 171-15-1425
Department of CSE
Daffodil International University

Milon Mahato
ID: 171-15-1472
Department of CSE
Daffodil International University

Md. Nazmul Hassan
171-15-1487
Department of CSE
Daffodil International University

Habibur Rahman
171-15-14 71
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty Allah for his divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our immense our indebtedness to **Mr. Md. Reduanul Haque**, **Senior Lecturer**, Department of Computer Science & Engineering (CSE), Daffodil International University, Ashulia, Dhaka. Deep Knowledge & enthusiastic interest of our supervisor in the field of "*Machine Learning*" to carry out this paper. His endless patience, authoritative guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this papers.

We would like to express our heartiest gratitude to **Mr. Md. Reduanul Haque, Mr. Mahfuzur Rahman Raju,** and Head, Department of Computer Science & Engineering, for his kind help to finish our papers and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the faithful support and patients of our parents.

A Deep Learning Approach to Learn Lip Sync from Audio

Abstract

With accurate lip-sync of speaker independent, we synthesize several high-quality videos for sake of generating expected target video clip from the composite synthesized video. In our work we explore several related works of lip-syncing, out-of-sync, talking face generation, speaker independent target video content creation from input audio stream containing some limitation & failure and we also implement as better lip-sync by training our models which is not rely on specific speaker. Besides, in this work we detect key reason for the mentioned problems and improve the difficult factors with new evaluation strategies then solve as better output returning like Wav2lip model. By the way, more realistic matched lip-syncing appearance of any individual speaking video from any voices or input audio clip along with mapping RNN is an incredible outcome since it can generate proper mouth texture.

KEYWORDS: Lip-sync, Synthesized video, Talking Face Generation, RNN, Video Creation

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

CHAPTER 1
Introduction

The advancement in new information and technology, the aspects of media from the few past years making an epoch-making change to the new era of audio and visual things with the popularity and essentiality of generating creative media contents. Now the fact is how can imagine and what interaction should be made with machine to work with datasets of speech or input audio signal, targeting any public video to combine with given audio for the sake of replication of random peoples with great aim.

Hence, this implementation of lip sync from voice signal or audio is an interesting topics which is also badly in needed in several video calling tools, in live conference for better experience in signal lost for continuing voice with video from pre deep learned model to provide proxy and also in emergency moment we need instant backup like for newsreader, during online appointment of different language speaker and also in increasing economic growth applying that technique in gaming, digital movie industries which will spread knowledge, culture as world-wide. Any language speaker can enjoy and develop themselves in all aspects being engaged with this deep learning. Recently, the research community also concentrating on the synthesizing speaking appearance of faces from audio [1].

By the way, making synthesized talking mouth or lip synced from inputted audio speech is remarkably an issue which is expressed in [2]. “In the ATR Interpreting Telecommunications Research Laboratory (ATR-ITL), research has been carried out since 1993 to achieve translation of speech dialogue among different languages. Related research has been done in the ATR Spoken Language Translation Research Laboratory since 2000, with the aim of expanding the technology to more extensive multiple domains, many languages, distant speech, and colloquial speaking style” [3] .

Especially, in the field of deep learning voice to video synchronization, language dubbing with accurate lip synchronization, 3d film or animation video creation, and also in gaming with random

famous characters are incredibly most demanded thing but in reality, creating or implementing these contents are more complex and quite challenging. There are so many online public conversation, videos, visual conferences are available where the behavior of their discussion, gesture, expression appearances and style of interaction are visible but analyzing them for replication to random people are difficult. Because, accurately lip sync with time and natural structure of face, mouth texture varies one people to another.

Though there is a difficulty of accurate lip syncing from input audio but in this paper, we show better than other previous related works ensuring the quality video of face movement, texture, talking generation with suitable accuracy from input stream audio. Moreover, ours work in this manner is not limited to specific people and more accurately synchronize the lip from speech based on our trained model as well. There was challenge but our methodology and chosen effective algorithm with effective steps provide desired outcome which is clearly expressed in the following below analysis and below mentioned figures sections along with remarked results in the table.

CHAPTER 2

RELATED WORKS

Initially, we observe on the implementations of talking face which are either obliged by the scope of personalities they can create whether the reach of lexicon they are restricted to. Some sequence of processes related precise lip-sync with audio clip have attained recently. From given audio input, the attempt is taken to make visual animation of faces in [4]. Few works in [5] creation of realistic Obama speaking with video was implemented but there was also difficulty due to texture variations of mouth imposing the restriction on capabilities of generalization. For new people or input audio track synthesize is unable where dataset training is performed only for particular speaker. From text this work [6] achieved lip-sync of photo-realistic manner where a mapping was learnt between input audio & analogous landmark of lip. Lip-sync from input audio stream, video of High-Quality (HQ) is synthesized by Suwajanakorn et al. [7] of Barack Obama talking accurately that is the closet to our works. But our works perform better in this regarded since instead using the model of tradition vision we use Neural Network.

A research work in 2019 along this line [8] is the mention to consistently alter recordings of individual speakers by adding or eliminating phrases from the talk [9]. They actually need an hour of information per speaker to accomplish this undertaking. By the way, another work proposes for sake of making video frames with lip-sync comprising with Long Short-Term Memory and Pix2Pix [10] resolves the Obama Net by the following [6].

Though the mentioned works perform well but there need impermanent dynamics. Because, without displaying it this work frequently brings about transiently mixed up recordings with unsuitable visual quality. To regard the mentioned reliance between video outlines, the proposal in [11] a transient GAN, equipped for producing a video of a talking head from a sound clip and a solitary actually picture. This strategy can catch the elements of the whole face to create transiently intelligible recordings by utilizing the RNN based generator and arrangement discriminator [12]. Nevertheless, the synthesized visual recordings experience the unfavorable effects of unaltered appearance of face, and the visual quality is also unsuitable.

A most current work [13] attempts to limit this information overhead by utilizing a approach of two phases, where they initially learn independently of any speaker attributes and afterward learn a delivering planning with approximately 300 seconds of information of the expected speech teller. But there is the vocabulary limitation of that current work and for new individual target speaker the previous training data has to clear to produce for the teller. Besides, the time-delayed procedure of LSTM is our motivation to proceed such kind of work like forecasting the facial landmarks and mapping that to target video frames. However, to enhance the accuracy rate of prediction, applied multimodal learning with proposed conventional computer graphics independent model of generating video from face.

Recently in [14] they train limited dataset sets of vocabulary or words which complicates the mapping with huge diversity of phoneme-viseme to the realistic videos. It's mentioned that a Face2vid network is introduced for realizing mouth landmarks along with Audio & Text mining for synthesis video based on facial texture and condition for generating talking face by a recent research work [12, 15]. Based on video frames and realization on lips & facial movement's translation, a model was proposed by [16] which inspired to [12]. Generating image and video handling with the help of Generative Adversarial Network and the advancement of GAN facilitates to map random vector's sequence towards a sequence of video frames. And that Motion and

Content decomposed GAN is also proposed by [17] for synthesizing video. Another work [18] formulates that if a short speech segment-S & a random reference face image becomes-R, then the task of the network is to make a lip-synced version Lg of the input face which matches the audio. Furthermore, the Lip Generative Adversarial Network model likewise inputs the objective face with base half concealed to go about as a posture earlier. This was vital as it permitted the created face yields to be consistently stuck once again into the first video minus any additional post-processing. It likewise prepares a discriminator related to the generator to separate in-a state of harmony or out-of-sync sound video sets. Both these works, notwithstanding, experience the ill effects of a huge impediment: they function admirably on static pictures of self-assertive personalities yet produce wrong lip age when attempting to lip-sync unconstrained recordings in nature.

To generate talking face from speech in the aspect of unconstrained manner, "Despite the rise in the number of works on speech-driven face generation, surprisingly, very few works 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 small vocabulary. This allows them to, at test time, lip-sync random identities for any speech" [9].

CHAPTER 3
METHODOLOGY

Our epic model creates altogether more precise lip-synchronization in powerful, unconstrained talking face recordings. Quantitative measurements show that the lip-sync in our produced recordings are nearly on a part with genuine synchronized recordings. The following visualizations dictates the proposed techniques to implement the desired lip-syncing video from given audio as input. By the way, the lip motion with uncanny valley problem and inaccurate mouth texture with blurry teeth appearance in different video environment and stock footages are concerned to make more realistic by synthesizing facial part with relevant speeches.

