Automated audio to video translation

Ahmed Mohamed Hassan
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
ahmed20191700063@cis.asu.edu.eg

Mohamed Ali Salama
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
mohamed20191700563@cis.asu.edu.eg

Ahmed Salah Mohamed
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
ahmed20191700039@cis.asu.edu.eg

Ahmed Mohamed Abu Alfarh
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
ahmed20191700060@cis.asu.edu.eg

Ahmed Mohmaed Wafik
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
ahmed20191700070@cis.asu.edu.eg

Yoman Ahmed
Computer Science Department ,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
yomna.Ahmed@cis.asu.edu.eg

Dina Khattab
Scientific Computing Department
,Faculty Of Computer And Information Sciences ,Ain Shams
Cairo,Egypt
dina.khattab@cis.asu.edu.eg

Abstract—This work explores the problem of automatically translating a video of a person speaking in one language into another language while maintaining realistic lip synchronization. Existing methods perform well on static images or specific individuals seen during training but struggle with accurately syncing arbitrary identities in dynamic talking face videos. To address this, we used wav2lip model that learns from a powerful lip-sync discriminator, resolving key challenges. Wav2lip paper introduced new evaluation benchmarks and metrics for measuring lip synchronization in unconstrained videos. Extensive quantitative evaluations demonstrate that wav2lip model [1] achieves lip-sync accuracy comparable to real synced videos. Additionally, we fine-tuned the model on two datasets (Lombard grid corpus, grid corpus) and evaluate its performance on a small dataset called (vidtimit), obtaining results close to real videos. To enhance the system's flexibility and applicability in various scenarios. We created system that can take different types of inputs like text with image or text with video and can take audio with image or audio with video

we created face to face translation system that takes input video in language A to produce another video in language B. This involves leveraging state-of-the-art solutions in speech recognition, speech synthesis, gender detection, and machine language translation.

COMPUTING METHODOLOGIES,COMPUTER VISION,LEARNING FROM CRITIQUES, DEEP LEARNING,GANS

I. INTRODUCTION

With the increasing consumption of audio-visual content, the need for rapid video content creation has become essential. However, a significant challenge lies in making these videos accessible in different languages. For example, translating a deep learning lecture series, a popular movie, or a public address into desired target languages would allow millions of new viewers to access them.

When individuals attempt to watch an important conference or presentation for an hour or more, the lack of synchronization between the video and the dubbed audio can lead to boredom.

A crucial aspect of translating such videos featuring talking faces or creating new ones is ensuring that the lip movements align with the target speech.

Consequently, the research community has shown considerable interest in lip-syncing talking face videos to match a given audio stream. Initial works in this area utilized deep learning to learn a mapping from speech representations to lip landmarks, relying on several hours of data from a single speaker.

More recent advancements generate images directly from speech representations, exhibiting exceptional generation quality for specific trained speakers. The motivation behind incorporating a visual module into a translation system-stems from the fact that most information streams today are increasingly audiovisual. This approach can be applied in various applications, such as translating lectures or TV series, dubbing movies into different languages, creating cut scenes for the 3D industry involving human characters, translating important conferences, generating missing video segments, and developing visual chatbots. The significant advantage of solving this problem is the ability to generate lip motion on a single, static image of any individual, with any voice, thereby facilitating the creation of social media content. The objective of this project is to develop an application that takes a recorded voice of a speaker and any video featuring a speaking person as input.

The application should generate a new synchronized video with the new voice, adjusting the mouth movements to match the target voice accurately. This lip-synced video will enhance accessibility for a larger portion of the public. Our aim is to create a comprehensive translation system that can transform a video in language A into a new video in language B.

II. Related Work

Lip Sync Discriminator:

The author showed that SyncNet model [2] that can be used as a discriminator to determine the lip-sync-error, also introduce models that generate lip-sync-videos like Speech2Vid model [3]. We saw that LipGan[4] and Wav2Lip [1] models are two networks consisting of generator which is like Speech2Vid model and modifies it and the other network is the discriminator which is like SyncNet [2] model and modifies it. The difference between LipGan and Wav2Lip models can also be seen and how Wav2lip improves the LipGan model. We talked about VispGan[6] that encode word labels into video and ensure that the words is said by using pre-trained lip reading model and by using cycle loss to preserve the content of the input video through a cycle application of generator.

