Synthesis of Talking Face with Emotion given an image and speech

A Project Report Submitted to the
Department of Computer Science of
Ramakrishna Mission Vivekananda Educational and Research Institute, Belur,
in partial fulfillment of the requirements for the degree of
MSc
in
Big Data Analytics

Submitted by
SUBHAJYOTI MAITY
ID No. B2230038

Supervisor:
Sw. Punyeshwarananda Maharaj
Department of Computer Science
Ramakrishna Mission Educational and Research Institute

Department of Computer Science
Ramakrishna Mission Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India
June 8, 2024

Synthesis of Talking Face with Emotion given an image and speech

By

SUBHAJYOTI MAITY

Declaration by student:

"I hereby declare that the present dissertation is the outcome of my project work under the guidance of Sw. Punyeshwarananda Maharaj and I have properly acknowledged the sources of materials used in my project report."

_____

(Subhajyoti Maity, ID No. B2230038)

A project report in the partial fulfillment of the requirements of the degree of MSc in Big Data Analytics

Examined and approved on

_____

by

_____

Sw. Punyeshwarananda Maharaj (Supervisor)
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

Countersigned by

_____

Registrar
Ramakrishna Mission Vivekananda Educational and Research Institute

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India

Acknowledgement

The present project work is submitted in partial fulfilment of the requirements for the degree of Master of Science of Ramakrishna Mission Vivekananda University (RKMVU). I express my deepest gratitude to my supervisor Sw. Punyeshwarananda Maharaj of Ramakrishna Mission Vivekananda Educational and Research Institute for his inestimable support, encouragement, profound knowledge, largely helpful conversations and also for providing me a systematic way for the completion of my project work. His ability to work hard inspired me a lot. I extend my heartfelt gratitude to the Vice-Chancellor of this University for his continuous encouragement and unwavering support during this journey. Last but not the least, this work would not have been possible without support of my fellow classmates. Also I am always indebted to my parents for their constant support and inspiration. Without them beside me, nothing would be possible.

Belur
June 8, 2024

Subhajyoti Maity
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

Contents

List of Tables

List of Figures

Chapter 1

Abstract

Visual emotional expression is crucial in communicating through audiovisual speech. Here, we present a novel approach for illustrating emotional facial expressions in speech-driven talking face synthesis[14]. Enhancing speech intelligibility involves integrating both lip and facial images into the speech signal, thereby emphasizing the significant role of lip image synthesis in achieving lifelike computer-generated faces[10].

we've developed a comprehensive system for generating talking faces. It operates by taking any random audio speech, an image of a chosen person's face, and a categorical emotion tag as inputs. The goal is to produces a video of a talking faces that mimic the spoken words of that individual while conveying the specified emotion[6].

Our network employs spatial-temporal discriminators for realistic images and videos, complemented by a lip-reading discriminator to refine lip synchronization accuracy, vital for convincing talking face videos[31]. The training process consists of three primary stages. Initially, we adjust the landmarks of the initial video frame to position the two eye points at predetermined locations, and this adjustment is applied to all subsequent video frames. Next, we eliminate identity-specific information by standardizing the landmarks to match the average face shape across the entire training dataset[13].

The code will be made publicly available at

`https://github.com/subhajyotirkmveri/research_intern`

Keywords: Talking face generation; Emotion; Audiovisual;Lip movements generation;Audio visual correlation

Chapter 2

Introduction

Speech communication does not solely depend on the acoustic signal. Visual cues, when present, also play a vital role[2]. The presence of visual cues improves speech comprehension in noisy environments and for the hard of-hearing population[3].

As a result, scientists have engineered systems capable of generating talking faces automatically from speech, thereby offering visual cues in their absence. These systems enhance the accessibility of abundant audio resources for individuals with hearing impairments and improve the quality of interactions with computers. Moreover, they find extensive utility in entertainment, education, and healthcare sectors[19].

In speech communication, emotions significantly influence the conveyed message and can substantially alter its meaning. Research indicates that predicting emotions based solely on speech audio is challenging for individuals without training, emphasizing the reliance on visual cues for interpreting emotions. Therefore, to enhance the authenticity of visual representation and advance speech communication, automated systems generating talking faces should integrate visual displays of emotions[12].