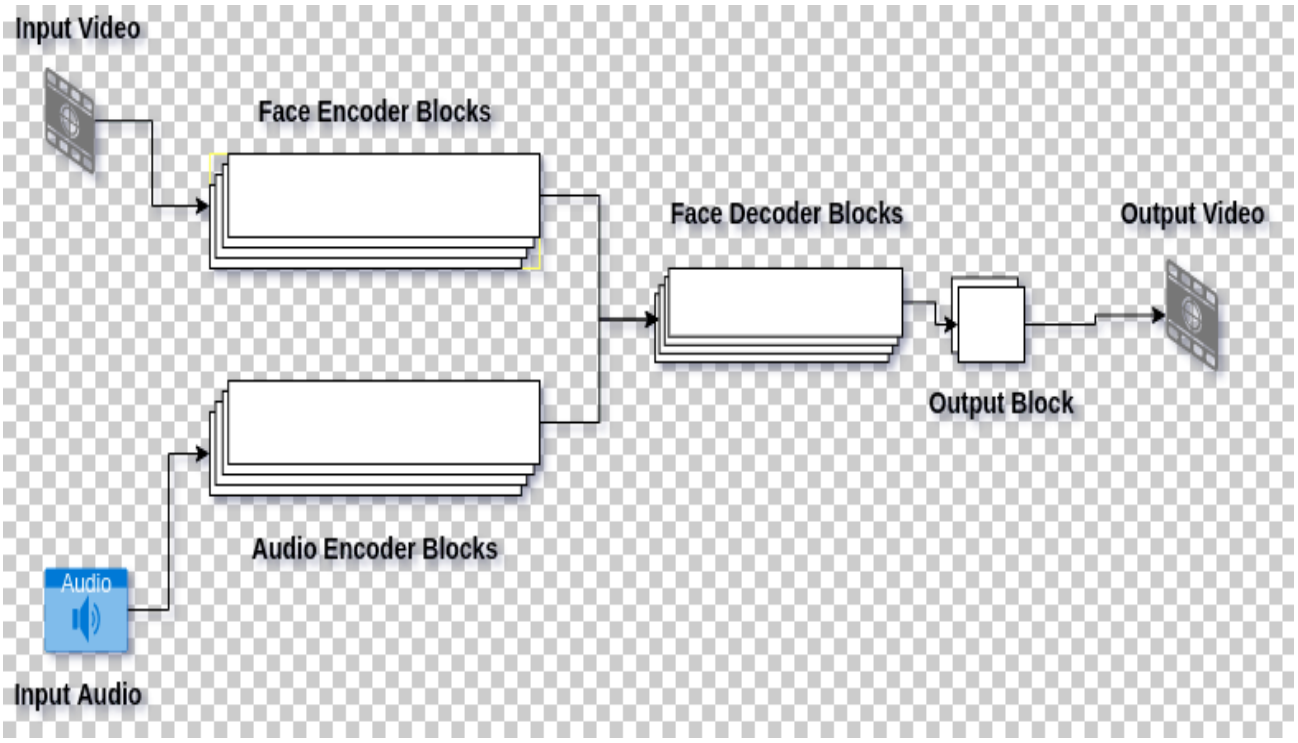
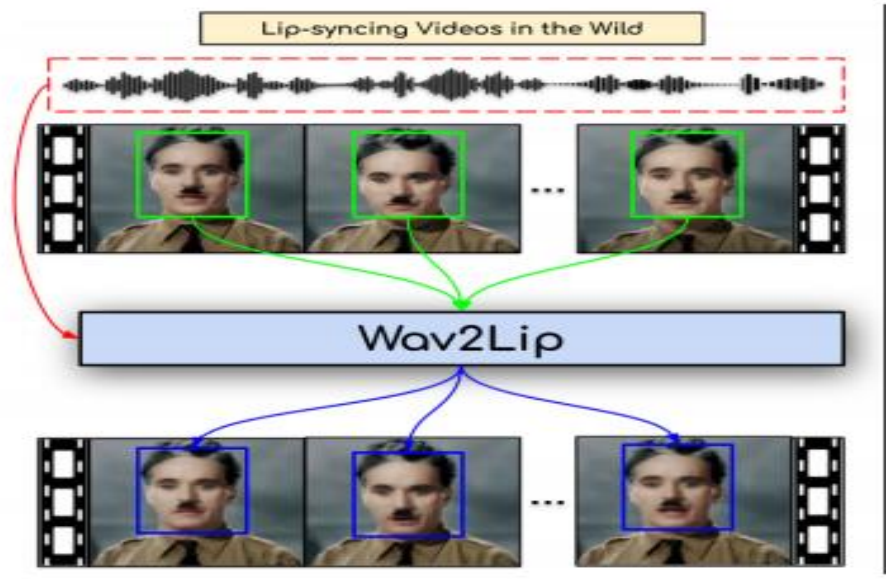


Figure.3.1.1: Model Structure of Audio to Video Encoding and Decoding.



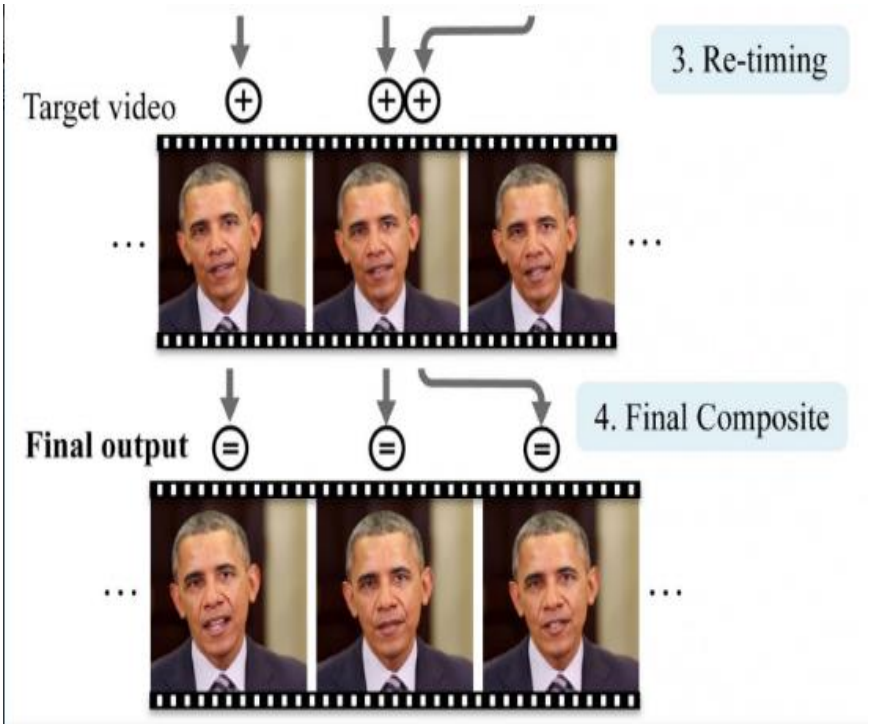


Figure -3.1.2: Lip-synced composite video

3.1 AUDIO TO ACCURATED LIP-SYNCED VIDEO

To synthesize individual's video track by learning and mapping audio input to video as output along with lower to higher dimensional signal is quite challenging though we performed with some sequence mapping. We made more efficient lip-sync from proper trained on several case studied expert of lip-sync. We focuses on the basic reasons for not returning the expected videos from given input audios, used to generate the lip-synced video from existing system where for inaccurate lip-sync the LipGAN's discriminator loss & L1 reconstruction loss are noticed. Hence, for the sake of making speaker independent lip-sync video, animation face talking with adequate quality, video synthesizing and facial different landmark predictions are done besides with input audio or text by detecting facial key points with the help of dlib face detector [19] from per frame of image for de-normalizing and time delayed Long Short Term Memory for predicting landmarks of faces. The inexplicit face regions like edge-background can be removed by extract or detecting the canny edge detector [20]where the combination these edges & facial landmarks create a sequence of synced face sketches.

3.1.1 What is Lip Sync?

Lip-sync intends to move one's mouth in a joint effort with a pre-recorded tune or soundtrack. The words lip-sync and lip-synchronize are shortenings of the term lip synchronization. Lip-synchronizing might be utilized in film, when naming English into an unfamiliar film or essentially embedding's changed discourse into a scene. A great many people partner lip-matching up with music. Initially, lip-synchronizing was utilized in TV, where it was savvier to just play a pre-recorded track as opposed to introduce artists playing live. As time went on, lip-synchronizing was utilized to mask an absence of ability.

For example, in singing perspectives, Lip Syncing is the specialty of emulating a melody to make it seem as though you are really singing. There is the craft of doing this which is extreme however when it is done well there is a great deal of significant worth in lip matching up for evident reasons. The music should be in time with the individual who is emulating if not there will be a slack between the two losing the sensible impact.

3.1.2 Why Lip Sync is used?

Lip Syncing can be truly advantageous to craftsmen. It can mean they don't need to put strain on their voices meaning they will have the option to accomplish more shows and thusly bring in more cash. The senseless factor is in motion pictures the entertainer will most likely be unable to sing. Therefore the lip matching up can help in making everything more practical and improving the nature of the film.

