Deepfake Creation Using Generative Adversarial Network

Hrutuja Satpute[1], Samruddhi Raut[2], Pragati Sapke[3], Dr. Mrudul Dixit[4]

[1,2,3,4]*Department of Electronics and Telecommunications Engineering*

[1,2,3,4]*MKSSS's Cummins College of Engineering for Women, Pune, India*

1,2,3 - Students 4 - Assistant Professor and Dean Alumni

Abstract - ***The usage of innovative and dynamic technology is gaining prominence in the innovative and entertaining sectors as well as the significance of storytelling expands in the digital-first strategy. Deep fake is one of these breakthroughs. Deep Fakes are images, movies, and audios that are made using deep learning techniques which seem genuine to the people around. Animation is a method in which motions are moved around in order to create moving visuals. For this methods are to be designed to create such an animation. Therefore, this paper gives a brief description for generating deep fakes which helps a still image to enact like the driving video. The network generates a video which does not exist in reality. It takes the motions of the target actor from the driving video and transfers it to the input image.The realism of this transition is achieved only through adversarial training, which results in the change of the target films that recreate the behavior of the artificially produced input. First order motion model is a model which is used for facial reenactment.This model consists of generator and motion extractor. A keypoint detector plays a vital role. Keypoint locates points on the body of the image. It locates the parts which are moving. The ability to recombine input image and input audio parameters can be utilized to reproduce the entire head through user-controlled editing, resulting in high visual dubbing. The basic goal of the model and visual dubbing is to synchronize the movements of the lips with the input as an audio which can be in any language for any moving video. It's made up of video and spoken elements. Talking head films are created using an image of a person plus an input audio having any dialogue.Eventually, the output will include lip movements as well as audio. Similarly, in this digital world there is a great demand of directly translating a video of a person from one language to the other including coordination of the lips. As a result, a pipeline is created in which the individual speaking Hindi in the input video will speak English with the output video.***

Keywords: **DeepFakes, First order motion model, Facial reenactment, GAN, Visual dubbing,Wav2Lip**

I. Introduction

In this era,the world is getting digital day by day. The areas like Deep learning and Artificial Intelligence have made remarkable progress and because of which creation of deep fakes is now possible. Video generation and image creation has increased demand in the world. Deep fakes are nothing but creation of a video which does not exist in reality. The main aim of deep fakes is to generate the images which look as realistic as possible. Deep fake,basically takes the features of the input video and transfers it to the source image. Motion extraction is done from the driving video and the motions extracted are transferred to the still image in order to make it live.

To make the image perform as similar to the driving video, the approach used is (GAN) General Adversarial Network. The GANs consist of the generator and the discriminator. The job of the generator is to generate fake images or images which do not exist in real life. The discriminator is used to distinguish whether the image is real or fake. Generator has one input which is an input noise vector. Discriminator consists of two inputs i.e the real image and the image generated from the generator. Both the inputs are given to the discriminator and then the discriminator classifies it as a real or fake. Generator maximizes the classification error whereas discriminator minimizes the error. GANs are very well known because they try to produce images that look like a xerox copy of the existing image. Making static photos to life has a variety of applications in teaching, filmmaking, and photography. It not only adds to the dynamic environment, but it also improves the user experience.

The term communication refers to the act of passing information from one person or group of people to another. People have long aimed to communicate across the globe in a variety of languages. People can better grasp the piece of note if the wording or voice is translated into multiple languages. The advancement of such technology has been remarkable. In terms of real life examples, dubbing of the recent film Pushpa-The Rise allowed people to see and appreciate the film in Hindi, despite the fact that it was originally released in telugu.

So, the input side consists of audio and video. The audio is in English language and the video is in Hindi language. The main job of the method is to convert the Hindi language of the video to English.Wav2Lip is a Generative Adversarial Network that generates a talking face in any language. This technique allows you to modify a person's lip movements in a video to match an audio sample of a target individual. Wav2Lip technology was utilized to offer a model for generating true speaking faces that can react to audio in any language.

II. Literature Survey

[1]"Deep Video Portraits" describes a method for combining input's real video portraits in front of immovable backdrops.A new translation network is used in this technique to convert a pattern of simple computer graphics representations into photo realistic video. Many applications, such as video reenactment for virtual and augmented reality and telerobotics, interactive video checking, and audio-visual dubbing, will benefit from it. This is a completely new approach that can be considered as a step toward a more realistic world of entire frame video format synthesis under the direction of relevant requirements.