In many scenarios such as telephony, however, speech communication is still acoustical. The absence of the visual modality can be due to the lack of cameras, the limited bandwidth of communication channels, or privacy concerns. One way to improve speech comprehension in these scenarios is to synthesize a talking face from the acoustic speech in real time at the receiver's side. A key challenge of this approach is to make sure that the generated visual signals, especially the lip movements, well coordinate with the acoustic signals, as otherwise more confusions will be introduced[11].

There are two approaches for audiovisual synthesis: synthesis from text and synthesis from speech. This paper focuses on speech synthesis. Lip movement synthesis requires various information, including phonetics, coarticulation, and timing. Similar to text-to-speech synthesis, controlling all these parameters solely based on text information is generally challenging. In contrast, speech inherently contains phonetic, coarticulation, and timing information. Therefore, lip movement synthesis from speech appears to be more promising than synthesis from text[40].

An alternative method for generating emotional talking faces involves initially inferring emotions from speech utterances and then reflecting them in the generated faces. However, this method is constrained by the accuracy of speech emotion recognition and lacks the ability to independently manipulate emotional expression in visual rendering. In our study, we adopt a distinct strategy: we disregard the emotions conveyed in the speech audio and instead base the generation of talking faces on a separate emotion variable[15]. This approach offers direct and adaptable management of visual emotion expression, potentially facilitating more customized applications in entertainment, education, and interactive assistive technologies.

Like in other audio processing tasks, end-to-end methods typically acquire more effective representations of the audio signal and produce superior outcomes when contrasted with manually engineered features like Mel Frequency Cepstral Coefficients (MFCCs). Furthermore, many of these techniques were neither formulated for nor tested in noisy environments, yet robustness to noise is crucial in real-world scenarios[14].

In this paper, we propose a system that can generate a talking face video with a frame rate of 25 frames-per-second (FPS) from an arbitrarily long noisy speech utterance and a single face image. The speech utterance and the face image do not need to belong to on, and neither identity is exposed to the system beforehand.

Compared to existing work, This study demonstrate a groundbreaking neural network framework designed to produce emotive facial expressions synchronized with speech. This system utilizes a combination of speech input, a reference facial image, and categorical emotion cues to generate expressive talking faces.

Chapter 3

Literature Survey

Generating a talking head automatically has been a great interest in the research community. Some researchers focused on text-driven generation. These methods map phonemes to talking face images[5, 36]. Compared to text, voice signals are surface-level signals that are more difficult to parse. Besides, voices of the same text show large variations across speakers, accents, emotions, and the recording environments[8]. On the other hand, speech signals provide richer cue for generating natural talking faces. For text, any plausible face image sequence is sufficient to establish natural communication. For speech, it must be a plausible sequence that matches the speech audio. Therefore, text-driven generation and speech-driven generation are different problems and may require different approaches[31].

The field of talking face generation has seen significant advancements in recent years, becoming a key area of research with a wide range of applications, including video conferencing, virtual assistants, and entertainment, among others. Recently, researchers have commenced incorporating emotion-conditioned facial expressions into talking face generation[18, 22, 15], leveraging emotional annotations in talking video datasets like MEAD[37]. However, none of these methods have yet succeeded in generating expressive and lifelike expressions in their talking faces.

Suwajanakorn and team (2017)[32] devised a system creating videos featuring President Barack Obama from speech data. Utilizing an LSTM network, the system forecasts PCA coefficients of mouth landmarks from speech attributes (13 MFCCs plus the energy). Following this, texture retrieval relies on projected PCA coefficients by selecting neighboring candidate frames from a dataset housing images of the target individual. These frames undergo amalgamation through a weighted mean.

It's notable that this approach is tailored for a solitary subject and necessitates a substantial dataset for accurate visual representation.

Chung and colleagues [21] proposed a convolutional neural network (CNN) system to generate a photo-realistic talking face video from speech and a single face image of the target identity. In contrast, Chen, Lele and Li and their groupmate (2018)[6] proposed a lip generation system incorporating an adversarial loss function to sharpen blurry outputs alongside a reconstruction loss. However, this system is limited to generating only the lip region, not the entire face. Compared to [32], the reduction from several hours of face videos to a single face image for learning the target identity is a great advance.