In film creation, lip-synchronizing is frequently important for the postproduction stage. Naming unknown dialect movies and causing enlivened characters to seem to talk both require expand lip-synchronizing. Numerous computer games utilize lip-synchronized sound records to establish a vivid climate in which on-screen characters have all the earmarks of being talking. In the music business, lip-synchronizing is utilized by vocalists for music recordings, TV and film appearances and a few kinds of live exhibitions. Lip-synchronizing by vocalists can be questionable to fans going to show exhibitions who hope to see a live presentation. There are so many significant reasons for which Lip-Syncing is badly in needed some of which are mentioned below:

- For dubbing educational and professional video tutorials internationally based on dubbing requirements. It’s mentioned that in COVID-19 pandemic situation any person/learner can learn, take online courses and even achieving online degree from international renowned institutions and foreign offering countries. But several language from several countries are not familiar to several country belonging people. Hence, language dubbing becomes more important thing for learning, sharing knowledge and to be skilled up proper understanding to adapt with world.
- For spreading movies widely as dubbed movies since the popularity of different cultural movies are increasing and also increases the economic health of correspond dubbed movie/film industries or related countries.
- In video conferencing or live online meeting during such kind of COVID-19 pandemic situation or even in normal time lip-syncing can help when translating with proper lip land-marking appearance.

- Creating missing video segments of video calling with uttered audio when connection signal losses for a temporal period due to network issues or others.
- Lip-syncing with most popular CGI characters in animation movie, tutorials, and also in Gaming sectors to get the demand in international marketplace.
- In social platforms recently the generated GIFs, memes, and stickers are getting most popular and also shared because of the combination of accurate Lip-syncing with these GIFs, stickers and memes.

Moreover, besides the advancement of new DL (Deep Learning) & ML (Machine Learning) strategies with AI applying the media contents and skills practices results into real VS fake content checking will contribute in the upcoming days.

Many individuals are finding their survey to some degree disturbing for reasons that aren't too self-evident. Stephen Dawson clarifies the significance of lip sync change and why you may have to do it physically. Some time ago we got our home amusement in basic, clear ways. The video and the sound were multiplexed together into a solitary RF stream. Your TV would de-mux them, send the sound off the speaker and the image off to the CRT. Both would regard their particular sign as a straightforward stream and follow up on it immediately. In the event that the sound and picture floated separated, at that point there was an off-base thing at the TV station. However, presently the standard they don't coordinate. The explanation is basic: while the preparing of sound has been eased back up somewhat because of current innovation, the handling of the image has been eased back significantly more. Luckily as people we have a reasonable piece of capacity to bear confounds among picture and sound, and our minds adapt to it. Yet, in actuality, the jumble is the opposite way around. Sound goes at around 340 meters for each second, light at around 300 million meters for every second. 340 meters isn't far: the length of three or four football fields, contingent upon the code you follow [21]. In the event that something makes a commotion that distant you will see it, all things considered, quickly. However, it will require an entire second for the sound to get to your ears. In the event that you are conversing with somebody on the other wide of the room – state, 5m away – of course you will see their lips moving at a similar moment that they are really moving, however their voice will be 15 milliseconds behind it. Presently 15 milliseconds – 15 one-thousandths of a second – may seem like scarcely any an ideal opportunity to make a whine about [21].Yet, our minds use timing prompts a lot more modest than that. We tell the heading of

a sound to a great extent by the contrasts between when the sound arrives at our left and right hears, and those distinctions are typically a small amount of a millisecond. Luckily in everyday life the different preparing circuits of our cerebrums do somewhat wizardry and make the voice of the individual conversing with us sound like it is absolutely coordinated to the development of that individual's lips. To a limited extent. On the off chance that they are excessively far away, at that point the change is not, at this point made, and the lips and voice don't coordinate. In any case, things are in reverse with regards to the image and sound in a home theater setup. The image gets to you slower than the sound does as a result of the idea of the playback framework. The sound – even with present day computerized frameworks – is postponed scarcely by any stretch of the imagination. Advanced disentangling and DSP control of the sound all occurs in a millisecond or two. Yet, the video is an alternate issue.

3.1.3 How Lip Sync works?

The Lip-syncing technique depends on various strategies including the term accurate Lip-sync, Synthesizing video, Talking Face Generating, RNN (Recurrent Neural Network) approach following, Composited Video Creation with sequence mapping, and GAN (Generative Adversarial Network) to Lip Syncing as LipGan and requirement dependent CNN (Convolution Neural Network) using for appearance uncanny problem and blurry facial segment like for un sharpening teeth appearance issues.However, in our paper which methodology and steps we followed are described within the whole paper and in details in below are the mentioned approaches with some background study related to our proposed research work.

Table 3.1.3: Face Encoder Block with torch Size Within different Network Layer

Block	Layer (type)	Output Shape	Params #
face_encoder_blocks	Conv2d	torch.Size([16, 6, 7, 7])	8171056
	BatchNorm2d	torch.Size([16])	
	ReLU	torch.Size([16])	
	Conv2d	torch.Size([32, 16, 3, 3])	
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([32, 32, 3, 3])	
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([32, 32, 3, 3])	
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([64, 32, 3, 3])	
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	
	Conv2d	torch.Size([64, 64, 3, 3])	
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	
	Conv2d	torch.Size([64, 64, 3, 3])	
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	

1. Find or record a video of the individual (or use video talk apparatuses like Skype to make another video) for the neural organization to gain from. There are a long period of time of video that as of now exist from interviews, video visits, motion pictures, TV programs and different sources, the scientists note.

2. Train the neural organization to watch recordings of the individual and make an interpretation of various sound sounds into essential mouth shapes.

3. The framework at that point utilizes the sound of a person's discourse to create reasonable mouth shapes, which are then united onto and mixed with the top of that individual. Utilize a

humble move to empower the neural organization to envision what the individual will say straightaway.

4. Presently, the neural organization is intended to learn on each person in turn, implying that Obama's voice — talking words he really articulated — is the lone data used to "drive" the integrated video. Future advances, nonetheless, incorporate assisting the calculations with summing up circumstances to perceive an individual's voice and discourse designs with less information, with just an hour of video to gain from as a sample example.

Table 3.1.4: Face Decoder Block with torch Size Within different Network Layer

face_decoder_blocks	Conv2d	torch.Size([512, 512, 1, 1])	25291072
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	Conv2d	torch.Size([1024, 512, 3, 3])	
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	Conv2dTranspose	torch.Size([512, 512, 3, 3])	
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	Conv2d	torch.Size([1024, 512, 3, 3])	
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	Conv2d	torch.Size([512, 512, 3, 3])	
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	Conv2d	torch.Size([512, 512, 3, 3])	
	BatchNorm2d	torch.Size([512])	
	ReLU	torch.Size([512])	
	ConvTranspose2d	torch.Size([768, 384, 3, 3])	
	BatchNorm2d	torch.Size([384])	
	ReLU	torch.Size([384])	

3.1.4 What's LipGAN?

LipGAN is associated with Generative Adversarial Network (GAN) which is the class of ML (Machine Learning) frameworks. By the way LipGAN permits us to modify the lip-developments of an individual in a video to coordinate a given objective brief snippet. This encoding is then used to remake a similar face however with various situation of the lips in a state of harmony with the info sound highlights.

Moreover, with LipGAN technology we can conceivably address lip sync blunders in named Animations, films, lectures. LipGAN can deal with discourse in any language and is strong to foundation commotion. It provides the facility with glue faces once more into the first video with negligible/no relics. Moreover, it can deal with in-the-wild face postures and looks of any individual.

Lip-Sync, or Lip synchronization may be defined by also Audio sound to Video synchronization but the absence of syncing denotes the lip-sync mistake. It alludes to the overall planning of sound track and video or picture parts during creation, after creation during blending, transmission, gathering and play-back handling. That synchronization can be an issue in live broadcasting, Television, video-conferencing, or film/movie dubbing and so on. In industry phrasing, the synchronization mistake is communicated as a measure of time the sound leaves from ideal sync with the video where the number of positive time shows the sound leads the video & a negative number demonstrates the sound slacks the video. This wording and normalization of the numeric lip-sync blunder is used in the expert transmission industry as proven by the different expert with some guidelines like and different references underneath.

Advanced or simple sound video transfers or video records as a rule contain a type of synchronization component, either as interleaved video and sound information or by unequivocal relative time stamping of information. The preparing of information should regard the overall information timing by for example extending between or introduction of got information. On the off chance that the handling doesn't regard the AV-sync blunder, it will increment at whatever point information gets lost in light of transmission mistakes or as a result of missing or confused preparing.

3.1.5 Gan (Generative Adversarial Network) on Lip-Syncing

3.1.5.1 Synthesis Video to Video

For Video2Video synthesizing generally Conditional GANs (Generative Adversarial Network) can be used to landmark the area of lip properly.