Advancements in lip sync problem

DAVS model [5] has several appealing properties: (1) A joint audio-visual representation is learned through audio-visual speech discrimination by associating several supervisions. The disentangled audio-visual representation significantly improves lip reading performance (2) Audio-visual speech recognition and audio-visual synchronizing are unified in an end-to-end framework; (3) Most importantly, arbitrary subject talking face generation with high-quality and temporal accuracy can be achieved by our framework both audio and video speech information can be employed as input guidance. Another improvement of the DAVS model which is called (PC-AVS) model which emphasize several appealing properties of the framework: 1) Without using any structural intermediate information. It devise a pose code and modularize audio-visual representations into the latent identity, speech content, and pose space. 2) The complementary learning procedure ensures more accurate lip sync results than previous works. 3) The pose of our talking faces can be freely controlled by another pose source video, which can hardly be achieved before. 4) model shows great robustness under extreme conditions, such as large poses and viewpoints.

Finally, we chose **wav2lip**[1] model as the main architecture as it has good accuracy among other models with the type of input, we want the system to handle.

III. Methodology

Our system consists of sub modules to get full translation system that handles different inputs to make the lip sync model more useful and has more features to the users with depending on wav2lip model as lip sync module.

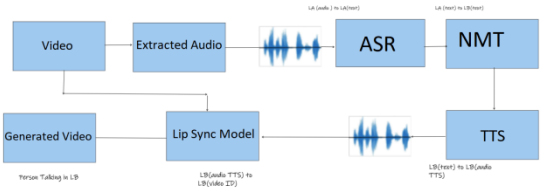


Figure 1:System Modules

Each module will be discussed in the following subsections.

A. Translation modules:

ASR:

Is Automatic speech recognition which is used to extract a transcript from any audio file from video, maybe the same video but to use transcript in automatic translation of the video in language A to a video in language B.

NMT (Natural machine translation):

This module helps to take the transcript from previous step and translate the language to the other language.

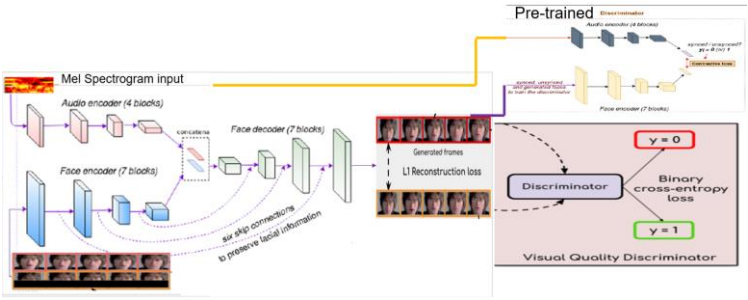
TTS (Text to speech):

This module helps generate a human like voice using most well-known APIs and deep learning models.

Lip Sync Model:

This is the **main** module that we developed to generate the new video based on any arbitrary audio.

B. Lip Sync module:



Generator Architecture Details. We used wav2lip' generator architecture as used by Lip GAN [1]. Our key contribution lies in training this with the expert discriminator. The generator G contains three blocks: (i) Identity Encoder, (ii) Speech Encoder, and a (iii) Face Decoder. The Identity Encoder is a stack of residual convolutional layers that encode a random reference frame R, concatenated with a pose-prior P (target-face with lower-half masked) along the channel axis. The Speech Encoder is also a stack of 2D convolutions to encode the input speech segment S which is then concatenated with the face representation. The decoder is also a stack of

convolutional layers, along with transpose convolutions for up sampling. The generator is trained to minimize L1 reconstruction loss between the generated frames Lg and ground-truth frames LG:

$$L_{recon} = \frac{1}{N}\sum_{i=1}^{N} ||Lg - LG||1$$

Thus, the generator is like the previous works, a 2D-CNN encoder-decoder network that generates each frame independently. Using pre-trained expert lip-sync discriminator that needs a temporal window of Tv = 5 frames as input starting from a random frame along with another five ground truth frames.