Another two-stage system was proposed by L. Chen and colleagues(2019)[7]. The system first predicts 68 face landmarks from speech using an LSTM-based network[13], and then extract face landmarks and align them across different speakers and transform their shapes into the mean shape to remove the identity information. They also employ a discriminator network to improve image quality. In another work, Egor Zakharov and his teammate [41] proposed a style-based landmark-to-image conversion method using generative adversarial networks (GANs) with a few shots of the target face. This method, however, lacks landmark adaptation methods to solve personality mismatch issues.

In contrast, Zhou, Hang and Liu and others[42] proposed a GAN-based approach modeling the entire face and introduced a temporal-GAN loss to enhance temporal dependency across frames. Additionally, Song et al. (2017)[31] introduced a method employing a conditional recurrent adversarial network to enhance realism in talking face generation. Vougioukas et al. (2018)[35] introduced a temporal-GAN approach to generate more authentic image sequences, augmented by three discriminators to en- hance video frame realism, continuity, and audio-visual synchronization. Similarly, Eskimez and his group members[14] proposed a robust end-to-end talking face generation system, featuring a frame discriminator to enhance image quality and a pair discriminator for improved lip-speech synchronization. They also introduced a mouth region mask (MRM) to refine lip-speech alignment, demonstrating superior alignment compared to standard approaches.

Straight-forward adaptations of GANs for videos are proposed in their work (2016)[34] and Saito et al. (2017)[29], replacing the 2D convolutional layers with 3D convolutional layers. Using 3D convolutions in the generator and discriminator

networks is able to capture temporal dependencies but requires fixed length videos. This limitation was overcome in Saito et al. (2017)[29] but constraints need to be imposed in the latent space to generate consistent videos.

In affective computing, researchers use emotion models to develop systems for automatic emotion detection. The most utilized emotion models for automatic systems are 1) categorical models 2) dimensional models. Ekman's model suggested six basic emotion categories: anger, disgust, fear, happiness, neutral, and sadness. Dimensional models[26, 28, 24] argue that emotions are correlated, and each emotion category can be represented with a combination of values from emotional dimensions.

In recent research endeavors, there has been a growing focus on the development of emotionally expressive talking faces[22, 33, 39]. ExprGAN[9] introduced an expression control module that enables the synthesis of faces with diverse expressions, allowing for continuous adjustment of expression intensity.

Chapter 4

Proposed Algorithm

In this work, we follow a end-to-end realistic video generation architecture given in [15] that takes an input image, a speech utterance, and a conditioned categorical emotion as inputs.

Figure [4.1] shows overview of the system. Specifically, the generator contains an image encoder, a speech encoder, a noise encoder and emotion encoder as input . It concatenates feature outputs from the encoders and uses a decoder to generate a talking face video with emotion. We utilize generative adversarial networks[16] in this system; specifically, we propose four discriminators. one discriminator to distinguish between emotions expressed in videos, while the frame discriminator evaluates the validity of every single frame of the video. Also we use MRM reconstruction loss and perceptual loss to improve the mouth-audio synchronization and to improve image quality accordingly. In the following, we describe each module in detail.

4.1 Network Architecture

4.1.1 Speech Encoder:

The speech encoder processes the input speech waveform and outputs a speech embedding. It follows the original implementation of [13] without any modification. Develop a speech encoder that can process an arbitrary-length speech waveform without any pre-processing, and output 25 feature vectors per second. It contains five convolutional layers with 1-D kernels operating in the time domain. The number of filters, filter sizes and strides for these convolutional layers are (64,63, 4), (128,