Because of the intrinsic multifaceted nature of the issue of video-to-video combination, this theme remained moderately neglected before. Contrasted with its picture partner – picture to-picture union, many less investigations have investigated and handled this issue. Contending that a broadly useful answer for video-to-video blend has not yet been investigated in the earlier work (instead of the picture partner – picture to-picture interpretation), the specialists contrast and

benchmark this methodology and a solid gauge that joins a cutting edge video style move calculation with a best in class picture to-picture interpretation approach.

Figure 3.2: Edge To Face Snippet for synthesizing Video to Video

A novel technique has proposed by the specialists from MIT's Computer Science and Artificial Intelligence Lab for video-to-video amalgamation, demonstrating noteworthy outcomes. The proposed technique video to video can integrate photorealistic, high-goal, transiently reasonable recordings on an assorted arrangement of info designs including division covers, outlines, and stances.

3.1.5.2 GAN Based Synthesis of Facial Texture & Surface

In any case, it is likewise notable that to prepare a truly mind boggling model, we'll need loads of information, which intently approximates the total information dispersion. Profound

organizations can be amazingly ground-breaking and viable in addressing complex inquiries. With the absence of true information, numerous scientists pick information increase as a technique for broadening the size of a given data set where he thought is to change the preparation models so that keeps their semantic properties flawless. That is not a simple undertaking when managing human mouth or facial appearance.

However, there some preprocesses to complete the mentioned strategies to synthesize and land marking image/video. Hence, the image data processing pipeline is dictated as below.

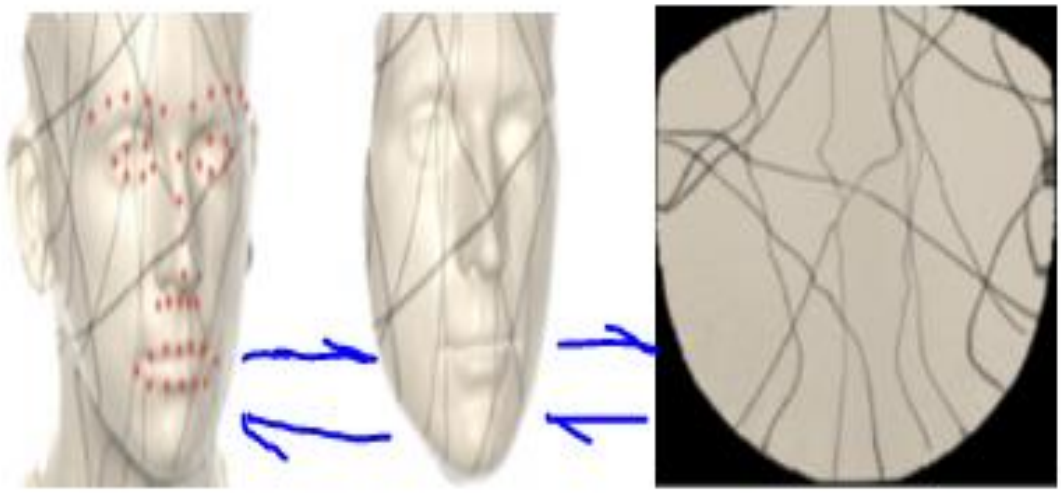


Figure 3.3 Facial Land marking Key points from Pipelining of Image Data Processing

Applying a pre-defined universal transformation this procedure contains the scans of vertex to vertex three Dimensional (3D) faces of human for aligning and textures mapping to two dimensional (2D) plane.

By the way, the main four (4) steps of image data preparation are precisely describe step by step below:

I. Collection of Data:

From a wide variety of age, gender, ethnic, and groups around 5000 scans were collected by the researchers whereas each subject was approached to perform 5 unmistakable articulations including an impartial one.

II. Allusion of Landmark:

By delivering the face and utilizing a pre-prepared facial milestone identifier, 43 landmarks were more appended to the cross segments semi-subsequently on the Two Dimensional pictures.

III. Performing Mesh Alignment:

As indicated by the mathematical structure of each scan the Mesh arrangement distorting a format face network was directed that was guided by the recently gotten facial landmark focuses.

IV. By transferring texture to template:

There is a technique called ray casting, which is built into the movement delivering tool stash of Blender. Afterwards, the textures and surface are moved from the output to the format of the template.

The following stage is to prepare Generative Adversarial Network (GAN) to learn and impersonate these adjusted facial surfaces. For this reason, the specialists utilize a reformist developing GAN with the generator and discriminator built as symmetric organizations. In this execution, the generator logically builds the goal of the component maps until arriving at the yield picture size, while the discriminator continuously lessens the size back to a solitary yield.

The last advance is to incorporate the calculations of the appearances. The specialists investigated a few ways to deal with finding conceivable math coefficients for a given surface. You can notice the subjective and quantitative [L2 geometric error] correlation between the different techniques.

3.1.5.3 Background Research of GAN

Our deep learning approach to learn Lip-sync from audio is a continuous process to implement the system with some Neural Network Learning and application of its classification with some effective model training and testing strategies. By the way, Generative Adversarial Network is also

one of them. Generative adversarial networks are the energizing ongoing development in AI. Since they are generative models, they make new information cases that take after our preparation data. For instance, GANs can generate pictures that resemble photos of human appearances, despite the fact that the countenances don't have a place with any real people.

Hence, firstly let's introduce with GAN and we should know what actually it is. Generative Adversarial Network (GAN) is such kind of generative model which is constitutes of 2 models named Generator, G & Discriminator, D. In the first one Generative model- the data distribution's random noise vector, z is captured and new data instances can be generated by Generative models whereas the Discriminator, D- separates between the arbitrary commotion vector and created information from the generator.

GAN takes in the mapping capacity from the generator over the noticed information x from the commotion appropriation pz(z) to information space as G(z;θg). Then again, the discriminator, D(x;θd), gauges the probability that the example is gotten from the preparation information in lieu of Generator. The Generator (G) and the Discriminator (D) are prepared at a time in the two-player minimax defined GAN. Generator and discriminator fakes to one another where boundary G limits log (D(G(z))) to fit the produced an incentive into D which gives the likelihood whether it is genuine or phony and D limits logD(X) that advances the probability of genuine data. So, the minimax game based GAN functions is V (D, G):

$$\min \max V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

Moreover, here we'll trying to express Wasserstein GAN (WGAN), Conditional Generative Adversarial Network (cGAN) and Deep Convolutional Generative Adversarial Network (DCGAN) for our Lip-Sync implementation purposes with LipGAN.

Generative models: p(X, Y) [The Joint Probability]; otherwise p(X) [For no labels].

The segment of the GAN that prepares the generator containing generator network to change the arbitrary contribution to the random data, discriminator network for grouping the produced data, output of the discriminator, generator loss that penalizes the generator for neglecting to trick the discriminator.

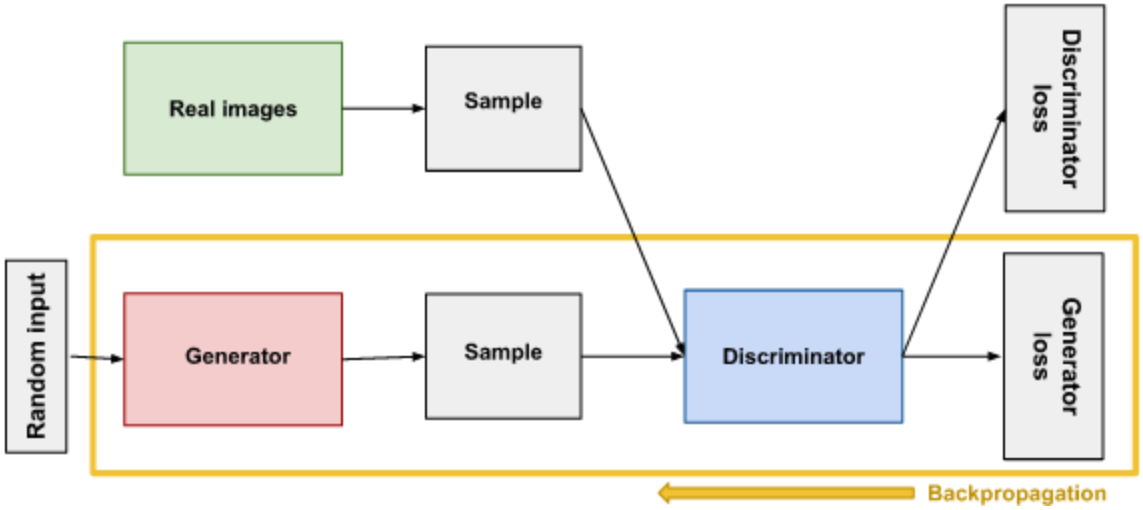


Figure 3.4: The Back propagation in training of discriminator.