[2]"DL for Deep Fakes Creation and Detection" is a research of deep fake generation algorithms. It reviews the background of deepfakes and state-of-the-art deepfake generation and discrimination methods before presenting extensive inventions on difficulties, research trends, and assumptions linked to deepfake technology. This research provides a thorough overview of deepfake approaches and facilitates the development of new, more robust strategies for dealing with the more difficult deep fakes.

[3] The Cycle-GAN network, which is a representation of two GAN networks, is used in "Deep Fakes Using GAN." Cycle-GAN is used to achieve two goals: object transformation and style transfer. Images of handbags and backpacks are mutually translated in the former, and photo conversions to other art types are mutually translated in the latter. The outputs of created photos using the PyTorch framework are relatively satisfying.

[4] "Deep Fake Creation using Deep Learning" explains how to make deep fakes using autoencoders. The autoencoder releases inert highlights from facial images, while the decoder reconstructs them. Between input image and target image, there is a requirement for two encoder and decoder sets where the encoders specifications are used between two system sets, where each pair is used to make a picture set.The software makes use of Google's TensorFlow AI Framework, which was previously used for the DeepDream programme, among other things.

[5] "Model used for adapting motions of the video is first order motion model" uses keypoints and local affine adjustments to demonstrate an innovative way to visual animation.The first order Taylor expansion approximation is used to efficiently construct the mathematics, which in brief describes the motion between two frames. The motion in the driver video which is an input video is stated by a blend of keypoints and local affine. The driver video consists of the motions and those motions are pulled out and transferred to the static image. All these things happen in the generator network. The responsibility of the generator is to generate fake images. Furthermore, it's a good idea to model occlusions explicitly so that the generator network knows which image regions to paint.

[6] Wav2Lip: A Lip Sync Expert Lip synchronization of any talking person irrespective of its identity to match a target speech segment was investigated in this paper. Major work increases at generating proper lip variation on a still image of a particular person or people. However, due to some reason failure rises to perfectly transform the lip movements of the identities which are not known, resulting in major chunks of the video being out of sync with the new audio.This study finds an important reason behind this and, as a result, resolves them by using a capable lip-sync discriminator. It is proposed that a new rigorous evaluation benchmark and criteria be developed to reliably measure lip-sync in fast videos. Large evaluations on difficult criteria reveal that the lip-sync accuracy of our Wav2Lip model-generated videos is nearly as excellent as real synchronized videos.

[7] Wav2Lip+GAN: Towards Automatic Translation focuses on audio and visual content.The input consists of a person who is speaking in one language and extends the challenge of automatic machine translation to a face translation. They developed a basic method for talking face generation in addition to proving the viability of a Face-to-Face translation pipeline.This paper also contributes to a number of language processing challenges for resource-constrained languages (such as textual machine translation). The face to face translation endeavor reveals a variety of new research areas like CV, multimedia processing, and ML. For example, when a speech is translated, the duration of the speech gets naturally changed .This necessitates a change in the associated motions, expressions, and background information.

III. Methodology

A deep fake video is a video of a person in which the face or body is to be digitally altered so that they look at someone else which does not exist in reality. Deep Fakes depend upon neural networks. Neural Networks can figure out huge sets of data samples so that it can learn to act or adapt facial expressions of a user. It can also adapt the eye movements, mouth and hand movements, etc. So, the method consists of an input image which is a still image and a driver video which consists of motions for example hand or mouth movements, eye movements. Taking the input image and the driver video and passing into a deep learning algorithm to train it. After this the motions of the driving video are extracted and transferred to the input image and the output will be a video which does not exist.

For transferring the motions or creating a deep fake video we require a methodology which includes GAN. General Adversarial Network plays a vital role in this technology.
GAN is nothing but a neural network which consists of a generator and a discriminator. The motto behind this is, to make sure that the deep fakes which are created should look real. Both the generator and discriminator fight for each other, both are competitive in nature.