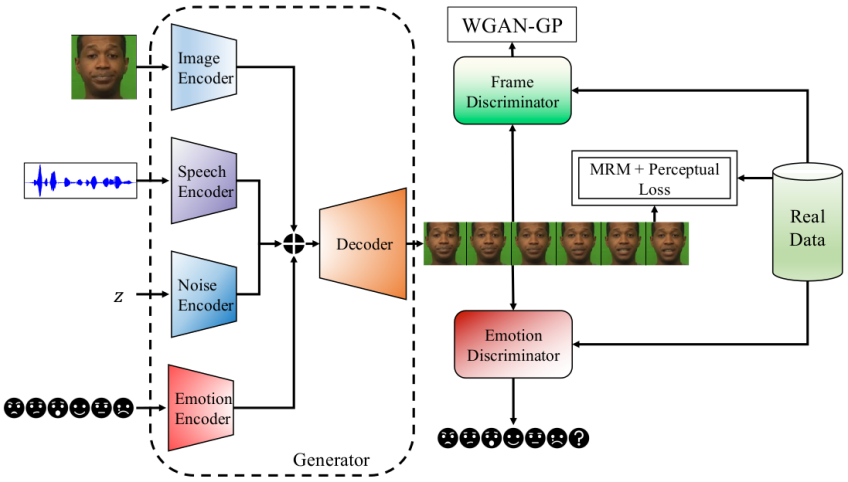

Figure 4.1: Overview of the proposed neural network system. The system takes as input a reference image, a speech waveform, a random vector sampled from a standard normal distribution, and a categorical emotion label. These inputs are processed and their embeddings concatenated to generate a synchronized talking face video that reflects the given emotion.(Image Source: [15])

31, 4), (256, 17, 2), (512, 9, 2), (16, 1, 1), respectively. Each convolutional layer is followed by a LeakyReLU activation with a 0.2 slope. The context layer then keeps every fifth feature vector and discards the rest. Finally, the resulting features are fed to a fully connected layer. For 1 second of speech in 8 kHz, the input size is 8000, and after these convolutional layers, it is reduced to 125. The context layer further reduces The context layer reduces the 125 time-steps to 25 time-steps by passing only every fifth frame to the next layer. Therefore, our generated videos are in 25 frames-perseconds (FPS). The output of the context layer is fed to a fully connected layer, followed by two LSTM layers, which output the speech embedding sequence.

4.1.2 Image Encoder:

The image encoder takes a single reference face image as input to encode the target identity. The architecture adheres to the original implementation without any changes[14]. It comprises six layers of 2-D convolutional layers with the following specifications: number of filters, kernel sizes, and down-sampling factors: (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), and (512, 4, 1), respectively. A LeakyReLU activation with a 0.2 slope follows each convolutional layer. It's im-

portant to note that nearest-neighbor interpolation is employed for downsampling instead of strides, which helps eliminate artifacts in the generated images. The final image embeddings, along with the intermediate representations, are transferred to the video decoder using U-Net style skip connections[27].

4.1.3 Noise Encoder:

In order to make the model robust to movements irrelevant to speech, and promote diverse mouth shapes, we follow [35] to add a noise input to the generator. We generate a 128- d Gaussian noise vector at every time-step with zero mean and unit standard deviation and pass it through an LSTM layer before passing it to the decoder. This noise input also allows for the addition of a temporal discriminator, as described in Section 2, without altering the architecture.

4.1.4 Emotion Encoder:

We assume 6 basic emotion categories: happiness, sadness, fear, anger, disgust, and neutral. To encode these categorical emotions, we add an emotion encoder block to the generator. The emotion label is first encoded as a one-hot vector and fed into the emotion encoder. The emotion encoder utilizes a two-layer fully connected (FC) neural network to transform the one-hot vector into an emotion embedding. It follows the original implementation of this paper[13]. This embedding is then replicated for each time step. Additionally, we apply a LeakyReLU activation with a 0.2 slope after each FC layer.

4.1.5 Decoder:

As shown in Figure 4.2, the decoder concatenates the speech, image, noise and emotion features calculated by their encoders, and feeds them to a fully connected (FC) layer. The resultant vectors are then reshaped to 2D images. They are then passed through 6 convolutional layers that are symmetric to the 6 layers in the image encoder. Before going through each of the first 5 layers, the output of the corresponding layer of the image encoder is concatenated as input to form U-Net style skip connections. To make the generated mouth shapes vary smoothly over time, and also use LSTM type recurrent connections to the 4th layer, making it an LSTM-CONV layer wihout any modification of this paper[15]. The number of

filters, filter sizes and up-sampling factors for these convolutional layers are (512, 3, 2), (256, 3, 2), (128, 3, 2), (64, 3, 2), (32, 3, 2), and (3, 3, 1), respectively.

A LeakyReLU activation with a 0.2 slope follows each convolutional layer, except for the last layer, where instead, hyperbolic tangent activation is used since the images are normalized to have values between -1 to 1.