The generator figures out how to create conceivable information. The created occasions become negative preparing models for the discriminator. For example, this model attempts to deliver persuading 1's and 0's by creating digits that fall near their genuine partners in the information space. It needs to show the appropriation all through the information space.

Discriminative models: Probability p(Y | X) [Conditional Probability]

It distinguishes the fake VS real data and also penalizes that generator for creating unrealistic outcomes.

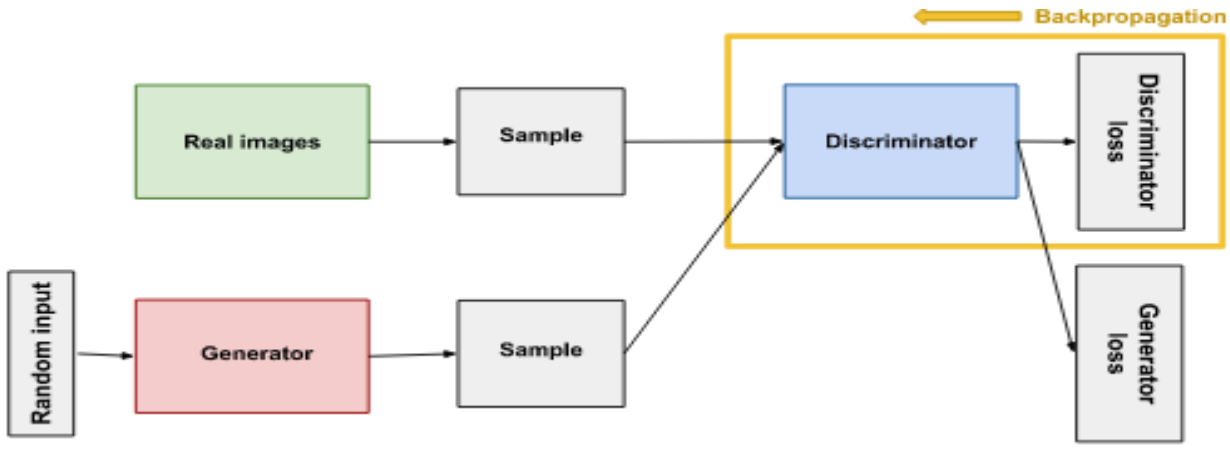


Figure 3.5: The Back propagation in Training of Generator.

The Discriminator: It is also a classifier to distinguish image data generated by the Generator.

Now as a whole system of the model is demonstrated a sbelow: **The Entire Framework for Both Model**

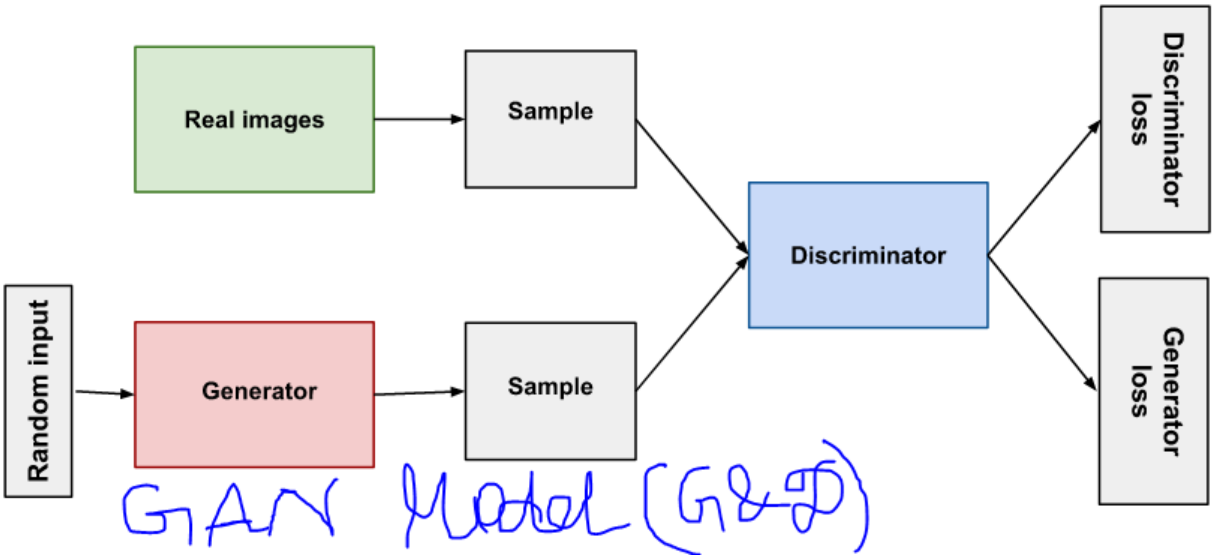


Figure 4: The Entire Framework for Both Model

3.1.5.3 The Concept of LipGan

All the models of speech recognition typically don't perform well in loud conditions. To help tackle the issue, specialists from Samsung & Imperial College in London built up a profound learning arrangement that utilizes computer vision acknowledgment. The model is fit for lip-reading, just as orchestrating sound it sees from the video.

Lip-reading is fundamentally utilized by individuals who are hearing impaired or can't hear properly or nearly deaf. Nonetheless, there are different applications in which it very well may be used, for example, video conferencing in loud or quiet conditions. GAN is fit for delivering characteristic sounding, understandable discourse which is synchronized with the video. IT maps video straightforwardly to crude sound and the first to create understandable discourse when tried on already concealed speakers. The model is comprised of three sub-networks including a

generator, which changes the video outlines into waveform, a 3D CNN that creates discourse like characteristic discourse, and a discourse encoder to preserve the discourse substance of the waveform. The model depends on the Wasserstein GAN, created by Facebook AI analysts, which limits the distances between the genuine and unreal conveyance.

3.1.6.1 Utilization of Lip-Sync in the Wild as Legal aspects

In Dubbing Language: In English-talking nations, numerous unfamiliar TV arrangement particularly anime, several serial dramas are dubbed for transmission. In any case, realistic arrivals of movies will in general accompany subtitle all things considered. The equivalent is valid for nations in which the neighborhood language isn't spoken broadly enough to make the costly naming industrially reasonable that means at the end of the day, there isn't sufficient market for it. However, different nations with a huge enough populace name all unfamiliar movies into their public language artistic delivery. Naming is favored by some since it permits the watcher to zero in on the on-screen activity, without perusing the caption.

Quality film naming necessitates that the discourse is first interpreted so that the words utilized can coordinate the lip landmarks developments of the entertainer. This is regularly difficult to accomplish if the interpretation is to remain consistent with the first exchange. Expound lip-synchronize of naming is likewise an extensive and costly cycle. The more improved non-phonetic portrayal of mouth development in numerous anime helps this cycle.

A sample example mentioned on the above like Video Conferencing can be briefly discuss and demonstrate basic practical procedure to accomplish lip sync on Video conferencing endpoints for the most part adopt 2 strategies are as followings:

1. The Lip-Sync of Helpless Man: This technique expects that delays eventually to-end media ways are known and consistent. It depends on bundle appearance times for synchronization.

2. The Lip-Sync of Normal Reference: This technique accepts that delays eventually to-end media ways are not handily known and may differ. It depends on a typical reference time base for both sound and video transfers.

However, the below sender side and receiver side signal transmitting and receiving understanding strategies will clarify the concept of video conferencing strategies to relate it with Lip-Sync:

From the Sender Side of video conferencing endpoint:

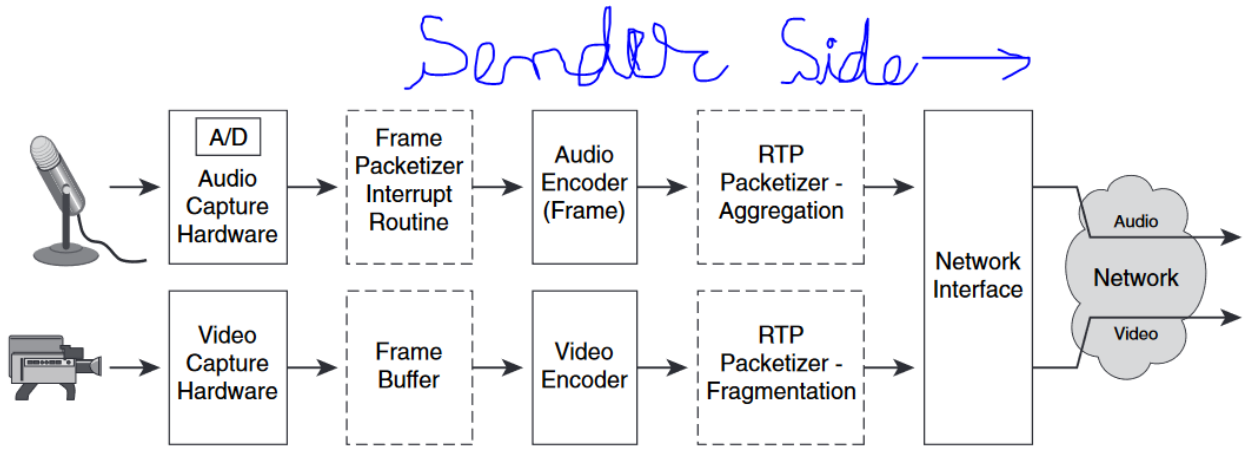


Figure 3.6: Sender Side's Signal Processing Strategies

To the Receiver Side of video conferencing endpoint:

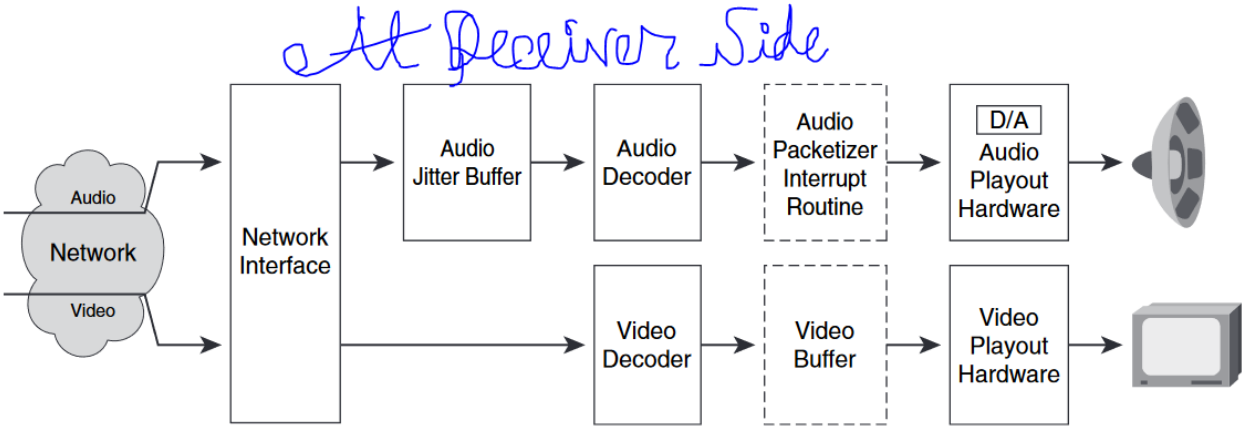


Figure 3.7: Receiver Side's Signal Processing Strategies

Lip-Sync In The Video: Lip-synchronizing is the strategy utilized when vivified characters talk, and lip synchronizing is fundamental when movies are named into different dialects. In film/movie creation, lip synchronizing is regularly essential for the after creation stage. Most film today contains scenes where the discourse has been re-recorded thereafter; In numerous melodic movies, entertainers sang their own tunes previously in a chronicle meeting and lip-synchronized during recording, however numerous additionally lip-synchronized to voices other than their own.

Lip-Sync In The Gaming Industries: On account of the measure of data passed on through the game, most of correspondence employments of looking over content. More established 'RPGs' depend exclusively on content, utilizing lifeless pictures to give a feeling of who is talking.

A few games utilize voice acting, for example, 'Diablo or Grandia II', yet because of basic character models, there is no mouth development to reproduce discourse. The RPG for hand-kept frameworks are still generally dependent on content, with the uncommon utilization of lip sync and voice records being held for full movement video cut-scenes. 'The Never winter Nights' arrangement are instances of momentary games where significant discourse and cut scenes are completely voiced, yet less significant data is as yet passed on in content. In games, for example, 'Jade Empire and Knights of the Old Republic', engineers made fractional counterfeit dialects to give the impression of full voice acting without having to really voice all discourse. Especially, in Gaming industries The Lip-Sync technology reduces the cost and save times and money along with increasing the popularity and AI practices with pure design creativity by the developers.

Lip-Sync In Animation: The incredible application of Lip-Sync is in the Animation video and to CGI character's Talking. Another sign of lip synchronizing is the specialty of causing an energized character to seem to talk in a prerecorded track of discourse. The lip sync method to cause the character as animated to seem to talk includes sorting out the timings of the discourse just as the real enlivening of the lips to coordinate the speech exchange track. The most punctual

instances of lip-sync in activity were endeavored by Max Fleischer in his Nineteenth Twenty Sixth short ‘My Old Kentucky Home’. The strategy proceeds right up 'til today, with energized movies and network shows, for example, ‘Shrek’, ‘Lilo’ and ‘Stitch’, and ‘The Simpsons’ utilizing lip-synchronizing to make their fake characters talk. Lip synchronizing is likewise utilized in comedies, for example, This Hour Has twenty two minutes and political parody, changing absolutely or just incompletely the first phrasing. It has been utilized related to interpretation of movies starting with one language then onto the next, for instance, Spirited Away. Lip-synchronizing can be an exceptionally troublesome issue in making an interpretation of unfamiliar works to a homegrown delivery, as a straightforward interpretation of the lines regularly leaves invade or underrun of high discourse to mouth developments.

3.2 Mapping Audio Using Recurrent Neural Network to Video

If we want to map audio to target video then firstly we have to determine the features of audio with mouth shape which vary time to time of speaker speaking. With the memory mechanism facilities of RNN, modifying hidden state of to solve time series issues are more popular for aspect of non-linear transitions. In some cases, assume one’s mouth moves before he/she states something like saying hurrahhh or yeahh then his mouth is as of now open. Henceforth, it's adequately not to condition his/her mouth shape on past sound info –the organization needs to investigate what's to come. So, there are several works point this out as a restriction of their LSTM strategy. One potential arrangement is to make the organization bidirectional. In fact [22] utilizations a bidirectional LSTM to misuse future setting. Bidirectional networks require significantly more figure force and memory to prepare, as they should be unfurled in the back propagation cycle, which typically restricts the length of preparing models, yet in addition the length of the yield. All things considered, a lot more straightforward approach to present a short future setting to a unidirectional organization is to add a period postponement to the yield by shifting the organization yield forward in time as investigated in the past as target delay. While bidirectional LSTMs are famous for recognition of speech issues [23], we notice that the less difficult time delay instrument is adequate for our assignment, likely because of the need to look less far later on for sound to video, contrasted and discourse acknowledgment which may require looking numerous words

ahead [5]. A period deferred RNN delay, d=2 resembles the followings:

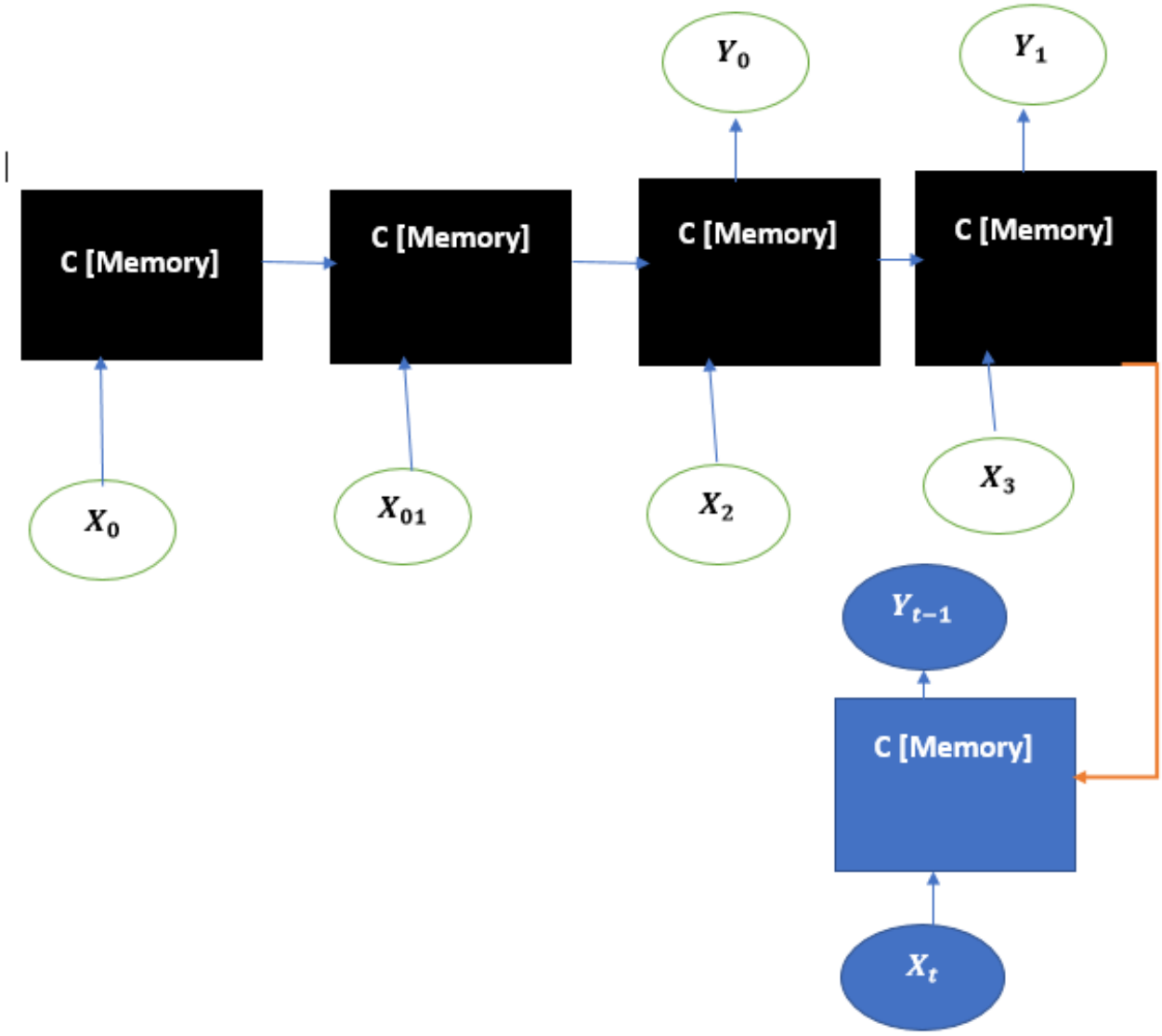


Figure 3.8: A time delayed (for d=2) The RNN Architecture

For more clarification it's mentioned that if input as [x1....xn], and the sequences of output vector, is [y1.....yn] the standard LSTM network is defined by the following functions:

$f_t = \sigma\,(W_f \cdot [h_{t-1}\,,\ X_t] + b_f)$ (1)

$i_t = \sigma\,(W_i \cdot [h_{t-1}\,,\ X_t] + b_i)$ (2)

$O_t = \sigma\,(W_0 \cdot [h_{t-1}\ ,\ X_t] + b_0)$ (3)

$c_t = c_{t-1} f_t + i_t \tanh\,(W_j\ .\ [h_{t-1}\ ,\ X_t] + b_j)$ (4)

$h_t = \tanh(c_t)\ O_t$ (5)

$\hat{y}_{t-d} = W_y h_t + b_y$ (6)

Here, f ; i; o; c; h are forget gate, input gate, output gate, cell state, cell output as introduced in [24, 5] σ is the sigmoid activation. Note that the cell state and cell output are transformed with tanh. (yˆt−d) represents the predicted PCA coefficients at time t − d (d-time-delay parameter. Learned parameters are weight matrices W and bias vectors b. [5]used a 60dimensional cell state c and a time delay d of 20 steps (200ms).

By utilizing L2-loss on the coefficients the network is minimized and trained using Adam optimizer [25]executed in TensorFlow [5, 26], on numerous hours of stock Obama weekly address video footage.

3.3 Computation of Reconstruction loss at the level of Pixel

3.3.1 What's Reconstruction loss?

Reconstruction loss expresses that the cross-entropy or the MSE (mean-squared error) between the output & input. It penalizes the network for generating outputs which are varies from the input. Reconstruction error is the separation between the first input and its auto encoder reproduction. Auto encoders compress the input into a lower-dimensional projection and after that remake the yield from this representation.

For de noising data like audio, images and detecting anomaly, image in painting and information retrieval the Auto encoders help in this matter as well. Auto encoders are a solo learning procedure wherein we influence neural organizations for the assignment of portrayal learning. In particular, we'll plan a neural organization

design with the end goal that we force a bottleneck in the organization which powers a packed information portrayal of the first info. On the off chance that the information highlights were every autonomous of each other, this pressure and ensuing recreation would be a troublesome errand. Nonetheless, if some kind of structure exists in the information this structure can be educated and therefore utilized while compelling the contribution through the organization's bottleneck.

Figure 3.9 3D faces reconstructed sample's expressive figure with different methods

Below is the visual graph demonstration of Rank (X-axis) VS Identification Rate(Y-axis) for recontructed 3d images.

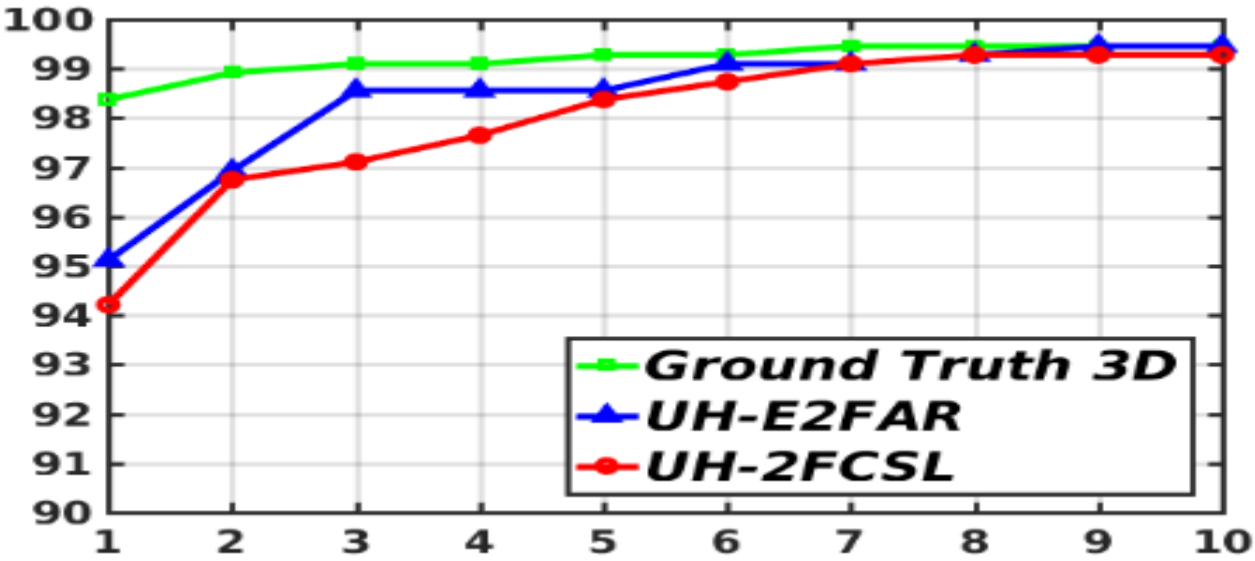


Figure 3.10 Rank VS Identification Rate Faceial Expressive Graph

To confirm the accurate pose creation, face's background, identity preservation in generally for the entire image the reconstruction loss is computed where the lip portion conrrespondingly to samller than in the percantage of four of the entire loss of reconstruction.. Hence, before starting of any network reconstruction images are optimized firstly for correcting the lip shape as smooth and sharpenning. This is additionally upheld by the way that the organization starts transforming lips just at around midway through its preparation measure. Subsequently, it is critical to have an extradiscriminator to pass judgment on lip-sync, as additionally done in LipGAN [18].Hoewever, we should also think about the discriminator whether it performing weel or as good enough in LipGAN.

3.3.2 Our Retrained expert Discriminator for Everything

It contains a face encoder and a sound encoder, including a heap of 2D-convolutions. L2 distance is registered between the embedding's produced from these encoders, and the model is prepared with a maximum edge misfortune to limit the distance between adjusted or un-synced sets.3.3.2

our master lip-sync discriminator. We make the accompanying changes to prepare a specialist lip-sync discriminator that suits our lip age task. Initially, rather than taking care of grayscale pictures linked divert savvy as in the first model, we feed shading pictures. Also, our model is altogether more profound. Thirdly, we utilize an alternate misfortune work: cosine-closeness with two fold cross-entropy error. That is, we register a spot item between the ReLU-initiated video and discourse embedding v, s to output a solitary incentive between several works for each example that demonstrates the likelihood that the info sound video pair is in a state of harmony.

So, Psync = $(v \cdot s)$ / (($\|v\|2 \cdot \|s\|2, \epsilon$). Our system return accurate output which is more than 90% with LRS2 test set and it's more better against the LipGAN since the LipGan returns the accuracy around 50%-60% for same data.

CHAPTER 4
EVALUATION OF LIP-SYNCING

The current assessment structure for speaker-free lip-syncing makes a decision about the models uniquely in contrast to how it is utilized while lip-synchronizing a genuine video. In particular, rather than taking care of the current outline as a source of perspective (as portrayed in the past area), an irregular casing in the video is picked as the reference to not release the right lip data during assessment. We firmly contend that the assessment structure in the past passage isn't ideal for assessing the lip-sync quality and exactness. Upon a closer assessment of the previously mentioned assessment framework, we noticed a couple of key constraints, which we talk about in the followings way.