Generative Adversarial Networks:

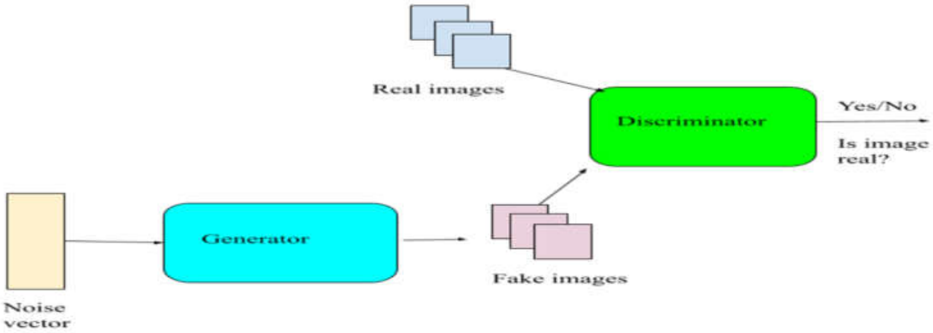


Figure 1. GAN model

The two main blocks of the above figure 1 are the generator and the discriminator. The generator generates images which are not real and the discriminator classifies whether the image generated is real or not.

The first step is to generate an image. So, the input noise is given to the generator. The generator will try to generate images i.e fake images. These fake images are given to the discriminator along with the real image data. Both will act as an input to the discriminator.The discriminator will classify real or not. The task of the generator is to maximize the error and to constantly fool the discriminator. The task of the discriminator is to minimize the error and to give the results in binary format.

This process continues until the discriminator identifies that the generated image from the generator is real.

First Order Motion Model
First Order Motion Model is a method where the motions of the driving video are transferred to the source image. Source image is nothing but a still image. The expressions of the source image are combined with the facial expressions of the driver video. First order motion model is a deep

learning model. It is also recognized as a State Of Art Model. This First Order Motion Model consists of Keypoints Detector, Local Affine Transformations, Dense Motion and a generator.

IV. Block Diagram

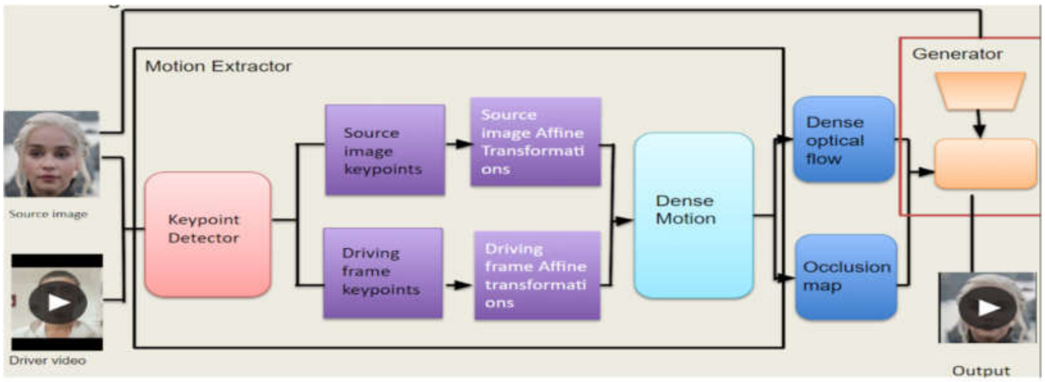


Figure 2. Block diagram

Motion extraction: Auto encoder technique is used by the Motion extractor to locate the key points. Keypoint detector **:** It uses an unsupervised approach. It extracts keypoints from input images as well as driving video. It also tries to locate the motions of the person in the driver video and then assign key points to them. These points are picture spatial points that grab everyone's attention. Even if the photo is twisted, translated, or lessened these points will remain unchanged.

So here, there is a need to support complex motion, which is why the network uses a depiction that consists of learned key points as well as local affine transformations used to alter the source image with respect to driver video.

From the above figure 2, First of all there will be an input image and a driving video. These will act as an input to the keypoint detector. It predicts key points from source images as well as driving video separately. Local affine transformations of the images and driver video are also produced. After this, the learned keypoints and local affine transformations are applied as an input to a dense motion network and it outputs a dense motion field from driving video to source image. In addition to that there is an occlusion mask which creates parts of objects that are not there in the input image. The dense motion network gives output as an occlusion map and dense optical flow which acts as an input to the generator network along with the source image and provides the required result.

Visual Dubbing:

For any actor in a video, visual dubbing focuses on matching the lip motions with any arbitrary audio as an input. Speech-driven face animation uses speech signals to create talking characters. It creates a link between audio and video elements.Using merely a still image of a person and an audio clip including some speech, this technique creates videos of talking heads.

The output video will have the lip movements that are in synchronization with the audio given.