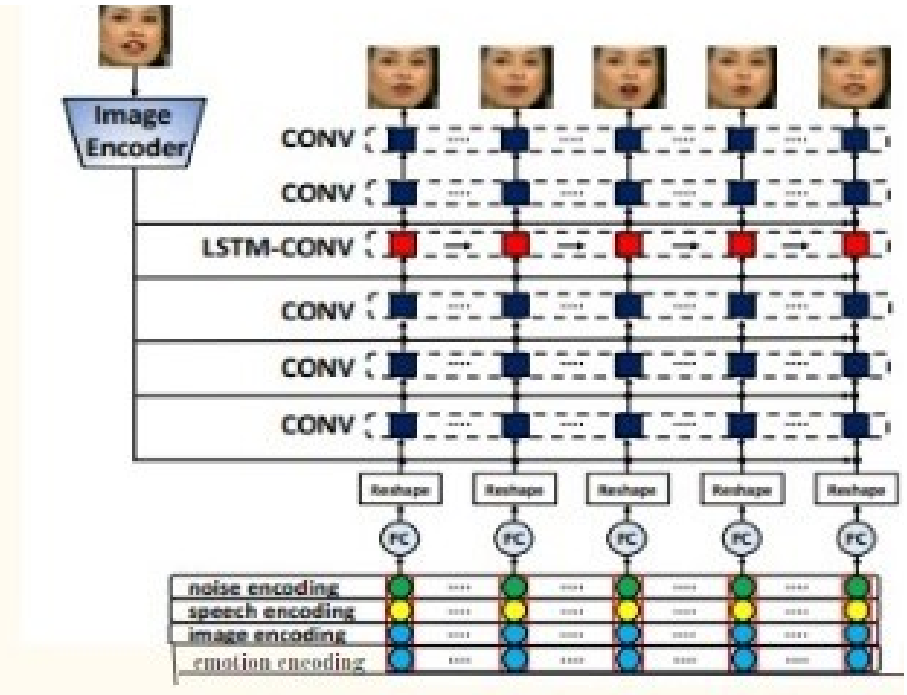

Figure 4.2: The decoder architecture. The input is the speech, condition image, emotion condition and noise features concatenated at each time-step. The features coming from the condition image are concatenated to each layer's output, except for the last layer.(Image Source:[14]; with a slight modification)

4.1.6 Frame Discriminator:

The Frame Discriminator achieves a high-quality reconstruction of the speakers' face throughout the video. The frame discriminator aims to improve the image quality of the generated video and to keep the target identity consistent throughout the video. The Frame Discriminator is a 6-layer CNN that determines whether a frame is real or not. Adversarial training with this discriminator ensures that the generated frames are realistic. Furthermore, the original still frame is concatenated channel-wise to the target frame and used as a condition, which enforces the identity onto the video frames. The number of filters, kernel sizes, and strides of these convolutional

layers are as follows: (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), respectively. The output is then flattened and fed into a two-layer FC network, which classifies the frame as fake or real. Each layer is followed by a LeakyReLU activation with a 0.2 slope except for the last layer, where we do not use an activation, We utilize the Wasserstein GAN[1] with a gradient penalty exactly as described without any modifications of this[35].

4.1.7 Pair Discriminator:

The pair discriminator's objective is to improve the synchronization between the mouth shape and the input speech in the generated videos As shown in Figure 4.3, its image encoder takes the generated or ground-truth videos that are mouth region masked and the condition image as input. It uses the same speech encoder architecture described earlier; however, the weights are not shared. A bidirectional LSTM (BLSTM) processes the output of the speech encoder, where another BLSTM processes the output of the image encoder for each frame. The outputs of the BLSTM's are then concatenated and fed into a third BLSTM followed by an FC layer that outputs the probability score. The hidden neurons for the image, speech and third BLSTMs are 512, 512, and 1024, respectively.I adhere to this paper's implementation[14].