- Doesn't mirror this present reality utilization. As examined previously, during age at test time, the model should not change the posture, as the produced face should be consistently stuck into the edge. Nonetheless, the current assessment structure takes care of irregular reference outlines in the info, along these lines requesting the organization to change the present. Along these lines, the above framework doesn't assess how the model would be utilized in reality

- Doesn't uphold checking for worldly consistency. As the reference outlines are haphazardly picked at each time-step, fleeting consistency is as of now lost as the casings are created aimlessly postures and scales. The current structure can't uphold another measurement or a future technique that intends to consider the fleeting consistency part of this issue.
- Conflicting evaluation returning. As the reference outlines are picked at irregular, this implies the test information isn't predictable across various works. This would prompt an unreasonable examination and ruin the reproducibility of results.
- Besides the mentioned aspects of evaluation for lip-syncing sometimes the metrics may not specific. Some metrics were promoted to judge the smoothness or quality of images but there was some error during lip-sync as a result the necessity of such kind of metrics which can determines the error for better quality and accurate face and landmark generation.

  And it’s our pleasure that we introduced and proposed on our system a metric & new benchmark to overcome the mentioned issues and also able to return better and accurate output compared to the other works.

Table 4.1: Audio Encoder Block Table with torch Size Within different Network Layer

Block	Layer (type)	Output Shape	Params #

audio_encoder	Conv2d	torch.Size([32, 1, 3, 3])	384
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([32, 32, 3, 3])	9312
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([32, 32, 3, 3])	9312
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([64, 32, 3, 3])	18624
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	
	Conv2d	torch.Size([64, 64, 3, 3])	37056
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	
	Conv2d	torch.Size([64, 64, 3, 3])	37056
	BatchNorm2d	torch.Size([64])	
	ReLU	torch.Size([64])	
	Conv2d	torch.Size([128, 64, 3, 3])	74112
	BatchNorm2d	torch.Size([128])	
	ReLU	torch.Size([128])	
	Conv2d	torch.Size([128, 128, 3, 3])	147840
	BatchNorm2d	torch.Size([128])	
	ReLU	torch.Size([128])	
	Conv2d	torch.Size([128, 128, 3, 3])	147840
	BatchNorm2d	torch.Size([128])	
	ReLU	torch.Size([128])	
	Conv2d	torch.Size([256, 128, 3, 3])	295680
	BatchNorm2d	torch.Size([256])	
	ReLU	torch.Size([256])	
	Conv2d	torch.Size([256, 256, 3, 3])	590592

Table 4.1.2: Output Block Table with torch Size Within different Network Layer.

output_block	Conv2d	torch.Size([32, 80, 3, 3])	23136
	BatchNorm2d	torch.Size([32])	
	ReLU	torch.Size([32])	
	Conv2d	torch.Size([3, 32, 1, 1])	99
	Sigmoid	torch.Size([3])	0

CHAPTER 5
LEGAL USES AND APPLICATIONS

Day by day the use of audio-video content is becoming increasing [27] which has become clearly apparent due to the coronavirus. Therefore long-length video generation, formation and translation have become highly needed. The videos of various online courses are usually in English language. It has become very important to spread the video all over the world with access to all languages but it is very complicated and time consuming to spread it in different languages. So, this has become a crucial challenge for us during this coronavirus epidemic situation. Our model can generate and translate automatically with accurate lip-synchronization into any language using our model. Watching movies with subtitles is very unpleasant, it can make accurate dubbing movies to watch it pleasantly.

Nowadays advanced AI technology makes video games more realistic and this has been made possible due to the effective use of lip synchronization. Our model can adjust any voice-over with accurate lip-sync that makes a huge impact in modern video game industry. Alike we can make cartoon characters also become more realistic by it. Live telecast is becoming very popular all over the world day by day. Live telecasts and video streaming, including public speaking and public conferences, are also on the rise. But sometimes it is seen that speaker's lips are going out with the translated speech thus which makes viewers unpleasant. Our model can solve this problem very easily accurately with the help of CGI (Computer Generated Imaginary) character. Speech can be translated CGI character automatically which can save many hours by automating manual effort and making it easier to understand.

We think it is our responsibility to discuss and promote fair use otherwise misuse of our work may cause serious destruction to any person or nation. The increasing capable of lip-sync works and advanced lip synchronization techniques change any audio-video contents that may be used in negatively. We strongly encourage to use our model in positive applications and we believe our model will contribute in future works to make better lip-sync in audio-visual contents.

CHAPTER 6

DISCUSSION & FUTURE WORK

We have seen that we get accurate lip synchronization while training by large scale video with sample audio but it was not possible to do perfect mouth synchronization in cases where teeth are seen due to audio which has appeared to us as a major problem. We think the target output would be perfectly realistic if there was proper synchronization of the teeth with lips of the target video. So, we think we have to do a lot of work in the future for realistic synchronization of the lip with the teeth.

Our model is not able to understand the transliteration of audio because it cannot read what the audio means cannot actually explain to itself and the meaning and significance of the audio. That's why the audio didn't seem to make the right expression visible on the face of the target video. We think there is a need for significant improvement in this area as well as the fact that the model is not capable properly of eye gesture observation.

CHAPTER 7
CONCLUSION

In this paper, we are mainly showing a new advanced technology that can generate accurate lip-synchronization from any given input audio with any sample videos. We can strongly say that our model is significantly better and advanced ability to match lip between input audio and target video comparing with others current works. When we worked on literature review, we tested and evaluated the current framework on lip-sync, we have found a lot of issues that we thought needed to be improved to make a realistic lip-sync. To solve these issues, we have created neural networks and frameworks with the help of efficient algorithms, which have tested and evaluated our framework through various sample of videos.

The work that is currently being done to make a lip movement is in the static image or video of a specific person and it is in the trial phase. Moreover, they failed to accurately determine the movement of the lips in dynamic arbitrary identifies, unconstrained talking face videos that making

most of the significant portion of the targeted videos out of synchronization with the sample audio. We identify these issues and solve these issues by our significant model. A demo video of lip synchronization is provided on our website. We encourage readers to get a clearer idea by watching the video of our works.

Although our model is mainly based on accurate lip synchronization, we think this model will open up new directions for future and making more advanced face syntheses

References

[1] A. Jamaludin, J. S. Chung and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision,* vol. 127, no. 11-12, pp. 1767-1779, 2019.

[2] Y. Chen, W. Gao, Z. Wang, J. Miao and D. Jiang, "Mining audio/visual database for speech driven face animation," in *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, 2001.

[3] S. Ogata, K. Murai, N. Satoshi and S. Morishima, "Model-based lip synchronization with automatically translated synthetic voice toward a multi-modal translation system.," in *IEEE International Conference on Multimedia and Expo*, 2001.

[4] T. Karras, T. Aila, S. Laine, A. Herva and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG),* vol. 36, no. 4, pp. 1-12, 2017.

[5] S. Suwajanakorn, S. M. Seitz and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG),* vol. 36, no. 4, pp. 1-13, 2017.

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

[7] S. Suwajanakorn, S. M. Seitz and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Transactions on Graphics (TOG),* vol. 36, no. 4, p. 95, 2017.

[8] 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),* vol. 38, no. 4, pp. 1-14, 2019.

[9] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri and C. and Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[10] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

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

[12] L. Yu, J. Yu and a. Q. Ling, "Mining audio, text and visual information for talking face generation," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019.

[13] J. Thies, M. Elgharib, A. Tewari, C. Theobalt and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European Conference on Computer Vision*, 2020.

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

[15] L. u, J. Yu, M. Li and a. Q. Ling, "Multimodal Inputs Driven Talking Face Generation With Spatial-Temporal Dependency," *IEEE Transactions on Circuits and Systems for Video Technology,* 2020.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, J. K. Andrew Tao and a. B. Catanzaro, "Video-to-video synthesis," *rXiv preprint arXiv:1808.06601,* 2018.

[17] S. Tulyakov, M.-Y. Liu, X. Yang and a. J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018 .

[18] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri and C. and Jawahar, "Towards automatic face-to-face translation," in *In Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[19] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research,* vol. 10, pp. 1755--1758, 2009.

[20] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 6, pp. 679--698, 1986.

[21] S. Dawson, "connectedmag," 2014 . [Online]. Available: https://connectedmag.com.au/importance-lip-sync/ . [Accessed 13 08 2020].

[22] B. Fan, L. Xie, S. Yang, L. Wang and F. K. and Soong, "A deep bidirectional LSTM approach for video-realistic talking head," *Multimedia Tools and Applications,* vol. 75, no. 9, pp. 5287--5309, 2016.

[23] A. Graves and J. and Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks,* vol. 18, no. 5-6, p. 602–610, 2005.

[24] S. Hochreiter and J. and Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

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

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin and a. others, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467,* 2016.

[27] NPD, "NPD," 29 3 2016. [Online]. Available: https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-usetheir-device-for-video-calling-according-to-the-npd-group/. [Accessed 1 12 2020].

[28] [Online].