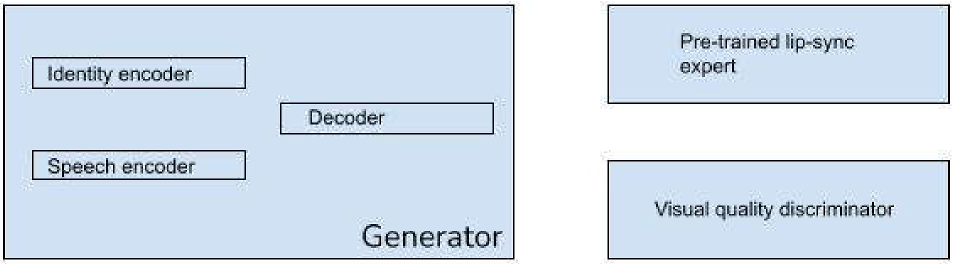


Figure 3. Block diagram

The figure 3 represents the diagram of the Wav2lip method.
A block consisting of a random segment of consecutive frames is used for speaker identification.
First block is the identity encoder. Its main purpose is encoding the random segments and ground truth frames.
Second block is the speech encoder. Its main purpose is encoding audio signals.
Third block is the face decoder. Its main purpose is decoding the combined similar vectors to a series of reconstructed video frames.

Features of the model

As we think about the face and its voice ,this model gives accurate flexibility. The best part of the model is that it can give the absolute best results with any face and can work very well with any audio or voice. Since it's a pre-trained discriminator, it gives about 91% accuracy.
Because it is trained on many videos in different languages, it has a good benefit to be fair towards any other new language.

Face-to-Face Translation

Most online courses i.e the coursera or udemy content is mostly available in English language. Due to language issues most of the people who are not familiar with this language can now see the content in Hindi because of this model. Most of the time the lip movements in the video are not in synchronization and thus it gives a very horrible experience. Therefore,this machine provides a method which takes a video of a person, let's say A, talking in a Hindi language and generates or outputs a video of the same person i.e person A talking in a target language i.e English.
The main motto is to produce a video with maximum synchronization in lip movement of any person talking in target language. To recognise head-to-head translation, the computer combines various modules from vision, speech, and language.

The network consists of five modules. Experimentation is performed on two local languages. The two local languages are English and Hindi.

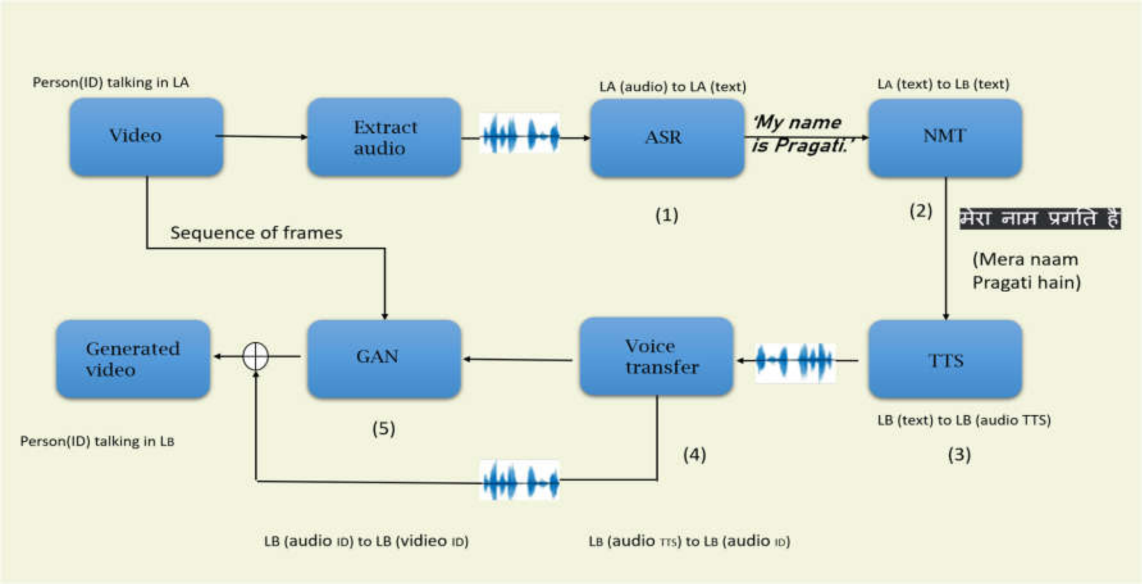


Figure 4. Block diagram of translation network

Input language is English, whereas output language should be Hindi.
There are some steps that needs to be followed:
1. Identification of speech which is given as the input.
2. Converting the identified text from input English language to Hindi language.
3. Synthesis of speech from the converted text.
4.After synthesis, talking faces in Hindi language are generated.