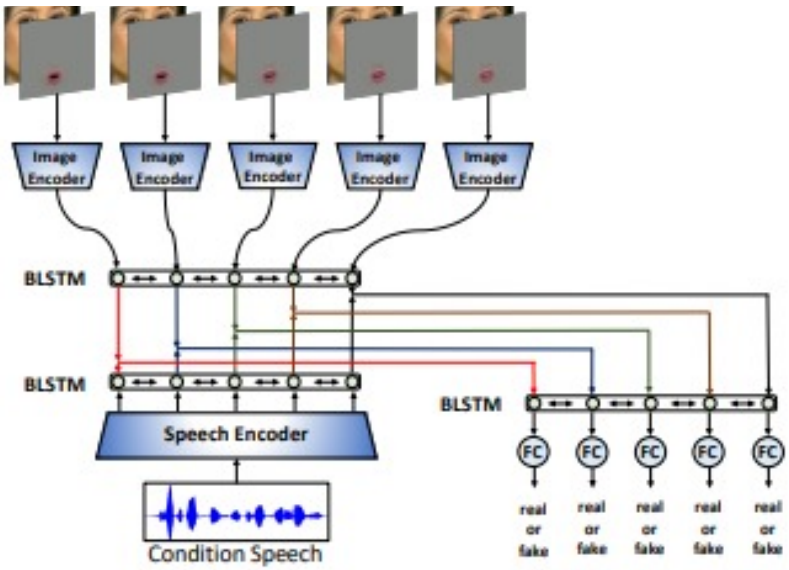

Figure 4.3: The pair discriminator architecture. The input is the speech, the condition image, and generated/ground-truth videos.(Image Source:[14])

4.1.8 Emotion Discriminator:

Our architecture for emotion discriminator is similar to that in [15]. Each frame is passed through five 2D convolution layers. The convolution layers are as follows (representing the number of filters, kernel sizes, and strides, respectively): (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), respectively. The output is then flattened and fed into a two-layer fully-connected network. The resulting sequence is fed into an LSTM layer[20]. The last time step of the LSTM layer's output is passed through a fully-connected layer that outputs probabilities for given categorical emotions. The last time step of the output of the LSTM layer is fed into an FC layer that outputs probabilities of the seven classes: six emotions (anger, disgust, fear, happiness, neutral, and sadness) plus the fake class similar to[25].

4.2 Objective Functions:

We used multiple objective functions that emphasize different aspects of the generated video, such as visual quality, accurate lip-sync, and emotion rendering.

Reconstruction Loss(Frame Discriminator Loss): The generator is trained to minimize the L1 reconstruction loss between the generated frames and the ground truth frames as described in subsection 4.1.6.

Mouth Region Mask (MRM) Loss: The MRM loss is a weighted L1 reconstruction loss between the generated and ground-truth videos around the mouth region. It uses a 2D Gaussian centered at the mean position of mouth coordinates as the weights. The intuition of MRM is to manually drive the attention of the network to the mouth region to improve the mouth-audio synchronization.

Perceptual Loss (PL): A pre-trained VGG-19 network [30] is exploited to calculate the intermediate features of the layers from the ground truth videos and the generated videos. The mean-squared loss between these intermediate features is defined as the perceptual loss (PL) [23].

Emotion Discriminator Loss: The emotion discriminator is optimized using a cross-entropy loss calculated between the emotion class predicted by the emotion discriminator for generated frames and the conditioned emotion class.

$$L_{\text{emo}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{4}$$

where, $N = 7$ (here, N signifies the number of emotion classes).

Combined Objective Function. The full objective function for the generator step is as follows:

$$J_{\text{GEN}} = \alpha L_{\text{MRM}_1} + \beta L_{\text{Perceptual}_2} + \gamma J_{\text{FD}} + \delta J_{\text{ED}}, \tag{4.1}$$

where L_{MRM_1} is the MRM loss, $L_{\text{Perceptual}_2}$ is the perceptual loss, J_{FD} is the frame GAN loss, J_{ED} is the emotion GAN loss, and α, β, γ, δ are the respective weights of each component.

4.3 Experiments:

We evaluate the effectiveness of the proposed method on popular datasets [37], and demonstrate our advantages over the state-of-the-art works. Specifically, the evaluation is conducted in terms of image quality, lip movement/shape accuracy, and video smoothness.

4.3.1 Dataset:

To incorporate the emotions, a dataset with emotion labels is required, and according to our approach, it should fulfill the requirement of a single face in every frame of each clip. Currently, only some such datasets are publicly available. We use CREMA-D[37] for our purpose. Here are the main attributes of the dataset:

- It contains 7442 clips from 91 actors (48 male and 43 female)
- Actors spoke from a selection of 12 sentences.
    1. It's eleven o'clock.
    2. That is exactly what happened.
    3. I'm on my way to the meeting.
    4. I wonder what this is about.
    5. The airplane is almost full.
    6. Maybe tomorrow it will be cold.
    7. I would like a new alarm clock.
    8. I think I have a doctor's appointment.
    9. Don't forget a jacket.
    10. I think I've seen this before.
    11. The surface is slick.
    12. We'll stop in a couple of minutes.
- Sentences were presented using one of the six emotions (happiness, sadness, fear, anger, disgust, neutral).