C. Generating Photo-realistic Faces

In wav2lip paper's experiments using a strong lip-sync discriminator forces the generator to produce accurate lip shapes. However, it sometimes results in the morphed regions to be slightly blurry or contain slight artifacts. To mitigate this minor loss in quality, a simple visual quality discriminator in a GAN setup along with the generator. Thus, the author[1] used two discriminators, one for sync accuracy and another for better visual quality. The lip-sync discriminator is not trained in a GAN setup. On the other hand, since the visual quality discriminator does not perform any checks on lip-sync and only penalizes unrealistic face generations, it is trained on the generated faces. The discriminator D consists of a stack of convolutional blocks. Each block consists of a convolutional layer followed by a Leaky ReLU activation . The discriminator is trained to maximize the objective function $Ldisc$ (Equation 3):

$$Lgen = E_{x\sim Lg}[\log(1 - D(x)] \quad (2)$$

$$Ldisc = E_{x\sim LG}[\log(D(x)] + Lgen \quad (3)$$

where Lg corresponds to the images from the generator G, and LG corresponds to the real images. The generator minimizes Equation 6, which is the weighted sum of the reconstruction loss (Equation 1) , so total loss will be :

$$L_{total} = (1 - s_w - s_g) * L_{recon} + s_w * E_{sync} + s_g * L_{gen}$$

where sw is the synchronization penalty weight, s_g is the adversarial loss which are empirically set to 0.03 and 0.07 in all our experiments. Thus, the complete network is optimized for both superior sync-accuracy and quality using two disjoint discriminators. The model only on the LRS2 train set [1], with a batch size of 80. The model used Adam optimizer [1] with an initial learning rate of 1e −4 and betas β1 = 0.5, β2 = 0.999 for both the generator and visual quality discriminator D. Note that the lip-sync discriminator is not fine-tuned further, so its weights are frozen. Similar to LipGAN [1], the model generates a talking face video frame-by-frame. The visual input at each time-step is the current face crop (from the source frame), concatenated with the same current face crop with lower-half masked to be used as a pose prior. Thus, during inference, the model does not need to change the pose, significantly reducing artifacts. The corresponding audio segment is also given as input to the speech sub-network, and the network generates the input face crop, but with the mouth region morphed.

IV. Experiments and Results

A. Enhancing the lip sync module:

We fine-tuned the model on two new datasets (Lombard Grid dataset, Grid Corpus dataset).

Which contains face videos about people talks and reading different passages and the different from the data (**LRS2**) which the model was trained on it is that the quality of the videos is high which will help in making the model learn more and produce high quality videos.

B. Evaluation of the lip sync module:

We test the wav2lip model on a small data called (vidmit) which contains some videos of people with one color background and has about 30 videos and we evaluate the model by two metrics (LSE-loss , LSE – confidence)

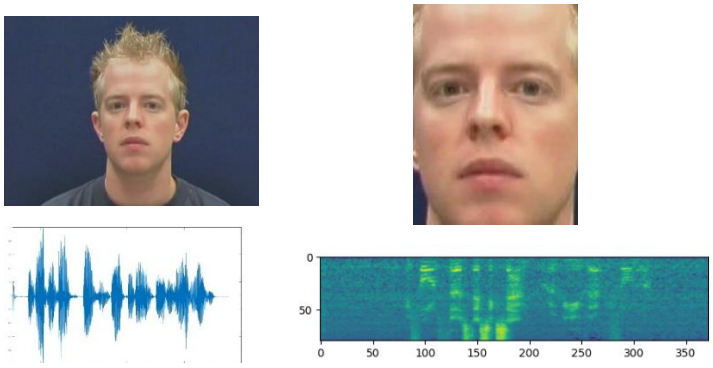
We evaluated the model by taking videos from dataset and calculate the two metrics (LSE-loss,LSE-confidence) and generating the same videos again with the same audio by wav2lip model to compare the results.

Preprocessing steps for generated the fake videos:

(1) Extracted cropped frames that contains face only.

(2) Extracted speech from the video.

Example input **Example output after preprocessing:**



After evaluating the wav2lip on vidtimit dataset by generating new videos

	Avg Confidence	Avg minimum distance
Real Videos	3.335	8.33
Fake Videos	2.633	8.75

C. Face to face translation:

We extend the problem of automatic machine translation to face to face translation with a focus on audio-visual content, i.e., where input and output are talking face videos. Beyond demonstrating the feasibility of a Face-to-Face translation pipeline.

Speech to text:

Used Wav2vec model to extract text from English videos to translate to transform the text to another language it later.

Text to speech module:

Used **salorie** models which contains many models with different speakers in different ages to have options in producing the sound and supports many languages except Arabic.

Used Microsoft azure api which has options to use a dialect in Arabic language to produce high quality audio to the wav2lip.

fine-tuned Tacotron2 text to speech model to produce another voice of a male in Arabic.

Gender Detection and Age:

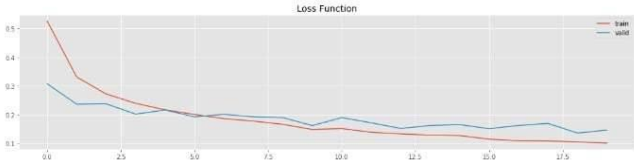
Trained two deep learning models one for detecting gender from a frame in video and the other model to detect the age to use it in producing the sound (on going producing audio based on age).