Automatic Speech Recognition: Uses automatic language recognition (ASR) to convert the language of the source language to the appropriate text. English speech recognition has been extensively studied due to the existence of big open source speech recognition datasets and trained models. In this study, DeepSpeech2 model is used to identify English speech.

NMT: Neural Machine Translation generates translated speech in the targeted language. This network gives better results and hence it is used.

Text to Speech: The system uses a sound wave generator to create audio. Finally, the frequency response of the phrase obtained from the acoustic model is loaded into the sound generator.

V. Result and analysis

Input image **Driving video** **Output video(Fake Video)**

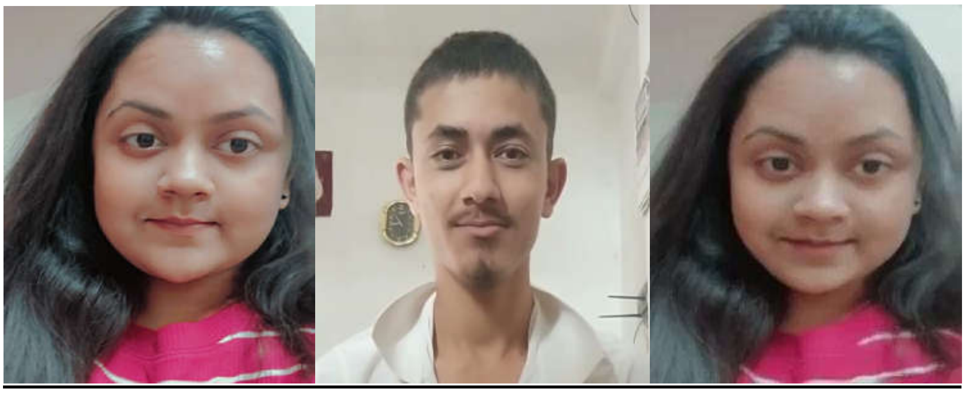


Figure 5.

The above figure 5 describes the input and output of the system.
The input is an image and driving video. The output is the generated video.

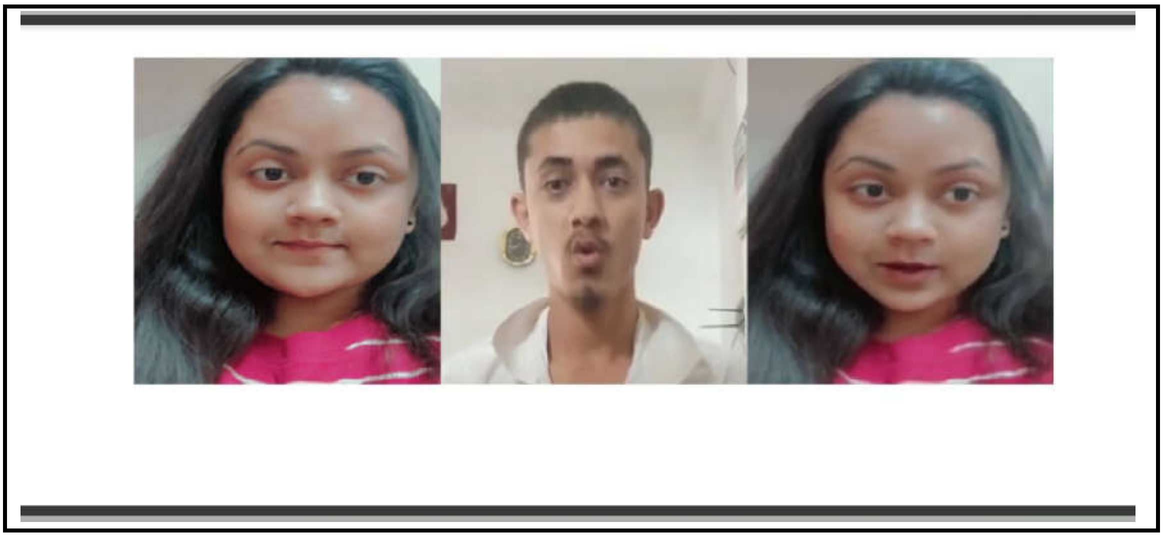


Figure 6.