The image resolution of the provided videos is 480x360, and the sampling rate is 30 frames per second (FPS). The audio is sampled at 44.1 kHz. We downsampled the video to 25 FPS and the audio to 8 kHz.

For each video of the actor, we estimated the similarity transform parameters between the template landmarks and extracted landmarks of the first frame using three points: the temporal mean points of the left eye, the right eye, and the nose. Note that we only took the first frame of each video to estimate the transformation, and used it to align the remaining frames to the template image. In this way, the faces in the resulting videos start from the same spatial location but can wander off to different parts of the scene.

4.3.2 Data Augmentation (DA):

During training, we randomly augmented the data using the Albumentations library[4] to improve the generalization capability of our network. The data augmentation includes randomly changing brightness, contrast, gamma, hue, saturation, and value. In addition, our algorithm includes contrast limited adaptive histogram equalization, adding random Gaussian noise to the image, and shuffling the channels, and shifting RGB values for each channel.

4.3.3 Implementation Details

We started by training with MRM and perceptual losses for 500 iterations. After that, we trained for another 500 iterations using the full objective function. This process was carried out on a small dataset with roughly 1,000 training samples and 200 validation samples. We used the Adam optimizer for all networks with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The learning rate for the generator was $1e-4$ during initialization and $1e-5$ during GAN training. Both discriminators' learning rates were $1e-4$. The constants α, β, γ, and δ mentioned in earlier were 100, 1, 0.01, and 0.001, respectively. The weight for the gradient penalty when training the frame discriminator was 10. All images were normalized between the -1 to 1 value range. During initialization, the mini-batch size was set to 2, and during GAN training, it was set to 2. The training took approximately one week using a GeForce RTX 2080 TI GPU. For the baseline method, we use the pre-trained model (trained with the CREMA-D dataset) provided by the authors.

Chapter 5

Objective Evaluation and Results

This section describes the metrics that are used to assess the quality of generated videos. The videos are evaluated using traditional image reconstruction and sharpness metrics. Although these metrics can be used to determine frame quality they fail to reflect other important aspects of the video such as audio-visual synchrony and the realism of facial expressions. We therefore propose using alternative methods that are capable of capturing these aspects of the generated videos.

- We employ Peak SNR (PSNR) and Structural Similarity (SSIM) [38] to evaluate the image quality of the generated videos. In order to evaluate the mouth synchronization, we extract the face landmarks of the generated and ground-truth videos and calculate their L_2 norm distance. Differently, however, we align the predicted and ground-truth landmarks to a template face landmarks for each frame using Procrustes analysis [17] before calculating the distance between them.

- In [6], this alignment is performed by matching the mean points of the mouth landmarks only. Simply matching the means of the mouth landmarks, however, is prone to errors caused by facial movements such as rotation and scaling. Procrustes analysis removes the translation, rotation, and scaling factors without changing the relative positions of the landmarks, resulting in a more meaningful evaluation metric. We call this metric normalized landmarks distance (NLMD).

Objective evaluation metrics for our proposed method calculated using 40 videos of both ground truth and generated videos given in 5.1.

Method	PSNR	SSIM	NLMD
our followed system evaluation	34.98	0.89	0.034

Table 5.1: Objective evaluation results for our proposed method. For PSNR, values above 30 dB are better; for SSIM, higher values close to 1 are better; and for NLMD, lower values are better.

Some result of this framework is given in the below link:

`https://drive.google.com/drive/folders/1jOgrdsasxNl_5VCvV2E_ETMlvbDkH_vR?usp=drive_link`

Amount and variety of expressions generated is dependent on the amount of expressions present in the dataset used for training and hence faces generated by models trained on expressive datasets such as CREMA-D will exhibit a wider range of expressions. This is illustrated in 5.1, where the facial expressions reflect nutral emotion of the speaker.

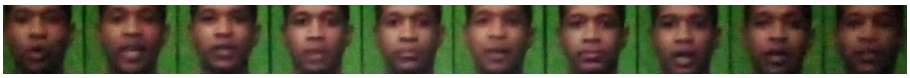

Figure 5.1: Videos produced by the proposed method using the same image taken from the CREMA-D test set and driven by the sentence "Congratulations on your promotion" spoken with a male voice taken from the Kaggle data source with neutral emotions.