Age detection model:

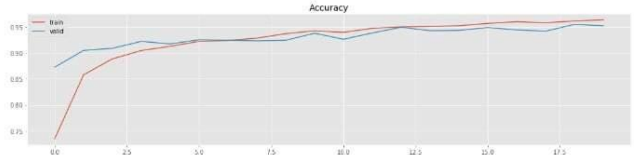
We trained the model on dataset: **CelebA** is a large-scale face with more than 200k celebrities and 10,177 number of identities.

Chosen Model for this task is (Inception model v3) as it has good accuracy in many tasks.

The loss after we trained the model:

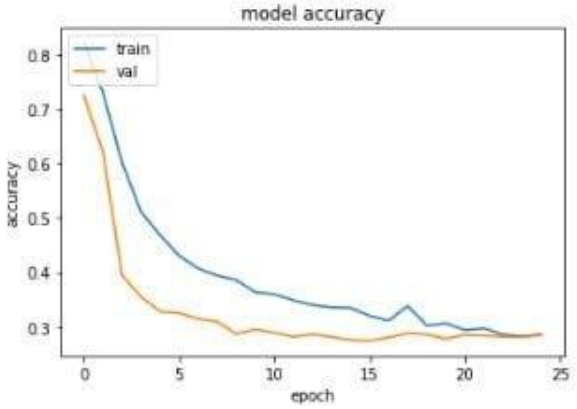


The accuracy:



Age detection model:

We trained vgg16 to predict the age as it's regression problem on the dataset **UTK faces.**



V. Conclusion

The presented work focuses on generating accurate lip-synced videos of unconstrained talking faces. The authors identify two key challenges in current approaches and propose a pretrained lip-sync "expert" to enforce natural lip motion generation, resulting in improved accuracy. The model is evaluated using new benchmarks and metrics proposed in the wav2lip paper[2], outperforming existing approaches in both quantitative measures and human evaluations.

Additionally, the work extends the concept of automatic machine translation to face-to-face translation, specifically audio-visual content involving talking face videos. In Wav2Lip paper [2], they introduced a novel approach for talking face generation and contribute to various language processing tasks, particularly for resource-constrained languages. Practical applications of the research include automatic dubbing of educational videos, movie clips, and interviews.

The "Face-to-Face Translation" concept opens up new research directions in computer vision, multimedia processing, and machine learning. Areas of exploration include modifying speech duration during translation, transforming gestures, expressions, and background content accordingly, and improving existing individual modules.

VI. Future Work

We believe our efforts and ideas in this problem can lead to new directions such as synthesizing expressions and head-poses along with the accurate lip movements. Making the model can run in real time will be important for video calls, live conferences, and online meetings. Voice Cloning for Multilingual Videos: Extending the model to generate videos in a new language while retaining the speech from the source language (voice cloning) would be an interesting direction. This could enable the creation of multilingual videos with. Summarizing large videos with a small of version of it by summarizing the transcript and generate new

summarized version. Fine tuning the model on another dataset in another language will help produce better lip-synced videos. Enhancing Realism with Facial Dynamics: Exploring techniques to incorporate facial dynamics, such as subtle muscle movements and micro expressions, can further enhance the realism of lip-synced videos. This could involve developing models that capture and synthesize these nuances to create more expressive and lifelike facial animations.

References

[1] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). 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 (pp. 484-492).

[2]Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV.

[3] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? arXiv preprint arXiv:1705.02966 (2017).

[4] KR, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., & Jawahar, C. V. (2019, October). Towards automatic face-to-face translation. In Proceedings of the 27th ACM international conference on multimedia (pp. 1428-1436).

[5] Wang, Ganglai, et al. "Attention-Based Lip Audio-Visual Synthesis for Talking Face Generation in the Wild." arXiv preprint arXiv:2203.03984 (2022).

[6] Doukas, Michail C., Viktoriia Sharmanska, and Stefanos Zafeiriou. "Video-to-Video Translation for Visual Speech Synthesis." arXiv preprint arXiv:1905.12043 (2019).

[7] Zhou, Hang, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. p. 4176-4186.

[8] Mira, Rodrigo, et al. "End-to-end video-to-speech synthesis using generative adversarial networks." IEEE Transactions on Cybernetics (2022).