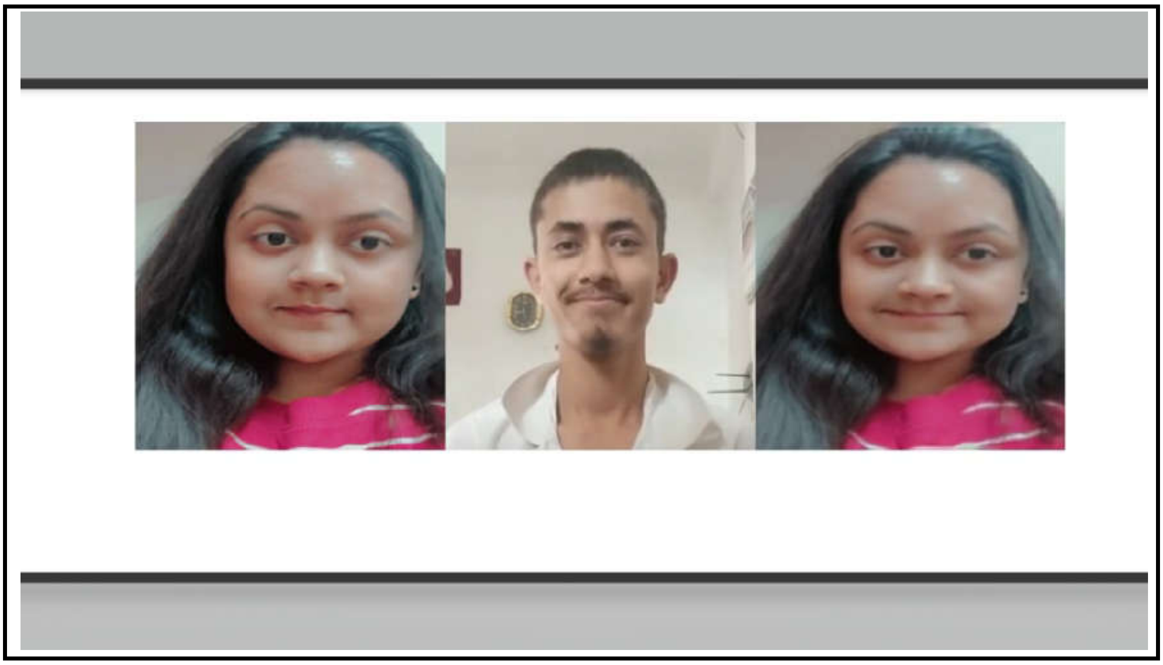


Figure 7.