Chapter 6

Conclusion and Limitations

In conclusion, this system represents a significant advancement in the field of talking face generation, offering the ability to create realistic and emotionally expressive videos from a static image, speech input, and specified emotional conditions. This technology has potential applications in various domains such as entertainment, education, virtual reality, and human-computer interaction, where realistic and expressive digital avatars can enhance user experience and engagement. For example, in the entertainment industry, this technology can be used to create more lifelike animated characters in movies and video games. In education, it can be utilized to develop interactive and engaging educational content where avatars can deliver lectures or tutorials with appropriate emotional expressions.

While talking face generation technology has many promising applications, there are also potential negative effects and ethical considerations to be aware of:

- Misuse for Misinformation: Generated talking faces could be misused to spread misinformation or fake news by creating videos of individuals saying things they never actually said. This could have serious consequences for public trust and social cohesion.

- Identity Theft and Impersonation: Synthesis of Talking Face generation could be used maliciously for identity theft or impersonation, where someone's likeness is used without their consent to create convincing videos that might be reduced their reputation or privacy.

- Manipulation in Media and Politics: There's a risk that generated talking faces could be used to manipulate media coverage or political discourse, producing

videos of public figures saying or doing things that they did not really say or do.

This proposed system may not produce satisfactory results for speech utterances longer than 8 secs.

However, it is crucial to consider and address the ethical implications and potential misuse of such technology. Safeguards and regulations should be put in place to prevent the creation and dissemination of deceptive or harmful content. It is also important to continue research into improving the system's performance on a wider range of faces and longer speech inputs, ensuring robustness and generalizability.

Bibliography

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Joshua GW Bernstein and Ken W Grant. Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 125(5):3358–3372, 2009.

[3] Carl A Binnie. Bi-sensory articulation functions for normal hearing and sensorineural hearing loss patients. *Journal of the Academy of Rehabilitative Audiology*, 6(2):43–53, 1973.

[4] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[5] Sarah A Cassidy, Björn Stenger, L Van Dongen, Kayoko Yanagisawa, Robert Anderson, Vincent Wan, Simon Baron-Cohen, and Roberto Cipolla. Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. *Computer Vision and Image Understanding*, 148:193–200, 2016.

[6] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018.

[7] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss. *arXiv preprint arXiv:1905.03820*, 2019.

[8] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[9] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[10] Yamamoto Eli, Nakamura Satoshi, and Shikano Kiyohiro. Lip movement synthesis from speech based on hidden markov models. *(No Title)*, pages 154–159, 1998.

[11] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. Generating talking face landmarks from speech. *arXiv e-prints*, pages arXiv–1803, 2018.

[12] Sefik Emre Eskimez, Kenneth Imade, Na Yang, Melissa Sturge-Apple, Zhiyao Duan, and Wendi Heinzelman. Emotion classification: how does an automated system compare to naive human coders? In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2274–2278. IEEE, 2016.

[13] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. Generating talking face landmarks from speech. In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14*, pages 372–381. Springer, 2018.

[14] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. End-to-end generation of talking faces from noisy speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1948–1952. IEEE, 2020.

[15] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480–3490, 2021.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[17] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[18] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20914–20923, 2023.

[19] Karen S Helfer and Richard L Freyman. The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, 117(2):842–849, 2005.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127:1767–1779, 2019.

[22] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[24] Penelope A Lewis, Hugo D Critchley, Pia Rotshtein, and Raymond J Dolan. Neural correlates of processing valence and arousal in affective words. *Cerebral cortex*, 17(3):742–748, 2007.

[25] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

[26] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[28] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[29] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.

[32] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[33] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023.

[34] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

[35] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018.

[36] Vincent Wan, Robert Anderson, Art Blokland, Norbert Braunschweiler, Langzhou Chen, BalaKrishna Kolluru, Javier Latorre, Ranniery Maia, Björn

Stenger, Kayoko Yanagisawa, et al. Photo-realistic expressive text to talking head synthesis. In *INTERSPEECH*, pages 2667–2669. Citeseer, 2013.

[37] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[39] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6609–6619, 2023.

[40] Eli Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano. Lip movement synthesis from speech based on hidden markov models. *Speech Communication*, 26(1-2):105–115, 1998.

[41] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[42] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019.