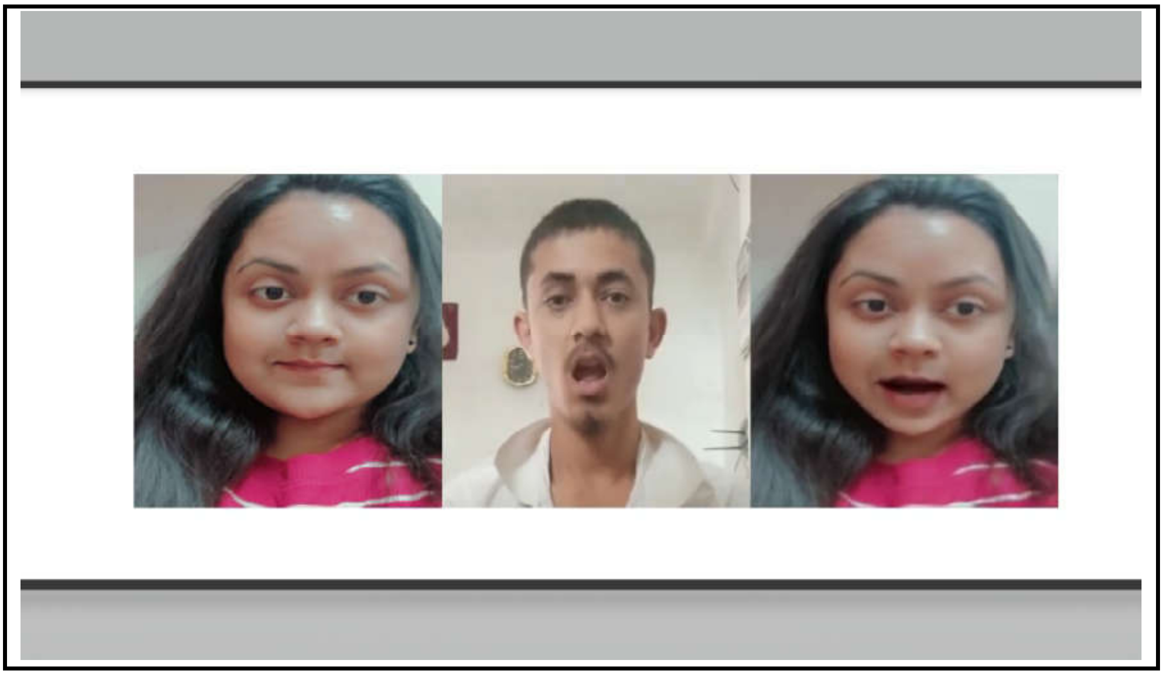


Figure 8.

The figure 6, figure 7 and figure 8 are the variations in the driving video which is adapted by the input image and thus the output is generated which is a fake one.

Parameters:

The performance is focused on the clarity of the output video.

After processing the quality of the generated output video seems to be decreased. The resolution and quality of the output video that is generated is found to be less.

Input_image1.jpg :- 53 kb is the size.

Resolution: 256*256 pixels.

Processing time: 5 minutes 9 seconds to process the video.

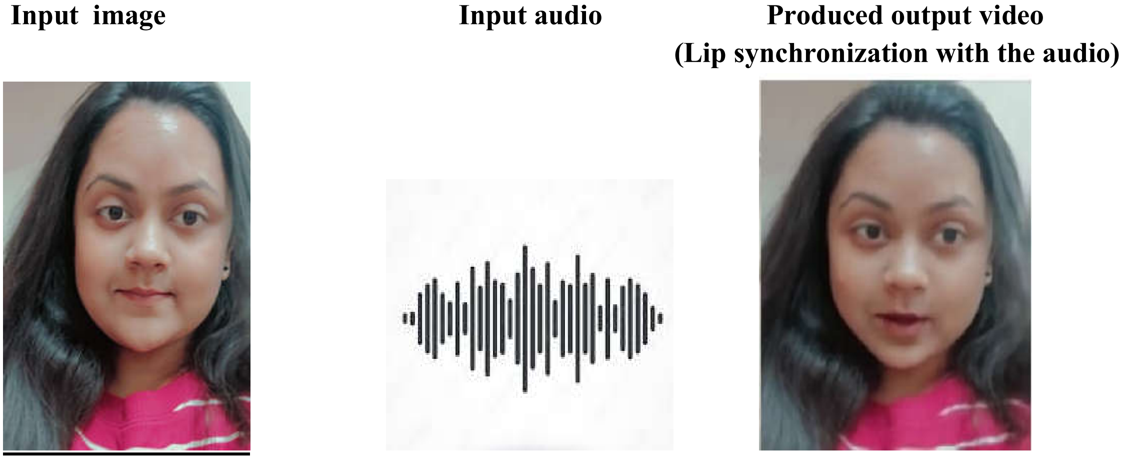
Input_image2.jpeg:- 5 kb is the size.

Resolution: 240*210 pixels.

Processing time: 4 minutes 24 seconds to generate a video.

This shows that it took less time to process small size images with low resolution. After processing the visual quality of the images deteriorated.

Results of Wav2lip model:



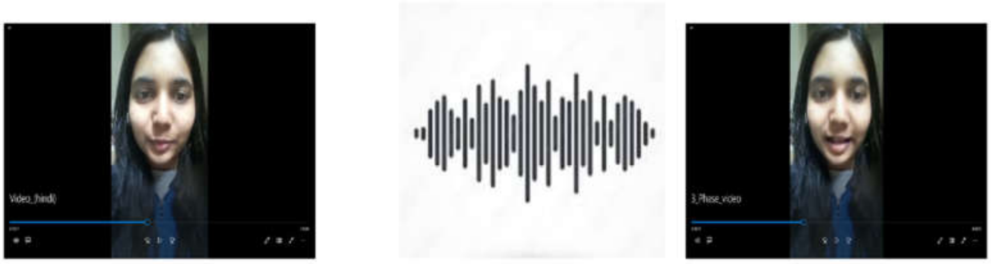
Input image

Input audio

Produced output video (Lip synchronization with the audio)

Figure 9.

Figure 9 demonstrates the lip synchronization of the input image. The inputs are the still image i.e the photograph and an audio in English language. The generated video is a taking image with proper lip synchronization based on the audio given.



Input Video (In hindi) *Input Audio (In English)* *Output Video (In English)*

Figure 10.

Figure 10 demonstrates translation of the language. The inputs contain a video which is in Hindi language and an input audio which is English language. Both the input audio and input video should necessarily be of the same length. The input video is of length 5 seconds and audio is of length 5 seconds.

The model is now trained and the translation will take place. The output video generated will contain the same input video.e the same face but the language will be changed. Previously the input video consisted of Hindi language but now the generated output video will have English language.This change of language happened because of the audio given which is in English language.

The visual quality was improved because of the addition of a visual quality discriminator. A large temporal window permits good lip synchronization. The rating on producing synchronous video frames for dubbed movies from 12.87%(LipGAN) to 11.84% (Wave2Lip + GAN), enhancing the common user-choice from 2.35% (LipGAN) to 60.2% (Wave2Lip + GAN).Since, pre-trained professional lip-sync discriminator is 91% accurate in detecting lip-sync errors, while LipGAN's lip-sync discriminator is 56% accurate.

VI. Applications

In this developing world , the consumption of audio-visual content is growing exponentially. Large scale video translation and generation are desperately needed. This model has the potential to meet these requirements. One such application is movie dubbing. It can also animate CGI characters' lips in response to sounds, which can save several hours of work when creating animated films. Artificial intelligence-generated synthetic media has the potential to open up a lot of doors in the entertainment industry. We may use this technology to increase the reach and magnify the message for social or charitable reasons.

Educational videos - A large portion of the educational content available on the internet is in English. They are frequently available with foreign language subtitles. However, this adds to the viewer's cognitive strain. This model can make films that are lip-synced to dubbed speech in several other local languages as a solution to this problem. It provides a pleasant viewing experience. This approach can be used to translate educational content while maintaining lip synchronization.

VII. Conclusion

A new method is projected for synthesizing full photo-real video portraits of target actors in front of generic static backgrounds. It is used to transmit an input actor's head movement , facial expression, and eye movements to a goal actor. It also takes into account the background movements or the movement due to clothes. All these movements are considered to be a part of dense motion. It expands the capabilities of various applications, including virtual reality and telepresence video reenactment, with video editing, and visual dubbing. As a result, our technique could be a step toward extremely realistic full-frame video content generation under the management of necessary parameters.

VIII. Future Scope

In this digital world as the progress is made further this method may give smooth results or generate smoothly. Using this method especially in the education sector people will easily understand the content, lectures delivered by the head of the universities. This will enhance the country's growth and make people understand the content in a better way.

Deep fakes are growing increasingly in various sectors like education ,news articles and entertainment.This approach majorly results in a few critical troubles like frauds and misusing political perspectives. Moreover, one more factor to not forget about deep fakes is that they play around with organization and identification of any people or character.There is a huge chance that a user is able to make anyone do something that he or she has not done prior.

Visual dubbing: These strategies are used to attain specific lip synchronization.Mostly online lectures or the lecture series available on any platform like coursera or udemy are mostly in English.So,it becomes very difficult to a person to understand the language.Hence,by using this visual dubbing technology help people to listen the same lectures in their local languages.

This technology in future can save numerous hours while developing or making animated movies or rich game content.

References

[1] Aliaksandr Siarohin DISI, Stephane Lathuiliere, Sergey Tulyakov, Elisa Ricci, Nicu Seb, "First Order motion model for Image Animation".

[2] Harshal Vyas, "Deep Fake Creation by Deep Learning", International Research Journal of Engineering and Technology (IRJET) , Volume: 07, July 2020

[3] Hyeongwoo Kim, Max Planck Institute for Informatics, Pablo Garrido, "Deep Video Portraits", ACM Trans. Graph., Vol. 37, No. 4, Article 163. Publication date: August 2018

[4] Abdulqader M. Almars, Deepfakes Detection Techniques Using Deep Learning: A Survey, *Journal of Computer and Communications*, **9**, 20-35. doi: 10.4236/jcc.2021.95003.

[5] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, Duc

Thanh Nguyen, Saeid Nahavandi, “Deep Learning for Deepfakes Creation and Detection: A Survey”, IEEE, 2021

[6] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild ”, 23 August 2020

[7] Bahar Uddin Mahmud and Afsana Sharmin, “Deep insights of Deepfake Technology :A Review”, DUJASE, September 2020.

[8] Rudrabha Mukhopadhyay, Abhishek Jha, Jerin Philip, “Towards Automatic Face-to-Face Translation”, MM ’19, October 21–25, 2019, Nice, France.